CHAPTER 1

Introduction.

1.1. Human genetic variation

Sequence variation arises naturally in the human genome due to copying errors during DNA replication. The frequency of a mutant allele in a population is dependent on a number of factors including natural selection. Human genetic variants are typically classified based on the frequency of the minor (less frequent) allele in the population. Common variants have a minor allele frequency (MAF) of at least 5% at the population level. Sequence variants with MAFs of 1–5% are referred to as low-frequency variants, whereas rare variants have a MAF of less than 1% (Frazer et al., 2009).

DNA sequence variants can be further classified with respect to their size, i.e. single-nucleotide variants and structural variants (Frazer et al., 2009). Single-nucleotide polymorphisms (SNPs) are single nucleotide changes. Structural variants range from a few bases to up to several megabases, and include insertion-deletion variants (indels), block substitutions, inversion variants and copy number variants (CNVs).

SNPs are the most prevalent type of sequence variation, with an excess of 15 million being identified thus far. In fact, most SNPs with MAFs of at least 1% across the genome and 0.1% in protein-coding regions have been identified in each of five major population groups, i.e. Europe, East and South Asia, West Africa and the Americas (The 1000 Genomes Project Consortium, 2010). The vast majority of SNPs do not contribute to phenotypic variation and are effectively neutral, therefore achieve commonness at the population level (Kruglyak & Nickerson, 2001). Along with SNPs, there are novel single-nucleotide variants that are under evolutionary constraint or occurred recently in the population, and may segregate in single individuals or families.

Genetic variants that reside at the same recombination interval are typically correlated with one another (**Figure 1-1**). This correlation structure, also known as linkage disequilibrium (LD), is based on the non-random association of alleles due to infrequent recombination, and varies across the genome and different populations (Reich et al., 2001; Pritchard & Przeworski, 2001). At regions of LD, the correlated set of variants defines a limited number of common haplotypes along the chromosomes that are separated by recombination hotspots (Daly et al., 2001; Gabriel et al., 2002).

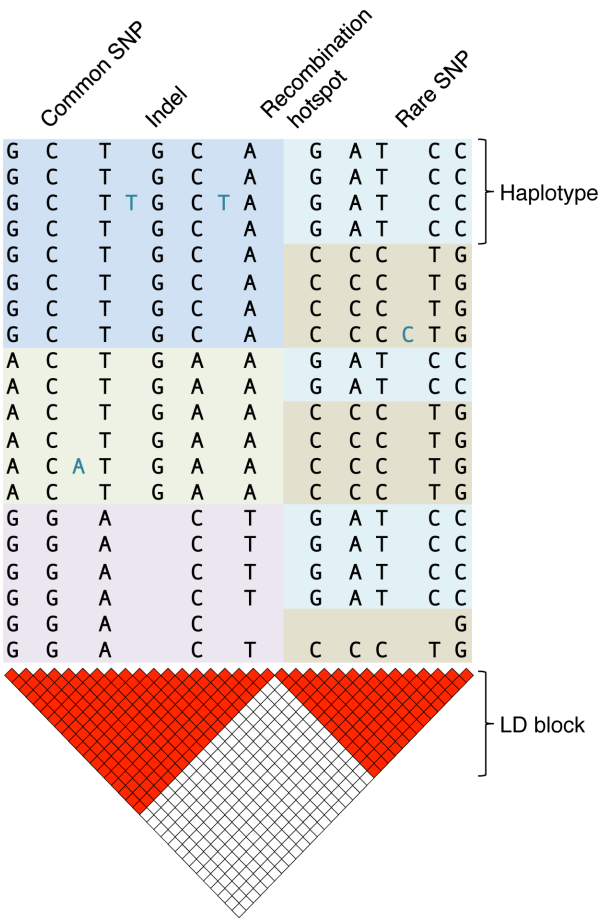

Figure 1-1. Human genetic variation, haplotype structure and linkage disequilibrium. Genetic variation, including common and rare single-nucleotide changes, as well as small indel polymorphisms, is shown in 10 individuals and their corresponding 20 haplotypes. In addition, five rare variants are shown in turquoise (reference nucleotide not shown). The six and five common polymorphisms on the left and right side, respectively, are correlated and organised in common haplotypes (indicated with different background colours). Common haplotypes are separated by a recombination hotspot, with little recombination on either side of the recombination hotspot. The pairwise correlation, or linkage disequilibrium (LD), between the common sites is shown, with red boxes indicating strong correlation ('LD block') and white boxes indicating weak correlation. Figure adapted from Altshuler et al., 2008.

Common haplotypes can be uniquely identified with 'tag' SNPs. Due to the low haplotype diversity in humans (The International HapMap Consortium, 2005), a relatively small number of tag SNPs are sufficient to scan most of the genome for most of the common variation.

The development of high-throughput genotyping arrays (so called 'SNP chips') has enabled the systematic, genome-wide characterisation of haplotype patterns. The International Haplotype Map (HapMap) Project defined these patterns by genotyping ~3 million tag SNPs in samples from Africa, Europe and Asia (The International HapMap Consortium, 2003; 2005; 2007). Over 80% of the SNPs with a MAF of at least 5% are strongly correlated with nearby proxies in ~550,000 LD bins for individuals of European or Asian ancestry and 1,100,000 LD bins for individuals of African ancestry, thereby capturing most of the common genomic variation in these populations (Barrett & Cardon, 2006; Pe'er et al., 2006; The International HapMap Consortium, 2007).

1.2. Genetics of complex traits in humans

1.2.1. Approaches to genetic mapping

Initially, naturally occurring sequence variants were used as markers to systematically trace the inheritance of rare genetic diseases through large affected families. Family-based linkage mapping led to the discovery of many hundreds of genes linked to rare Mendelian diseases (Botstein et al., 1980; Gusella et al., 1983; Donis-Keller et al., 1987). Most Mendelian diseases are characterised by deleterious phenotypes and are caused by rare variants or private mutations in a single gene. However, the approach proved unsuccessful in mapping genes linked to common diseases that show complex inheritance in the general population. Complex traits, including common diseases, are characterised by allelic and locus heterogeneity, whereas gene-environment interactions also play an important role.

The common disease–common variant hypothesis postulates that because the vast majority of genetic variation in the population is due to common variants, the susceptibility alleles for a common complex trait will thus likely be common and of small effect size (Risch & Merikangas, 1996; Lander, 1996; Reich & Lander, 2001). For example, the susceptibility alleles for common diseases, such as diabetes and heart disease, are only moderately deleterious, owing to recent expansion of the population and adaptation to living conditions and lifestyle (Reich & Lander, 2001). The opposing theory, known as common disease–rare variant hypothesis, suggests that common complex traits are the summary of low-frequency, high-penetrance variants (Pritchard, 2001; Bodmer & Bonilla, 2008).

As an alternative to linkage mapping for identifying common genetic variants associated with complex traits, the concept of systematic genome-wide association (GWA) studies was proposed, i.e. the comparison of frequencies of genetic variants among affected and unaffected individuals (Risch & Merikangas, 1996; Lander, 1996; Collins et al., 1997). This concept required the preparation of a catalogue of common variants, which was composed by the International HapMap Project, utilising technological advances in assaying SNPs (The International HapMap Consortium, 2005; 2007). The SNP chips applied in GWA studies typically contain 0.3–2.5 million tag SNPs to assay differences in allele frequencies between case and control samples from a population. Denser chips with up to 5 million tag SNPs are also in use, but their additional content mainly consists of low-frequency variants.

Despite the debate on the two hypotheses, i.e. common disease–common variant vs. common disease–rare variant (Gibson, 2012), GWA studies have been tremendously successful in identifying genetic loci

that are robustly associated with a wide range of clinical conditions and complex traits, including biological measurements.

The Wellcome Trust Case Control Consortium published a landmark GWA study of 14,000 patients, i.e. 2,000 for each of seven major diseases, and a shared set of 3,000 controls. The study led to the discovery of 24 independent association signals for these diseases, which included bipolar disorder, coronary artery disease (CAD), Crohn's disease, rheumatoid arthritis, type 1 diabetes and type 2 diabetes (The Wellcome Trust Case Control Consortium, 2007). In addition, the consortium set guidelines for appropriate statistical analyses for future GWA studies.

1.2.2. Allelic spectrum of genetic variants and missing heritability

A marked difference has been observed for almost all common complex traits between the extent of overall familial aggregation (heritability) and that attributable to common variants identified by GWA studies – despite the use of tens of thousands of samples and meta-analyses of GWA studies (McCarthy et al., 2008; Maher, 2008; Manolio et al., 2009). Some of the 'missing heritability' is expected to be due to low-frequency and rare variants with larger effects (**Figure 1-2**). Such variants are neither frequent enough to be captured by current genotyping arrays (McCarthy & Hirschhorn, 2008), nor is the effect size sufficient to be detected by family-based linkage mapping (Bodmer & Bonilla, 2008). In addition, rare variants are often in low LD with common variants at the same recombination interval, and therefore are not captured by the conventional GWA arrays based on tag SNPs.

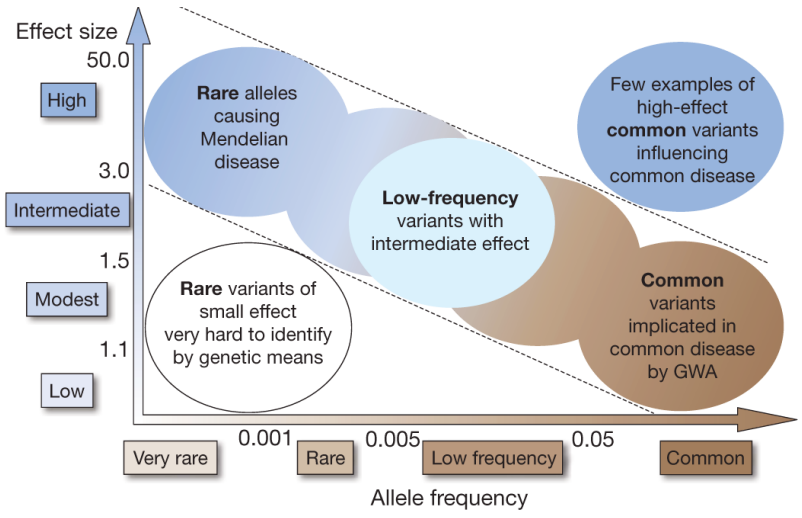

Figure 1-2. Spectrum of genetic variants with respect to allele frequency and genetic effect size. Figure taken from Manolio et al., 2009.

Strategies to capture such variants rely on targeted, deep resequencing of the association locus (Lowe et al., 2007; Rivas et al., 2011; Trynka et al., 2011). For example, resequencing of the *PCSK9* gene at the low-density lipoprotein (LDL) cholesterol and CAD risk locus (Teslovich et al., 2010; Schunkert et al., 2011) revealed rare variants with MAFs of 0.5–1%, which were related to a 15% reduction in plasma levels of LDL cholesterol (Cohen et al., 2006). It has to be noted however, that although the inferred effect sizes of such rare variants are larger, the overall contribution to the heritability may still be small because of their low frequency (Lander, 2011).

Common diseases are usually manifested by a heterogeneous group of clinical events. The study of such heterogeneous disorders may be facilitated by using intermediate quantitative phenotypes that are implicated in the disease aetiology (also known as endophenotypes), for example LDL cholesterol levels and blood pressure in heart disease. The genetic study of these more homogeneous phenotypic subsets may facilitate the analysis of the genetics of multifactorial disease risk and refine its definition – even before disease onset (Plomin et al., 2009). Indeed, sequencing candidate genes in individuals at the extremes of a quantitative trait distribution can identify additional associated variants, both common and rare (Cohen et al., 2004; Kryukov et al., 2009; Johansen et al., 2010; Guey et al., 2011).

Association studies using whole-genome sequencing on moderately large cohorts will soon be feasible, but the analytical challenges to detect disease-associated low-frequency variants in such studies are significant. It is possible to increase power by imputing available large-scale GWA data sets (Li et al., 2009). However, this approach also has limitations, as imputation accuracy drops as MAF decreases and overall, it depends on the density of the initial SNP chip used for genotyping. In parallel, extending the approach to include structural variants, epistasis, gene-gene and gene-environment interactions, may explain a larger fraction of the missing heritability (Maher, 2008; Manolio et al., 2009).

1.2.3. Consensus and challenges in genome-wide association studies

Currently, the catalogue of published GWA studies contains over 1,300 published studies for more than 670 different traits (http://www.genome.gov/gwastudies/; accessed: August 1, 2012). An important inference from this catalogue is that the reported findings instantly suggest new biological hypotheses regarding the molecular basis of the complex trait or disease of interest. For example, some genes are connected to biological pathways not previously suspected to play a role in the disease aetiology. And some diseases with distinct aetiology appear to have a common molecular basis. An example for this pleiotropic overlap is a locus at chromosome 9p21, associated with coronary heart disease, myocardial

infarction (MI), coronary artery calcification, abdominal aortic aneurysm and intracranial aneurysm (Helgadottir et al., 2008; Gretarsdottir et al., 2010; Schunkert et al., 2011; O'Donnell et al., 2011).

More generally, after five years of GWA studies of complex traits and diseases, the following key conclusions can be drawn (McCarthy et al., 2008; Donnelly, 2008; Frazer et al., 2009; Lander, 2011; Visscher et al., 2012). First, a large number of independent genetic loci can influence complex traits. For example, over 180 genetic loci influence adult height (Lango Allen et al., 2010). Second, the majority of the associated common variants exert only small to moderate effect, with an increase in risk of 1.1–1.5 per associated allele. Few exceptions have been reported, for example the association of a coding variant in the *CFH* gene with age-related macular degeneration with an effect of at least 2 per allele (Klein et al., 2005; Haines et al., 2005; Edwards et al., 2005). Third, an association locus can harbour multiple independent susceptibility alleles of different frequency and effect size. For instance, eight common SNPs in five LD blocks at chromosome 8q24 are independently associated with prostate cancer risk (Al Olama et al., 2009). At the *IFIH1* type 1 diabetes association locus (Todd et al., 2007; Barrett et al., 2009), in addition to common variants, rare variants were identified lowering disease risk independently of each other (Nejentsev et al., 2009). Fourth, the impact of common susceptibility variants can vary across ethnic groups, because their allele frequencies are population-specific. And finally, causal sequence variants (i.e. those variants with a direct or indirect functional effect on the phenotypic variation of a trait or disease risk) are usually not instantly identified in GWA studies. As common sequence variants are genetically correlated and located in LD blocks of typically 50–250 kb, the precise functional variants are either not tested or cannot be distinguished. Therefore, the GWA index SNPs identified usually act as proxies for the true causal variant.

Indeed, it is a formidable challenge to provide the molecular and biological explanation for why a locus is associated with a complex trait. The optimal way to translate an association signal into knowledge of the causal variant is unclear. In some cases, the associated SNP may be in LD with nearby coding variants that alter the gene product. However, in the vast majority of cases, the association signal implicates non-protein coding regions (Hindorff et al., 2009). Therefore, the underlying causative variant is likely to affect gene regulatory sequences (discussed in **Section 1.8**). As a consequence, understanding gene regulation and generating comprehensive maps of regulatory elements, which are embedded in the chromatin structure, across all cell types is becoming central to the quest to annotate non-coding GWA signals.

1.3. Primary structure of chromatin

The basic unit of chromatin, known as the nucleosome core particle, contains 147 bp of DNA wrapped in 1.7 turns around a core histone octamer. This octamer is composed of two molecules of each of the histones H2A, H2B, H3 and H4 (Luger et al., 1997). Each core histone encompasses two functional domains; a distinct histone-fold motif required for both histone-DNA and histone-histone contacts within the nucleosome, and polypeptide tail domains allowing for covalent post-translational modifications at specific amino acid residues (Horn & Peterson, 2002; Zhang & Pugh, 2011). Nucleosome core particles are repeated at intervals, with linker DNA of variable length between the units. In most cases, each nucleosome particle is also associated with a linker histone, such as H1, which protects the DNA from nuclease digestion at the core particle boundary. Linker histones contain globular domains that are flanked by NH_2- and COOH-terminal tail domains, where the globular and tail domains are important for binding to the nucleosomes and chromatin folding, respectively (Horn & Peterson, 2002).

Eukaryotic genomes are organised into condensed heterogeneous chromatin fibres, enabling the compaction of DNA into the nucleus. Nucleosomes are arranged as a 10 nm fibre that confers a 5- to 10-fold compaction of the genomic template. This structure is known as 'beads-on-a-string' (**Figure 1-3**). Linker histones can stabilise the 10 nm fibre to form the 30 nm fibre, a higher-order structure characterised by levels of compaction of 50-fold and higher (**Figure 1-3**). Further compaction of the condensed fibres leads to the formation of heterochromatin.

1.4. Determinants of chromatin accessibility

Despite the dense packaging of nucleosomes, regulatory factors and transcriptional machinery must still gain access to the DNA template in order to extract genetic information. Indeed, chromatin structure allows for dynamic changes, where local modulation of chromatin accessibility provides an opportunity to influence the fundamental processes of DNA transcription, replication and repair (Bell et al., 2011).

The key determinants of chromatin accessibility are DNA sequence, ATP-dependent remodelling, histone variants and modifications, as well as competitive protein binding. These determinants are discussed in the following sections.

1.4.1. DNA sequence-dependent nucleosome positioning

Individual nucleosomes can be highly positioned with respect to a specific DNA sequence (Bai & Morozov, 2010). The DNA molecule has to bend sharply around the core histone octamer (Luger et al., 1997). Therefore, nucleosome assembly is facilitated by flexible sequences such as GC-rich sequences, and disfavoured by relatively rigid poly-AT sequences (Field et al., 2008; Kaplan et al., 2009). GC dinucleotides and AA/TT dinucleotides tend to contract and expand the major groove of DNA, respectively (Jiang & Pugh, 2009). Periodic 10 bp intervals of AA/TT/AT dinucleotides contribute to the rotational setting of the DNA helix on the surface of the histone octamer and stabilisation of the nucleosome through bending of the DNA molecule (Segal et al., 2006).

The basic principles described above, of how DNA sequence influences nucleosome assembly, have been established in lower organisms such as budding yeast (*Saccharomyces cerevisiae*) and worms (*Caenorhabditis elegans*). The extent to which DNA sequence can influence chromatin structure and function in human cells remains unclear (Stein et al., 2010; Valouev et al., 2011). A recent study showed that the actual position of nucleosomes is heavily influenced by the activity of ATP-dependent *trans*-acting factors, indicating that DNA sequence alone is not sufficient to predict nucleosome positioning *in vivo* (Zhang et al., 2011).

1.4.2. ATP-dependent chromatin remodelling

Chromatin remodelling complexes can perturb intrinsic histone-DNA interactions via different mechanisms. Generally, these multiprotein complexes use ATP hydrolysis to disassemble or slide histone octamers (Clapier & Cairns, 2009). ATP-dependent remodelling proteins mainly catalyse the replacement of histone subunits, translational repositioning of nucleosomes, and nucleosome removal and deposition.

At active genes, the histones H2A and H3 may be replaced by the histone variants H2A.Z and H3.3, respectively (Jin et al., 2009). The replacement of the H2A and H3 histones in budding yeast is facilitated by the chromatin remodelling complexes SWR1 (Mizuguchi et al., 2004; Kobor et al., 2004) and CHD1 (Konev et al., 2007), respectively. To expose or cover DNA regulatory sites, nucleosomes are repositioned onto A/T-rich DNA tracts via complexes containing ISW2 (Whitehouse et al., 2007). In contrast, the SWI/SNF complex creates DNA loops on the nucleosome surface to control access to DNA regulatory sites (Smith & Peterson, 2005). The activity of SWI/SNF and related complexes can be

enhanced by acetylation of histone tails (Hassan et al., 2001; Suganuma et al., 2008). Through neutralising positively charged lysine residues, acetylation may reduce histone-DNA electrostatic interactions and subsequently disrupt higher-order, repressive chromatin structures (Dion et al., 2005; Wang & Hayes, 2008). Furthermore, nucleosomes may be removed from or deposited onto DNA by the chromatin structure remodelling complex (RSC) or histone chaperons (Jiang & Pugh, 2009).

1.4.3. Histone variants and modifications

Histone variants and modifications confer functionality to nucleosomes, in particular through control of DNA accessibility and regulation of gene expression, but also compaction of chromatin into higher-ordered structures.

The composition of the histone octamer can vary depending on the incorporation of histone variants, which are encoded by different genes and differ in amino acid sequence compared to canonical histones. Only small differences in sequence can have profound effects on histone properties. For example, the histone variant H3.3 differs from the canonical H3 by four amino acid substitutions (Henikoff, 2008; Talbert & Henikoff, 2010). However, H3.3 incorporation into nucleosomes facilitates the eviction and/or repositioning of nucleosomes during transcription. Indeed, H3.3 is highly enriched for modifications associated with transcription, such as H3ac, H3K4me2, H3K4me3 and H3K79me2 (McKittrick et al., 2004; Schwartz & Ahmad, 2005; Wirbelauer et al., 2005; Chow et al., 2005). Histone octamers containing the histone variant H2A.Z form less stable octamers, in turn facilitating chromatin accessibility for transcription initiation at gene promoters (Raisner et al., 2005). In contrast to the canonical histones H3 and H2A, both H3.3 and H2A.Z are incorporated primarily in a DNA replication-independent manner (Talbert & Henikoff, 2010).

Histones are subject to numerous post-translational modifications. Depending on the chemical modification and the amino acid residue targeted, histone modifications regulate chromatin structure, recruit ATP-dependent chromatin remodelling enzymes, influence transcription and affect many other DNA processes, such as repair, replication and recombination (Kouzarides, 2007; Bannister & Kouzarides, 2011). The modifications of the N-terminal histone tails can be grouped into at least eight distinct classes (**Table 1-1**). Importantly, these histone modifications are dynamic and change rapidly with respect to the intracellular signalling conditions and stimuli at the cell surface.

Table 1-1. Overview of different types of histone modifications. Table adapted from Kouzarides, 2007.

Histone modification	Amino acid residues modified	Function(s) regulated
Acetylation (ac)	Lysine	Transcription, repair, replication, condensation
Methylation (me)	Lysine (mono-, di- or trimethyl)	Transcription, repair
Methylation (me)	Arginine (mono- or dimethyl)	Transcription
Phosphorylation (ph)	Serine, threonine	Transcription, repair, condensation
Ubiquitylation (ub)	Lysine	Transcription, repair
Sumoylation (su)	Lysine	Transcription
ADP ribosylation (ar)	Glutamic acid	Transcription
Deimination	Arginine → peptidyl citrulline	Transcription
Proline isomerisation	*Cis*-proline ↔ *trans*-proline	Transcription

Histone modifications function via either the disruption of contacts between nucleosomes to disentangle chromatin or the recruitment of non-histone proteins. Acetylation is the most potent histone modification to affect the electrostatic interactions between histones and DNA or between histones of neighbouring nucleosomes, because it neutralises the basic charge of the lysine residue and thus impacts higher-order chromatin structure (Shogren-Knaak et al., 2006). Non-histone proteins bind to modified histone residues via specific domains. For example, proteins bind to methylated residues via chromodomains, Tudor domains and MBT domains (Royal-superfamily modules), as well as WD40 repeats and PHD finder domains (Taverna et al., 2007). Acetylation and phosphorylation are recognised by bromodomains and a domain within 14-3-3 proteins, respectively. The recruited proteins have enzymatic activities, such as ATPases, that modify chromatin and ultimately activate gene expression (Wysocka et al., 2005).

Both active and silent chromatin, termed euchromatin and heterochromatin, respectively, associate with a distinct set of histone modifications. The euchromatic environment facilitates gene transcription, DNA repair and replication. Actively transcribed euchromatin has high levels of acetylation and is trimethylated at H3K4, H3K36, and H3K79, whereas low levels of acetylation, methylation and phosphorylation are detected in genes that are poorly expressed. In contrast, the heterochromatic environment is transcriptionally inactive and is associated with high levels of methylated sites, for example H3K9me3 and H4K20me3 (Schotta et al., 2004), as well as low levels of acetylation. Heterochromatin structure is important for the protection of chromosome ends and the separation of chromosomes during the cell cycle.

Genome-wide analyses of a subset of histone modifications indicate that histone variants and modifications are selective to specific nucleosome positions along the genome (Kouzarides, 2007; Wang et al., 2008; Hon et al., 2009). Therefore, nucleosomes are likely to serve position-relevant functions, whereby the combinatorial configuration of histone variants and modifications regulates chromatin and the transcription machinery (Kouzarides, 2007; Schones & Zhao, 2008). Indeed, such a 'histone code' may exist, in which these specific organisational combinations provide markers for gene regulatory proteins (Jenuwein & Allis, 2001). As a consequence, assessment of these markers may not only give information about the gene start and end, but also its transcriptional status.

1.4.4. Competitive protein binding

Transcriptional regulation is mostly achieved through sequence-specific binding of transcription factors. Transcription factor binding to nucleosomal DNA can lead directly to histone displacement *in vitro* (Workman & Kingston, 1992), but most transcription factors require exposure of their binding sites (Lomvardas & Thanos, 2001). These binding sites may already be exposed, for example at linker DNA between nucleosomes or A/T-rich tracts that disfavour stable nucleosome formation. Indeed, genome-wide mapping studies of nucleosome occupancy indicate that gene regulatory elements are marked by nucleosome depletion. These are referred to as nucleosome-depleted regions (NDRs) or sites of 'open chromatin', with a high rate of histone replacement at their boundaries (Mito et al., 2007; Henikoff, 2008).

For example, the glucocorticoid receptor binds largely to pre-existing NDRs upon hormone induction. The glucocorticoid receptor binding patterns appear to be pre-determined by cell type-specific differences in baseline (pre-hormone) chromatin accessibility patterns (John et al., 2011). Alternatively, binding sites covered by nucleosomes can become accessible to transcription factor interaction by nucleosome mobilisation, or during spontaneous unwrapping and rebinding of the histone octamer (Li et al., 2005). Repressed promoters often harbour at least one exposed binding site, whereas additional binding sites are inaccessible within nucleosomes, occluding interaction. Initial binding of a 'pioneer' transcription factor can lead to the recruitment of chromatin modifiers, which in turn exposes binding sites for secondary transcription factors required for transcription initiation (Zaret & Carroll, 2011).

1.5. Transcriptional regulation

Biological processes such as development, differentiation, proliferation and apoptosis, depend on the precise spatial and temporal expression of genes. Gene expression is controlled by transcriptional regulation, a cell type-dependent process that is mediated by distinct classes of regulatory elements and factors. This process 'functionalises' the genome, and is tremendously complex.

Eukaryotic expression of protein-coding genes can be regulated at several steps, including transcription initiation and elongation, as well as mRNA processing, transport, translation and stability. Most regulation however occurs during transcription initiation and usually requires general (basic) transcription factors (GTFs), promoter-specific activator proteins (activators) and non-DNA binding co-activators. GTFs, comprising RNA polymerase II, TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, along with the large multisubunit complex Mediator (Malik & Roeder, 2010), collectively form a pre-initiation complex (PIC) at the core promoter.

Transcriptional activity is greatly enhanced by activators, which bind upstream of the core promoter at the proximal promoter and exert their function by facilitating formation or increasing performance of the PIC. Activator proteins, referred to as transcription factors, consist of a sequence-specific DNA-binding domain and an activation domain. Different classes of DNA-binding domains have been described, including basic leucine zipper (bZIP), cysteine-rich zinc finger, ETS, helix-loop-helix (HLH), homeobox and forkhead. The DNA-binding sites for activators are between 6–12 bp, with certain positions of the consensus sequence being relatively constrained and others more variable. Co-activators can form a link between GTFs and DNA-bound activators, thereby modulating activity of the activator and stimulating PIC assembly or modifying chromatin. Importantly, activators can stimulate transcription synergistically, i.e. the regulatory effect of multiple factors is greater than the summed effect of the individual factors (Lin et al., 1990; Carey et al., 1990).

As for activator proteins, transcriptional activity can be repressed by DNA-binding repressor proteins, and non-DNA binding co-repressors. Mechanistically, repressors may compete with activators for DNA-binding sites, restrict the activity of activators by binding in proximity, or bind to silencer and insulator regions. All of these actions may interfere with the assembly and activity of the PIC (Maston et al., 2006).

Different combinations of GTFs and cell type-specific transcription factors bind to multiple transcriptional regulatory elements, some of which are distal from the target genes. This complex

combinatorial control enables the transcription regulation of a large number of protein-coding genes (n=20,000–25,000; International Human Genome Sequencing Consortium, 2004) by a relatively small number of transcription factors (n<3,000; Babu et al., 2004; Vaquerizas et al., 2009; Farnham, 2009). The rate of transcription is dependent on the relative concentration of transcription factors in each cell type.

1.6. Transcriptional regulatory elements

Genes transcribed by RNA polymerase II usually feature two distinct classes of transcriptional regulatory elements, which influence transcriptional activity: a core and proximal promoter; and distal regulatory elements, including enhancers, silencers and insulators (**Figure 1-3**). Transcriptional regulatory elements also include locus control regions (LCRs), which consist of multiple different regulatory elements acting together on a cluster of genes in a specific cell type (Li et al., 2002).

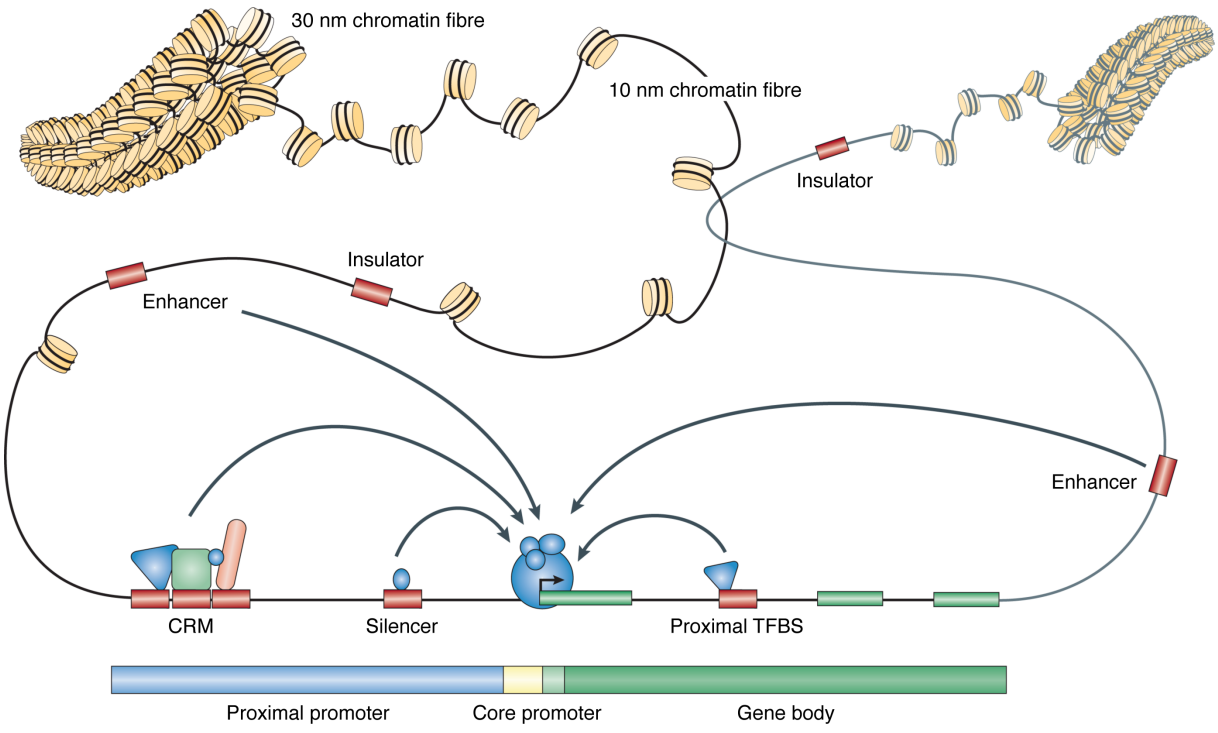

Figure 1-3. Chromatin structure and transcriptional regulatory elements. Nucleosomes are arranged as a 10 nm fibre, known as 'beads-on-a-string'. Stabilisation of this structure through linker histones leads to the formation of a 30 nm fibre. The transcription start site is marked by the core and proximal promoter elements, usually spanning less than 1 kb. Distal regulatory elements, including enhancers, silencers and insulators, are located within 1 Mb of the promoter and may form a complex with the promoter regions through DNA looping to modulate transcriptional activity. Regulatory elements can be physically mapped locally/proximal (*cis*) or distal (*trans*) with respect to transcription start sites. Multiple transcription factor binding

sites (TFBS) at regulatory elements can be arranged in *cis*-regulatory modules (CRMs), also referred to as 'enhanceosomes' (Farnham, 2009). Figure modified from Lenhard et al., 2012.

1.6.1. Promoters

The core promoter element is located within 100 bp of the transcription start site of a gene transcript, containing binding sites for GTFs that are crucial in recruiting the RNA polymerase II, and serving as docking site for the PIC (Hahn, 2004). In metazoa, core promoters may be composed of several distinct motifs (Maston et al., 2006; Sandelin et al., 2007), including TATA box, initiator element (Inr), downstream promoter element (DPE), downstream core element (DCE), TFIIB-recognition element (BRE) and motif ten element (MTE). BRE interacts with TFIIB, whereas TATA box, Inr, DPE, DCE and MTE are recognised by TFIID (Lenhard et al., 2012). The proximal promoter element is located within 1 kb of the transcription start site and contains multiple binding sites for activator proteins (Maston et al., 2006). The chromatin state at promoters is largely invariant across cell types (Heintzman et al., 2009).

In general, nucleosomes frequently contain the histone variants H3.3 and H2A.Z at the promoter and 5'-regions of actively transcribed genes, and exhibit H3ac and H3K4me3 (Schneider et al., 2004; Liang et al., 2004; Kim et al., 2005; Heintzman et al., 2007). The architecture of RNA polymerase II promoters can differ substantially, thus affecting promoter function. Recent studies indicated three main functional promoter classes in vertebrates (**Table 1-2**), which feature different configurations of nucleosomes and preferentially associate with subsets of histone marks (Lenhard et al., 2012).

Another discriminative feature of promoters is the presence of CpG islands (**Table 1-2**). These are genomic sequences of 0.5–2 kb in length with relatively high CpG dinucleotide content compared to bulk DNA, and are located in proximity to about 60% of promoters (Ioshikhes & Zhang, 2000; The ENCODE Project Consortium, 2007). The CpG dinucleotides in CpG islands are usually unmethylated at the fifth carbon position of the cytosine base (Bird, 1987), in contrast to many CpG dinucleotides across the genome. Methylation at CpG dinucleotides, commonly referred to as DNA methylation, is associated with transcriptional silencing. Mechanistically, DNA methylation prevents transcription factors from binding to their recognition sequences, but also enables methylation-specific binding proteins, such as MeCP2, to bind and recruit chromatin silencers (Jones et al., 1998).

Table 1-2. Overview of different functional types of gene promoters in vertebrates. Table modified from Lenhard et al., 2012.

Promoter	Gene function	Correlation with chromatin state	Other common properties
Type I	Tissue-specific transcription in peripheral, terminally differentiated tissues (Carninci et al., 2006; Ernst et al., 2011).	▪ Disordered nucleosome positioning; ▪ TSS covered by nucleosomes; ▪ H3K4me3, H3K4me2 and H3K27ac only downstream of TSS (Ernst & Kellis, 2010); ▪ No RNA polymerase II binding when the genes are not active (Ernst & Kellis, 2010).	▪ Low GC content (Carninci et al., 2006); ▪ Narrow transcription initiation span (Ponjavic et al., 2006); ▪ Enrichment of TATA box; ▪ Mostly no CpG islands; ▪ Depend on *cis*-regulatory modules for regulation; ▪ Key regulatory elements close to promoter (Roider et al., 2009); ▪ Less diversity across promoter states (Ernst & Kellis, 2010).
Type II	Ubiquitously expressed genes or developmentally regulated genes (Carninci et al., 2006; Ernst et al., 2011).	▪ Ordered, precise nucleosome positioning (Rach et al., 2011); ▪ H3K4me3 and H3K27ac at TSS; ▪ NDR at TSS and the immediate upstream region, even when the gene is not expressed (Ernst & Kellis, 2010).	▪ High GC content (Carninci et al., 2006); ▪ Broad, dispersed TSS (Yoshimura et al., 1991); ▪ CpG islands overlap (Deaton & Bird, 2011); ▪ TATA box-depleted; ▪ Short CpG island that typically only overlaps the 5'-end of the gene (Akalin et al., 2009); ▪ Few enhancers nearby (Ernst & Kellis, 2010).
Type III	Differentially regulated genes, often regulators in multicellular development and differentiation (Ernst et al., 2011).	▪ Ordered, precise nucleosome positioning (He et al., 2011); ▪ Bivalent promoter pattern, i.e. broad H3K27me3 (repression) and H3K4me3 (activation) marks at TSS (Bernstein et al., 2006).	▪ High GC content; ▪ Multiple large CpG islands extending into the body of gene; ▪ TATA box-depleted; ▪ High number of enhancers nearby; ▪ Associated with multiple long-range enhancers and with highly conserved non-coding elements (Visel et al., 2009); ▪ Diversity across promoter states (Ernst & Kellis, 2010).

1.6.2. Enhancers

Enhancer elements activate transcription in a spatial and temporal manner by acting on a promoter, where the enhancer function is independent of both distance and orientation relative to the promoter (Ong & Corces, 2011). Enhancers are usually long-distance transcriptional regulatory elements and can be located several hundred kilobases distal from a promoter. For example, mutations in a conserved *cis*-acting regulatory region of *SHH*, which is located 1 Mb upstream of the target gene within intron 5 of *LMBR1*, ultimately cause pre-axial polydactyly, a common limb malformation in humans (Lettice et al., 2002; Lettice et al., 2003; Sagai et al., 2005; Furniss et al., 2008). Typically, enhancers are composed of a dense cluster of transcription factor binding sites (Panne, 2008), with binding of transcription factors working in a cooperative manner to establish gene expression.

Enhancers may function by promoting DNA looping, which brings bound activators in close proximity to the core promoter. This enables PIC formation, followed by gene activation (Vilar & Saiz, 2005; Sexton et al., 2009). There is evidence that these long-range interactions are facilitated and stabilised by Mediator and cohesin (Kagey et al., 2010).

Similar to promoters, enhancer elements are associated with distinct post-translational histone modifications (Barski et al., 2007; Mikkelsen et al., 2007). Importantly, enhancers are marked with highly cell type-specific histone modification patterns (Heintzman et al., 2009). Enhancers are enriched with H3K4me1, H3K4me2 and H3K27ac (Heintzman et al., 2007; Barski et al., 2007; Heintzman et al., 2009), where these marks exhibit cell type specificity and correlation with gene expression patterns (Heintzman et al., 2009). In addition, several studies confirmed the correlation of enhancer elements with both cyclic AMP-responsive element-binding (CREB) protein (CBP) and p300 binding events (Heintzman et al., 2007). CBP and p300 are transcriptional co-activators that feature histone acetyltransferase activity. *In vivo* mapping of p300 binding sites in murine embryonic forebrain, midbrain, limb, as well as human foetal and adult heart tissue, accurately identified enhancer elements that showed tissue-specific gene expression patterns in transgenic mouse assays (Visel et al., 2009; Blow et al., 2010; May et al., 2012). Furthermore, the presence of H3.3 and H2A.Z histone variants, as well as nucleosome depletion, allows for the identification of enhancer sequences (Crawford, Holt, et al., 2006; Barski et al., 2007; Boyle, Davis, et al., 2008; Wang et al., 2008).

Enhancers play an important role during cellular development through the spatiotemporal regulation of gene expression. Specific chromatin signatures associate with enhancers of target genes at certain cellular stages (Cui et al., 2009; Levine, 2010). For example, in human embryonic stem cells, enhancer

elements marked by H3K4me1 and H3K27ac are located in proximity to actively expressed genes, whereas enhancers marked by H3K4me1 and H3K27me3 are linked to inactive genes (referred to as poised enhancers). During cellular differentiation and embryonic development, these inactive genes are then switched on, resulting in the replacement of H3K27me3 with H3K27ac (Orford et al., 2008; Creyghton et al., 2010).

1.6.3. Silencers

Silencer elements are regulatory regions that confer a negative effect on the transcriptional output of a target gene. As for enhancer elements, many silencers act in a distance- and orientation-independent manner with respect to the promoter, and may reside as part of the proximal promoter or distal enhancer, or act as independent modules (Ogbourne & Antalis, 1998; Narlikar & Ovcharenko, 2009). Silencer elements consist of binding sites for transcription factors, referred to as repressors, and generally share many similar features to enhancers (**Section 1.6.2**).

The levels of the histone modifications H3K27me2 and H3K27me3 correlate with gene silencing. Furthermore, silencers are weakly associated with H3K9me3 and H3K9me2 (Barski et al., 2007). In contrast to its correlation with gene activation at promoter regions, the histone variant H2A.Z associates with gene silencing at genic regions.

1.6.4. Insulators

Insulator elements (also known as boundary elements) are regulatory regions that interfere with the activating or repressing transcriptional activity between adjacent loci (Maston et al., 2006; Narlikar & Ovcharenko, 2009). Insulators are 0.5–3 kb in length, act in a position-dependent but orientation-independent manner and usually contain multiple binding sites for transcription factors. There are two functional classes, enhancer-blocking and barrier insulator elements (Recillas-Targa et al., 2002; Gaszner & Felsenfeld, 2006). Enhancer-blocking insulators prevent the interaction and communication of an enhancer with a promoter when placed in-between. In contrast, barrier insulators prevent the spread of repressive heterochromatin structure, or favour the formation of active euchromatin structure, thereby creating independent structural domains.

The CTCF transcription factor (CCCTC-binding factor) is a highly conserved zinc finger protein with diverse roles in gene regulation, and organises global chromatin architecture by recruiting chromatin modifying enzymes (Phillips & Corces, 2009). The mediation and stabilisation of intra- and interchromosomal interactions is facilitated by cohesin (Wendt et al., 2008). Importantly, CTCF has been implicated in the blocking of enhancer activity and heterochromatin spreading (Bell et al., 1999; Hark et al., 2000; Cuddapah et al., 2009). The binding sites of CTCF are relatively invariant across different cell types, and show enrichment for H2A.Z but are not correlated with other histone modifications (Kim et al., 2007; Barski et al., 2007; Heintzman et al., 2009; Hon et al., 2009; Ernst & Kellis, 2010).

1.7. Methods for mapping gene regulatory elements

The systematic genome-wide identification of sequences with regulatory potential, particularly enhancer elements, was first accomplished by comparative genomic strategies, e.g. cross-species sequence alignment and comparison (Woolfe et al., 2005; Prabhakar et al., 2006; Pennacchio et al., 2006; Visel et al., 2008). Most of these studies relied on the assumption that the sequences of gene regulatory elements are under evolutionary constraint (Loots et al., 2000; Nobrega et al., 2003; Pennacchio et al., 2006). However, this approach has limitations. Most importantly, while an enhancer sequence may be conserved, its activity is dependent on many different factors. For example, spatial and temporal activity patterns in the developing or adult organism. Therefore, deletion of such conserved enhancer sequences often does not result in an apparent phenotype (Ahituv et al., 2007; Pennacchio & Visel, 2010).

Alternative methods are required to identify newly evolved human regulatory elements. Complementary to comparative genomic methods, biochemical assays using isolated cells and cell nuclei can define gene regulatory elements by revealing common patterns of nucleosome arrangements and modifications. Indeed, May et al. showed that heart enhancers, which were neither evolutionarily nor functionally conserved between the human and mouse genomes, could be identified using genome-wide occupancy profiling of p300/CBP (May et al., 2012). Among many others (Schones & Zhao, 2008; Zhou et al., 2011; The ENCODE Project Consortium, 2011), such assays include endonuclease digestion, DNA methylation footprinting, formaldehyde-assisted isolation of regulatory elements and chromatin immunoprecipitation of histone modifications and related proteins. These methods are discussed in the following sections. Recent advances in high-density microarray and massively parallel next-generation sequencing technologies (Metzker, 2010) have enabled the application of these

experimental strategies on a genome-wide scale (Schones & Zhao, 2008). Here, next-generation sequencing has the advantage of greater coverage, larger dynamic range, higher resolution and less noise compared to microarrays.

Large-scale efforts, for example initiated by the ENCODE Project (The ENCODE Project Consortium, 2004), BLUEPRINT (Adams et al., 2012) and NIH Roadmap Epigenomics Mapping (Bernstein et al., 2010) consortia, generate high-resolution genome-wide maps of chromatin states using various biochemical assays across a multitude of cell types, physiological conditions and developmental stages.

1.7.1. Endonuclease digestion

Nucleases are sensors of accessible chromatin structure (**Figure 1-4 A**). In the nucleosome structure, DNA-protein interactions protect chromosomal DNA from digestion by endonucleases, such as deoxyribonuclease I (DNase I). Conversely, sites of reduced nucleosome occupancy and high histone-turnover are preferentially digested (Wu et al., 1979; Wu, 1980; Mito et al., 2007; Boyle, Davis, et al., 2008). These sites are referred to as DNase I hypersensitive sites, and frequently represent active regulatory elements, such as promoters, enhancers and insulators (Crawford et al., 2004; Sabo et al., 2004; Dorschner et al., 2004; Crawford, Davis, et al., 2006; Sabo et al., 2006; Boyle, Davis, et al., 2008). Recent studies applied high-throughput DNA sequencing to map DNase I hypersensitive sites with single-nucleotide resolution. Within these sites of DNase I hypersensitivity, the 'footprints' of regulatory proteins can be detected, i.e. the exact position of the DNA-protein interaction (Hesselberth et al., 2009; Pique-Regi et al., 2011; Boyle et al., 2011).

Another endonuclease-based method uses micrococcal nuclease (MNase), which preferentially cleaves DNA in linker regions between nucleosomes, as well as in NDRs (**Figure 1-4 A**). This method generates mono- and oligonucleosomes, therefore allowing for accurate mapping of nucleosome positioning. Genome-wide MNase digestion profiles indicate that nucleosomes tend to be characteristically positioned or depleted at gene regulatory regions, in particular at promoters and 3'-ends of transcription units (Schones et al., 2008; Valouev et al., 2011).

Both DNase I and MNase digestion methods involve enzyme titration, followed by characterisation of the digested DNA by microarrays or high-throughput sequencing. Nucleolytic methods involve many handling steps and are complicated by variations in commercial enzyme activity. In addition, nucleases are generally unable to resolve more condensed structures of chromatin.

1.7.2. DNA methylation footprinting

Differential DNA accessibility can be measured by methylation footprinting with exogenous DNA methyltransferases (Singh & Klar, 1992; Gottschling, 1992; Fatemi et al., 2005). In this approach, methyltransferases, such as *M.CviP* I or *M.Sss* I, methylate cytosines in exposed linker DNA sequences, but less efficiently in sequences that are bound up in nucleosomes (**Figure 1-4 B**). Unlike cytosine, methyl-cytosine is protected against bisulphite conversion to uracil *in vitro*. Therefore, methylation footprints can be identified by DNA sequencing following bisulphite treatment, whereby GC dinucleotides and GU/T-rich sequences are inferred to be nucleosome-free and nucleosomal, respectively. This method presents a complementary approach to nuclease-based assays, as it offers a different range of chromatin sensitivity (Bell et al., 2010).

1.7.3. Formaldehyde-assisted isolation of regulatory elements (FAIRE)

Nucleosome depletion can be assessed by the formaldehyde-assisted isolation of regulatory elements (FAIRE) assay. Formaldehyde preserves protein-protein and protein-DNA interactions *in vivo* (Fragoso & Hager, 1997). FAIRE involves fixation of chromatin with formaldehyde, chromatin shearing by sonication, and subsequent phenol-chloroform extraction (**Figure 1-4 C**). The technique is based on the differential segregation of nucleosomal and nucleosome-free soluble DNA in the phenol-chloroform and aqueous phase, respectively (Nagy et al., 2003). Isolated DNA can be either fluorescently labelled and hybridised to a high-density microarray or subjected to next-generation sequencing (Giresi et al., 2007; Giresi & Lieb, 2009; Gaulton et al., 2010; Simon et al., 2012).

The major advantages of FAIRE are its relatively simple protocol and cost efficiency, providing an attractive method for NDR mapping in many cell types or under various conditions. Furthermore, FAIRE makes prior treatments of cells unnecessary, with other methods requiring nuclei preparation. Determination of the appropriate nuclease concentration for each procedure is omitted. However, its resolution in nucleosome mapping is lower compared to nuclease-based methods.

1.7.4. Chromatin immunoprecipitation (ChIP)

The genome-wide distribution of histone modifications, chromatin-associated proteins and transcription factors can be monitored by using specific antibodies in chromatin immunoprecipitation

assays (**Figure 1-4 D**). The method uses cross-linked sheared chromatin as starting material. Protein-DNA complexes are selectively immunoprecipitated using specific antibodies to a protein of interest. This is followed by isolation of precipitated DNA and detection on microarrays, high-throughput sequencing or mass spectrometry (Barski et al., 2007; Johnson et al., 2007; Robertson et al., 2007; Mikkelsen et al., 2007; Park, 2009).

The resulting distribution profile of protein binding events appears to be static. However, modifications on histones and the binding of regulatory proteins are dynamic and rapidly changing. Immunoprecipitation experiments also heavily rely on the availability and specificity of the antibody used, as antibodies may cross-react with the same or a similar modification located on other histones (Kouzarides, 2007; Zhang & Pugh, 2011; Egelhofer et al., 2011).

As mentioned in **Section 1.6.2**, the mapping of binding sites of the co-activator p300 using ChIP followed by next-generation sequencing provides a powerful, conservation-independent strategy of discovering tissue-specific enhancer sequences (Visel et al., 2009; Blow et al., 2010; May et al., 2012).

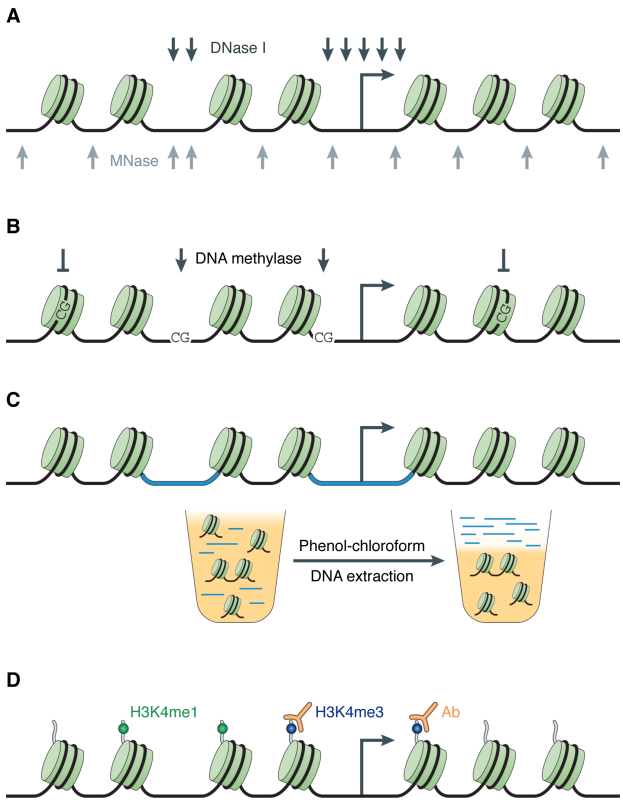

Figure 1-4. Experimental methods for mapping gene regulatory elements. Experimental strategies for identifying gene regulatory elements by mapping chromatin accessibility include (**A**) endonuclease digestion, (**B**) DNA methylation footprinting, (**C**) formaldehyde-assisted isolation of regulatory elements (FAIRE) and (**D**) chromatin immunoprecipitation (ChIP) of histone modifications. Figure modified from Bell et al., 2011.

1.8. Gene regulatory elements in human disease

Chromatin accessibility correlates with DNA susceptibility to mutations, including insertions and deletions. Analysis of the ENCODE pilot regions (The ENCODE Project Consortium, 2007) revealed that the small indel rate is reduced up to two-fold in open chromatin regions, but not the SNP density rate (Clark et al., 2007). In Bushmen genomes, SNPs tend to be enriched in NDRs at promoter elements near nucleosome borders (Schuster et al., 2010). Therefore, nucleosome organisation may contribute to the evolution of gene regulatory elements, and ultimately impact diversity and human disease susceptibility (Zhang & Pugh, 2011).

Rare Mendelian diseases are usually caused by mutations in protein-coding genes, including non-synonymous nucleotide substitutions (i.e. those causing an amino acid change), insertions and deletions. However, rare disorders and common complex diseases have also been associated with genetic variants in transcriptional regulatory elements, as well as members of the transcriptional machinery (Kleinjan & van Heyningen, 2005; Wray, 2007; Epstein, 2009). For example, mutations in the proximal promoter element of *GP1BB* result in reduced binding of the transcription factor GATA1 and reduced *GP1BB* expression, causing Bernard-Soulier syndrome (Ludlow et al., 1996), a rare bleeding disorder characterised by giant platelets. Conversely, mutations in *GATA1* itself have been linked to a number of haematopoietic disorders (Cantor, 2005). In a second example, common low-penetrance variants located intronic of *RET* in an enhancer region are associated with Hirschsprung disease risk (a congenital defect of the colon). Of note, the contribution of these variants to disease risk is 20-fold greater than the rare coding mutations (Emison et al., 2005; Grice et al., 2005). In a third example, an integrative analysis of ChIP and DNase I hypersensitivity data sets, multispecies sequence alignment, as well as gene expression profiles, revealed a pathogenic mechanism underlying a form of the inherited blood disorder α-thalassemia. A gain-of-function regulatory SNP in a non-coding region was shown to create a new promoter-like element that modulates α-globin gene expression (De Gobbi et al., 2006).

The advent of GWA studies led to the rapid discovery of a large number of sequence variants associated with complex traits, including common diseases, of which the majority are located at non-protein coding regions (**Section 1.2.3**). Many of these associations were found to influence the expression levels of nearby genes in cell types and tissues of biological relevance to the phenotype of interest (Cookson et al., 2009; Nicolae et al., 2010; Nica et al., 2010). The genome-wide assaying of quantitative levels of gene expression and its correlation with genetic variation has facilitated the interpretation of non-coding GWA signals (Libioulle et al., 2007; Moffatt et al., 2007; Barrett et al., 2009). Indeed, such variation in gene transcript abundance is highly heritable and can be mapped as a

quantitative trait (Dixon et al., 2007; Emilsson et al., 2008; Cheung & Spielman, 2009). These are referred to as expression quantitative trait loci (eQTLs). Intersection of GWA index SNPs with eQTLs revealed that 10–15% of these SNPs act via a known eQTL (Cookson et al., 2009).

However, in order to identify causal non-coding variants from GWA signals and establish their molecular mechanism, it is necessary to integrate a number of different functional data types (Harismendy & Frazer, 2009; Hawkins et al., 2010; Freedman et al., 2011; Cooper & Shendure, 2011; Baker, 2012). Annotating the association locus using chromatin accessibility, histone modification, transcription factor binding and other data sets in relevant cell types can identify putative active regulatory elements (**Figure 1-5**). Intersection of these 'regulatory maps' with GWA signals may in turn identify candidate functional variants. Ideally, such correlation should be done using complete sequence information through resequencing of the GWA locus and dense genotyping (**Section 1.2.2**). Candidate functional variants can then be validated in experimental assays.

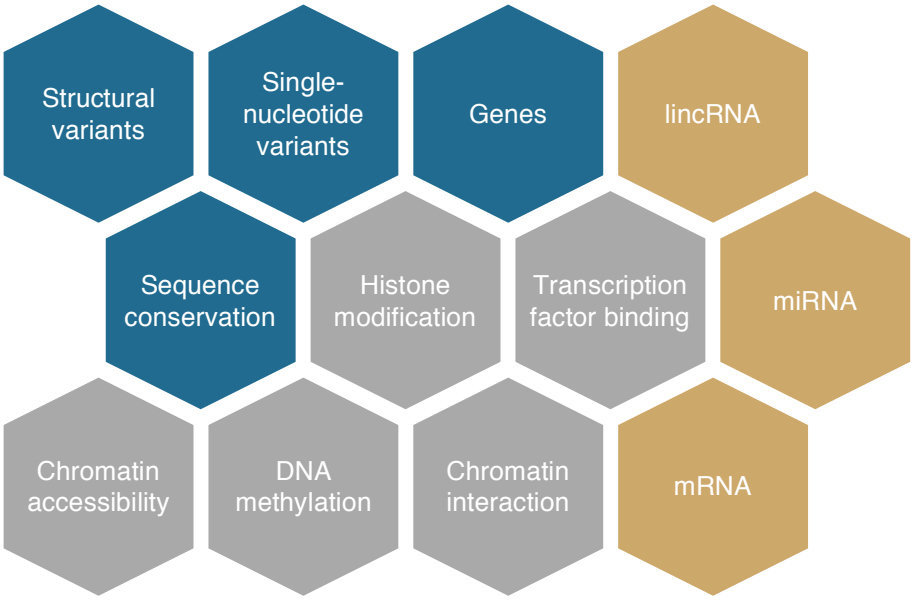

Figure 1-5. Integration of multiple data sets for annotating GWA loci. Functional annotation is based on genomic data (blue), i.e. gene annotation, sequence variation and conservation. Both epigenomic (grey) and transcriptomic (gold) data provide means to assigning function to the sequence information, but dependent on specific cell types. Many additional data sets can be applied for functional characterisation (The ENCODE Project Consortium, 2011). Abbreviations: lincRNA: large intergenic non-coding RNA; miRNA: microRNA; mRNA: messenger RNA.

Four recent landmark studies described such integrative analyses and demonstrated how associations at non-coding regions can be taken down to a molecular level. First, common variants at chromosome 8q24 were found to be associated with colorectal cancer susceptibility, despite being located over

300 kb away from the nearest gene (Haiman et al., 2007; Tomlinson et al., 2007; Zanke et al., 2007; Yeager et al., 2008). Two studies provided a mechanistic explanation for the association signal, applying sequence conservation, ChIP, chromosome conformation capture (van Steensel & Dekker, 2010) and *in vitro* and *in vivo* reporter assays. The studies showed that the GWA tag (index) SNP rs6983267 affects TCF4 transcription factor binding at an enhancer region, which physically interacts with the *MYC* oncogene (Pomerantz et al., 2009; Tuupanen et al., 2009). Second, Hardismendy et al. examined the molecular basis underlying the chromosome 9p21 GWA signal associated with CAD and MI (McPherson et al., 2007; Helgadottir et al., 2007; Schunkert et al., 2011). Two common risk alleles were located over 150 kb from the nearest gene at an enhancer region, where they disrupt a binding site for the transcription factor STAT1 (Harismendy et al., 2011). Using a novel technique to detect long-distance chromosomal interactions, this enhancer was shown to interact with the genes *CDKN2A*, *CDKN2B* and *MTAP*, as well as an interval downstream of *IFNA21*, in a vascular cell type. Thus, the study provided a link between CAD genetic susceptibility and the response to inflammatory signalling (Harismendy et al., 2011). Third, the common non-coding polymorphism rs12740374 at chromosome 1p13, associated with plasma LDL cholesterol and MI risk (Willer et al., 2008; Kathiresan et al., 2008; Teslovich et al., 2010; Schunkert et al., 2011), was shown to create a novel C/EBP transcription factor binding site and alter the expression of *SORT1* in human-derived hepatocytes. Using both small interfering RNA knockdown and viral overexpression in mouse liver, *Sort1* was shown to alter atherogenic plasma low-density and very low-density lipoprotein levels (Musunuru et al., 2010). Finally, Gaulton et al. mapped sites of open chromatin at type 2 diabetes risk loci using FAIRE. An islet-specific NDR was found to contain the susceptibility variant rs7903146 (Grant et al., 2006) located intronic of *TCF7L2*. The variant showed allelic imbalance in the FAIRE signal in heterozygous human islet samples, and was found to alter enhancer activity (Gaulton et al., 2010). As discussed in **Chapters 3**, **4** and **5**, we have contributed further examples to this list, investigating index SNPs from GWA studies of haematological traits.

The main difficulties in identifying non-coding functional variants relevant to a particular trait are our limitations to recognise functionally active non-coding sequences, laborious detection of active regulatory regions, inaccessibility of relevant cell types and tissues in humans, and a lack of tools available for establishing functional consequences. In addition, such variants are less likely to have a profound phenotypic effect. This is because regulatory variants affect the expression level of a gene but not the structure and function of a protein, and hence their impact may be absorbed by secondary, redundant regulatory elements.

1.9. Haematopoietic system and genetics of haematological traits

The haematopoietic system is among the best-characterised cellular differentiation systems in mammals. Multipotent haematopoietic stem cells (HSCs), which are capable of self-renewal, reside in the bone marrow. Here, HSCs have the potential to reconstitute the entire haematopoietic system through differentiation into various progenitor cells that become progressively restricted to single lineages (**Figure 1-6**). After maturation and differentiation along specific pathways, these progenitors are destined to become mature cells in the peripheral blood (Orkin, 2000).

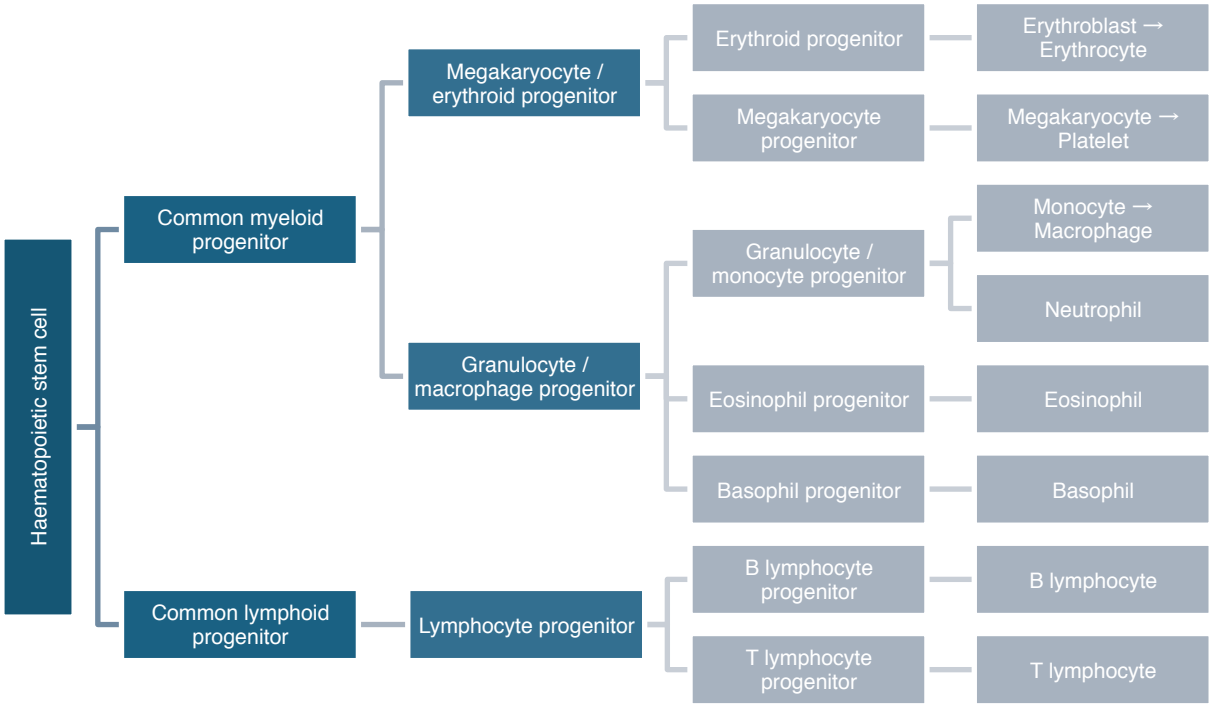

Figure 1-6. Simplified scheme of lineage determination in human haematopoietic hierarchies. In haematopoiesis, there are two major programmes, myeloid and lymphoid. HSCs, common myeloid (CMP) and lymphoid (CLP) progenitors are multipotent. From these progenitors, committed precursors for the various lineages arise, eventually forming mature blood cells. These mature cells can be experimentally distinguished by cell surface and other markers. Figure adapted from Orkin & Zon, 2008.

The differentiation of HSCs into mature haematopoietic lineages is tightly regulated by combinatorial networks of transcription factors that provoke differentiation and maturation along lineages (Miranda-Saavedra & Göttgens, 2008). For example, enforced expression of GATA1/2 transcription factors in murine myeloid cells causes conversion to a megakaryocyte phenotype (Visvader et al., 1992; Visvader & Adams, 1993; Visvader et al., 1995).

Mature blood cells are responsible for a wide range of cellular functions, including the transport of oxygen to tissues by haemoglobin-containing red blood cells (erythrocytes), homoeostasis and wound repair by platelets (arise by budding from large, polyploid megakaryocytes), and innate and adaptive immunity by white blood cells (lymphocytes, granulocytes and monocytes).

Haematological traits, including the count and volume of blood cells in peripheral blood, are highly heritable and vary widely between individuals (Garner et al., 2000). For example, platelet count has a heritability of 80%, as determined by twin studies (Evans et al., 1999). Haematological parameters have widespread clinical relevance and deviations outside normal ranges are indicative of several disorders, such as infectious and immune diseases. In addition, several studies indicated that elevated white cell count is an independent risk factor for CAD and MI (Ensrud & Grimm, 1992; Danesh et al., 1998; Hoffman et al., 2004). Likewise, increases in both spontaneous platelet aggregation (Trip et al., 1990) and mean platelet volume (Boos & Lip, 2007; Chu et al., 2010; Slavka et al., 2011) have been shown to confer genetic risk for cardiovascular disease in epidemiological studies. Indeed, larger platelets carry greater pro-thrombotic potential, and are metabolically and enzymatically more active (Kamath et al., 2001). In addition, genetic loci associated with platelet count, e.g. *SH2B3-ATXN2* and *PTPN11*, have been reported to be associated with CAD and MI, suggesting a possible role for platelets as an intermediate phenotype (Soranzo, Spector, et al., 2009).

Studying the genetic architecture of haematological traits experimentally is particularly appealing, because of the simple phenotypes at the cellular level, relatively easy access to primary cell types from peripheral blood, and suitable animal models.

1.10. Thesis aims and objectives

Common sequence variation at non-protein coding regions of the genome has been driving most of the association signals in GWA studies of complex traits thus far. In most cases, neither the causative variant at the GWA locus nor its exact molecular mechanism are known. It is likely that some of the causative sequence variants exert their effect on phenotype through regulation of gene expression levels.

The aim of this thesis is to identify functional sequence variants at genetic loci associated with haematological and cardiovascular-related traits, and to establish their molecular mechanism.

Figure 1-7 illustrates the general concept of how genetic signals can be translated into molecular mechanism, in which this thesis focuses on the first four elements.

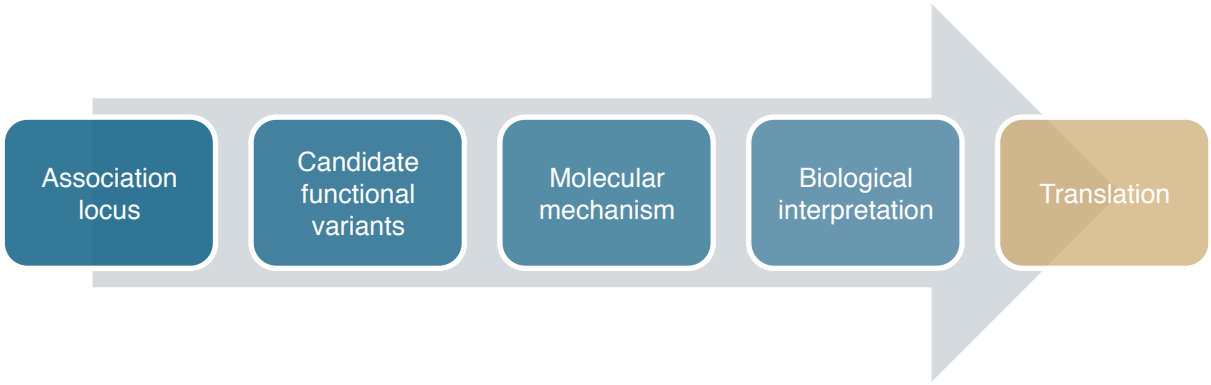

Figure 1-7. Translation of genetic signals into molecular mechanism and biological understanding. These biological insights can then be used to further aid the development of new treatments and diagnostic tools, such as reliable biomarkers.

To address this aim, I set the following objectives:

1. To generate maps of open chromatin indicating sites of regulatory activity in cell types of the myeloid lineage using the FAIRE technique (**Chapters 3** and **4**);
2. To intersect these cell type-specific open chromatin maps with GWA signals of haematological and cardiovascular-related traits to identify candidate functional variants (**Chapters 3** and **4**);
3. To provide a proof-of-concept by defining the molecular mechanism of a GWA locus associated with platelet volume and function, using both experimental and computational methods (**Chapter 5**);
4. To explore the use of open chromatin maps to annotate low-frequency variants linked to Mendelian diseases (**Chapter 6**).