MAPS OF OPEN CHROMATIN – FROM GENETIC SIGNALS TO FUNCTION.

Dirk Stefan Paul

A thesis submitted for the degree of
Doctor of Philosophy

Darwin College, University of Cambridge
Wellcome Trust Sanger Institute

September 2012

DECLARATION.

This dissertation is the result of my own work and does not contain the outcome of work done in collaboration with others, except where indicated in the text. The work described here has not been submitted for a degree, diploma or similar qualification at any other university or institution. I confirm that this dissertation does not exceed the word limit specified by the Biology Degree Committee at the University of Cambridge.

Dirk Stefan Paul
August 31, 2012

DARWIN COLLEGE
CAMBRIDGE

wellcome trust
sanger
institute

Maps of open chromatin – from genetic signals to function.

Dirk Stefan Paul ▪ Darwin College

Genome-wide association (GWA) studies have been very successful in identifying genetic loci associated with complex traits, including common diseases. Many GWA signals map outside protein-coding regions suggesting that the underlying functional variants may influence phenotype through regulation of gene expression. This thesis aims to address the challenge of identifying functional variants at these regions, and interpreting their biological consequences.

I applied the formaldehyde-assisted isolation of regulatory elements (FAIRE) method to map nucleosome-depleted regions (NDRs), marking active regulatory elements. First, I used FAIRE-chip to map NDRs at known genetic loci associated with haematological and cardiovascular traits in a megakaryocytic and an erythroblastoid cell line. Then, I used FAIRE-seq to map NDRs genome-wide in primary human megakaryocytes and erythroblasts. I showed that (i) cell type-specific NDRs can guide the identification of regulatory variants; (ii) sequence variants associated with the corresponding platelet and erythrocyte traits were enriched in NDRs in a cell type-dependent manner; (iii) the majority of candidate regulatory variants in NDRs at known platelet quantitative trait loci affected protein binding, suggesting that this is a common mechanism by which sequence variation influences quantitative trait variation. As a proof-of-concept, I established the molecular mechanism of the 7q22.3 platelet volume and function locus. I identified a megakaryocyte-specific NDR harbouring the non-coding GWA index SNP rs342293, found to differentially bind the transcription factor EVI1 and affect *PIK3CG* gene expression in platelets and macrophages. Gene expression profiling of *Pik3cg* knockout mice indicated that PIK3CG is associated with gene pathways with an established role in platelet function. Lastly, I used the FAIRE data sets to characterise two low-frequency SNPs at the *RBM8A* locus, identified through exome sequencing of patients with thrombocytopenia with absent radii (TAR), a rare congenital malformation syndrome. This work revealed that compound inheritance of one of these two SNPs and a rare null allele causes TAR. The two regulatory variants located in an NDR resulted in reduced *RBM8A* transcription *in vitro* and reduced expression of the encoded Y14 protein in platelets from individuals with TAR. These data implicate insufficient Y14, a subunit of the exon-junction complex, as the cause of TAR syndrome.

This thesis demonstrates the utility of maps of open chromatin for identifying regulatory variants associated with genetic traits, and highlights through two examples how such data sets can be used to establish a functional mechanism. This information can aid the development of new treatments and diagnostic tools.

PUBLICATIONS.

The work described in this thesis resulted in the following publications (* indicates equal contribution):

1. Paul, D.S., Nisbet, J.P., Yang, T.P., Meacham, S., Rendon, A., Hautaviita, K., Tallila, J., White, J., Tijssen, M.R., Sivapalaratnam, S., Basart, H., Trip, M.D., Cardiogenics Consortium, MuTHER Consortium, Göttgens, B., Soranzo, N., Ouwehand, W.H. & Deloukas, P. (2011). Maps of open chromatin guide the functional follow-up of genome-wide association signals: application to hematological traits. **PLoS Genet.** *7*, e1002139.

   Research highlight: Open chromatin and hematologic traits (2011). **Nat. Genet.** *43*, 728.

2. Albers, C.A.*, Paul, D.S.*, Schulze, H.*, Freson, K., Stephens, J.C., Smethurst, P.A., Jolley, J.D., Cvejic, A., Kostadima, M., Bertone, P., Breuning, M.H., Debili, N., Deloukas, P., Favier, R., Fiedler, J., Hobbs, C.M., Huang, N., Hurles, M.E., Kiddle, G., Krapels, I., Nurden, P., Ruivenkamp, C.A., Sambrook, J.G., Smith, K., Stemple, D.L., Strauss, G., Thys, C., van Geet, C., Newbury-Ecob, R., Ouwehand, W.H.* & Ghevaert, C.* (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit *RBM8A* causes TAR syndrome. **Nat. Genet.** *44*, 435-439.

   Research highlight: Deficiency of the Y14 protein is a critical factor underlying the etiology of thrombocytopenia with absent radii syndrome (2012). **Clin. Genet.** *82*, 29-30.

3. Nürnberg, S.T.*, Rendon, A.*, Smethurst, P.A., Paul, D.S., Voss, K., Thon, J.N., Lloyd-Jones, H., Sambrook, J.G., Tijssen, M.R., HaemGen Consortium, Italiano, J.E., Jr., Deloukas P., Göttgens B., Soranzo N., Ouwehand W.H. (2012). A GWAS sequence variant for platelet volume marks an alternative *DNM3* promoter in megakaryocytes near a MEIS1 binding site (2012). **Blood.** *In press.*

4. van der Harst, P.*, Zhang, W.*, Leach, I.M.*, Rendon, A.*, Verweij, N.*, Sehmi, J.*, Paul, D.S.*, Elling, U.*, HaemGen Consortium (2012). 75 genetic loci influencing the human red blood cell (2012). **Nature.** *In press.*

5. Paul, D.S.*, Albers, C.A.*, Rendon, A.*, Voss, K., Stephens, J., HaemGen Consortium, van der Harst, P., Chambers, J.C., Soranzo, N., Ouwehand, W.H.* & Deloukas, P.* (2012). Maps of open chromatin highlight cell type-specific patterns of regulatory sequence variation at hematological trait loci. **Genome Res.** *Under review.*

ACKNOWLEDGMENTS.

First and foremost, I would like to thank my supervisor and mentor Panos Deloukas for giving me the opportunity to work in his group, and for his guidance and exceptional support throughout my thesis. Many thanks also to my thesis committee, Willem Ouwehand, Nicole Soranzo and Stephan Beck, for great discussions, advice on my work and putting me in touch with the right people. I am grateful for financial support from the Marie-Curie Training Network 'NetSim' (215820).

I want to thank Christina Hedberg-Delouka, Annabel Smith and the Sanger Institute Board of Graduate Studies, as well as Debra Fletcher at the Department of Haematology at the University of Cambridge, for their great support throughout my studies.

I was very fortunate to work with a remarkable group of people, who not only helped me with my research projects, but have become close friends, in particular Kees, Augusto, Stephane, George, James, Kathy and Pelin. Great thanks to the rest of the Deloukas and Ouwehand groups.

For making my time at the institute memorable and very entertaining, I thank the PhD classes of 2012 and 2013, especially Jovana, Aparna, Lars, E-Pien, Jenn, Johan, Jolene, Madushi, Laure, Jared and Sammy. A big thank you to my close friends at Darwin College, in particular Meenal, Nam, Tim, Hubi and the 2010 May Ball and 2011 Student Association Committees.

Last but not least, I would like to thank my parents and grandparents for their love and unreserved support throughout my studies. And finally, I thank Laura for proofreading my thesis… and because she is sooo amazing – du bist einfach die Beste!

Dedicated to my parents and grandparents.

CONTENTS.

LIST OF FIGURES.

LIST OF TABLES.

5. Functional follow-up of the platelet volume and function locus at chromosome 7q22.3 95

6. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in *RBM8A* causes TAR syndrome 110

8. Appendix 130

NOMENCLATURE.

ac	acetylation
ADP	adenosine diphosphate
ATP	adenosine triphosphate
bp	base pair
CAD	coronary artery disease
CBP	cyclic AMP-responsive element-binding (CREB) protein
CEU	HapMap 'European' population: Utah residents with Northern and Western European ancestry from the CEPH collection
ChIP	chromatin immunoprecipitation
chr	chromosome
CNV	copy number variant
CRM	*cis*-regulatory module
CTCF	CCCTC-binding factor
DBP	diastolic blood pressure
DNA	deoxyribonucleic acid
DNase I	deoxyribonuclease I
EB	erythroblast
EB cell	erythroblastoid cell line
EJC	exon-junction complex
EMSA	electrophoretic mobility shift assay
ENCODE	Encyclopedia of DNA Elements
eQTL	expression quantitative trait locus
FAIRE	formaldehyde-assisted isolation of regulatory elements
FDR	false discovery rate
GO	Gene Ontology
GREAT	Genomic Regions Enrichment of Annotations Tool
GTF	general (basic) transcription factor
GWA	genome-wide association
HapMap	International Haplotype Map Project
Hb	haemoglobin
HPC	haematopoietic progenitor cell

HSC	haematopoietic stem cell
HYP	hypertension
indel	insertion-deletion variant
iPS cell	induced pluripotent stem cell
kb	kilobase
LCL	lymphoblastoid cell line
LD	linkage disequilibrium
LDL	low-density lipoprotein
lincRNA	large intergenic non-coding RNA
MAF	minor allele frequency
Mb	megabase
MCH	mean cell/corpuscular haemoglobin
MCHC	mean cell/corpuscular haemoglobin concentration
MCV	mean cell/corpuscular volume
me	methylation
MEP	megakaryocyte-erythrocyte progenitor
MI	myocardial infarction
miRNA	microRNA
MK	megakaryocyte
MK cell	megakaryocytic cell line
MNase	micrococcal nuclease
MO	monocyte
MPV	mean platelet volume
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
NDR	nucleosome-depleted region
NMD	nonsense-mediated RNA decay
OMIM	Online Mendelian Inheritance in Man
PCV	packed cell volume
PIC	pre-initiation complex
PLS	platelet signalling
PLT	platelet count
QC	quality control
QTL	quantitative trait locus

RBC	red blood cell count
RNA	ribonucleic acid
RRM	RNA-binding domain
s.d.	standard deviation
SBP	systolic blood pressure
SNP	single-nucleotide polymorphism
STS	sequence-tagged site
TAR	thrombocytopenia with absent radii
TSS	transcription start site
UCN	unique case number
UTR	untranslated region
VWF	Von Willebrand Factor
WBC	white blood cell count

CHAPTER 1

Introduction.

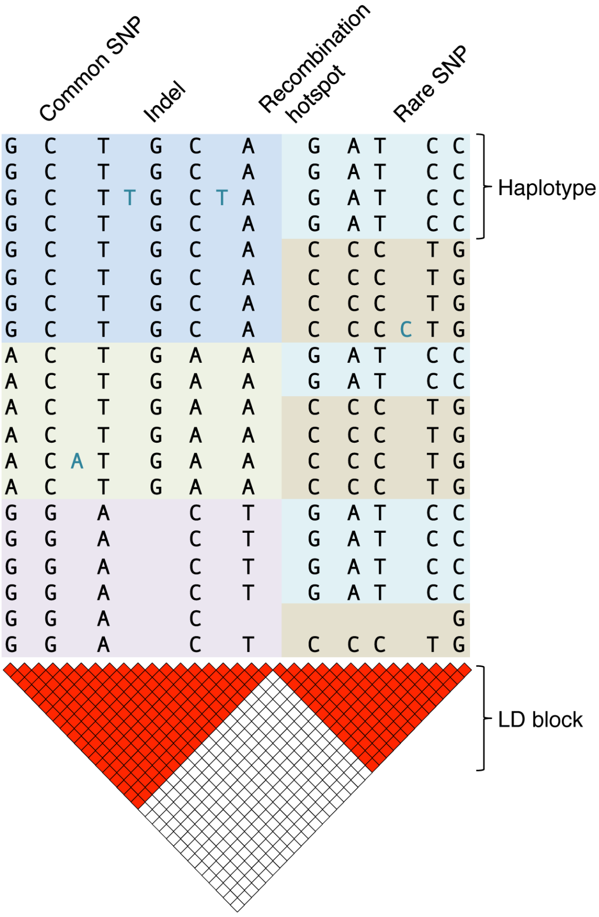
1.1. Human genetic variation

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).



```
G C   T   G C   A G A T   C C
G C   T   G C   A G A T   C C
G C   T T G C T A G A T   C C
G C   T   G C   A G A T   C C
G C   T   G C   A C C C   T G
G C   T   G C   A C C C   T G
G C   T   G C   A C C C   T G
G C   T   G C   A C C C C T G
A C   T   G A   A G A T   C C
A C   T   G A   A G A T   C C
A C   T   G A   A C C C   T G
A C   T   G A   A C C C   T G
A C A T   G A   A C C C   T G
A C   T   G A   A C C C   T G
G G   A       C   T G A T   C C
G G   A       C   T G A T   C C
G G   A       C   T G A T   C C
G G   A       C   T G A T   C C
G G   A       C                 G
G G   A       C   T C C C   T G
```

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.

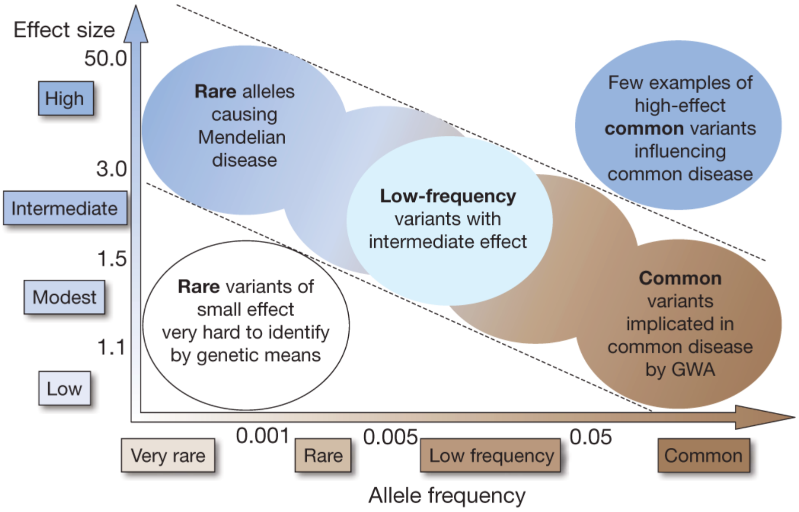

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

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).

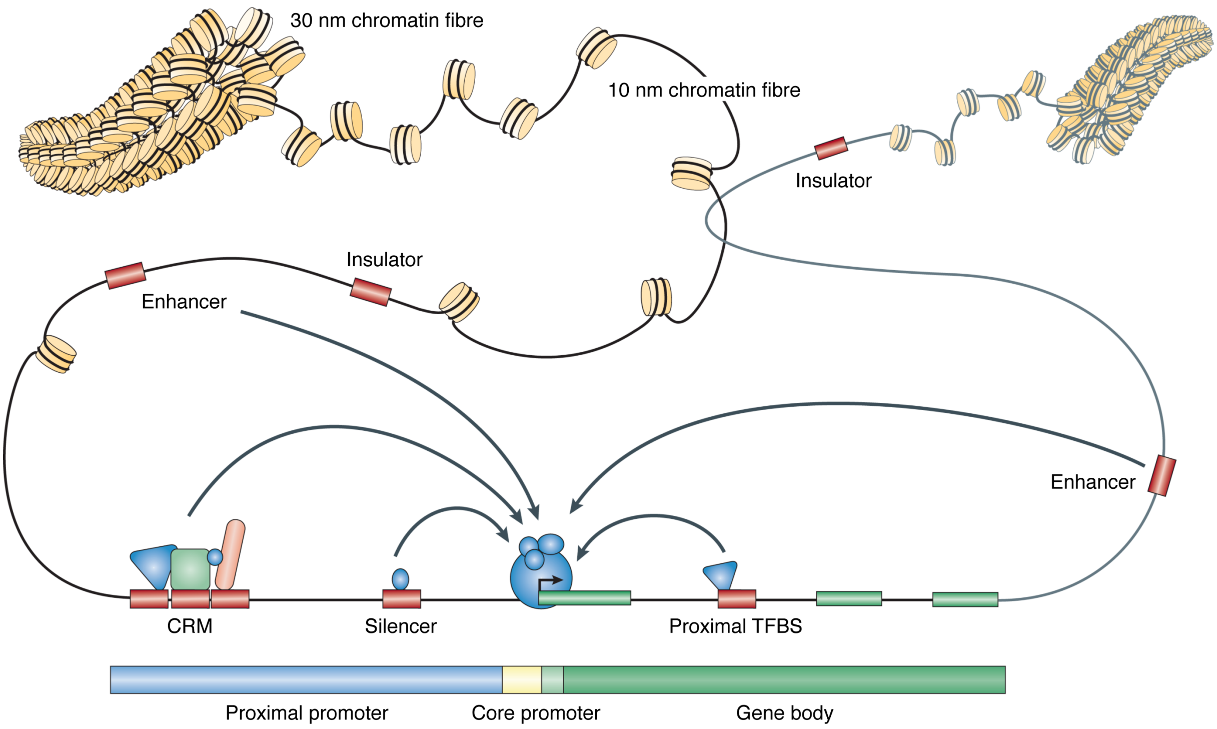


Figure 1-3. **Chromatin structure and transcriptional regulatory elements.** Nucleosomes are arranged as a 10 nm fibre, known as 'beads-on-a-string'. Stabilisation of this structure through linker histones leads to the formation of a 30 nm fibre. The transcription start site is marked by the core and 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).

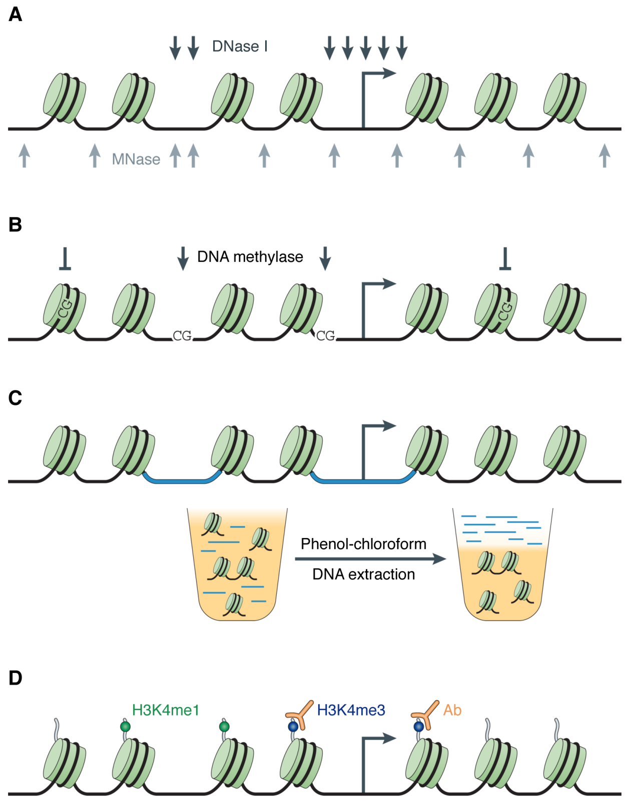


Figure 1-4. Experimental methods for mapping gene regulatory elements. Experimental strategies for identifying gene regulatory elements by mapping chromatin accessibility include (A) endonuclease digestion, (B) DNA methylation footprinting, (C) formaldehyde-assisted isolation of regulatory elements (FAIRE) 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.

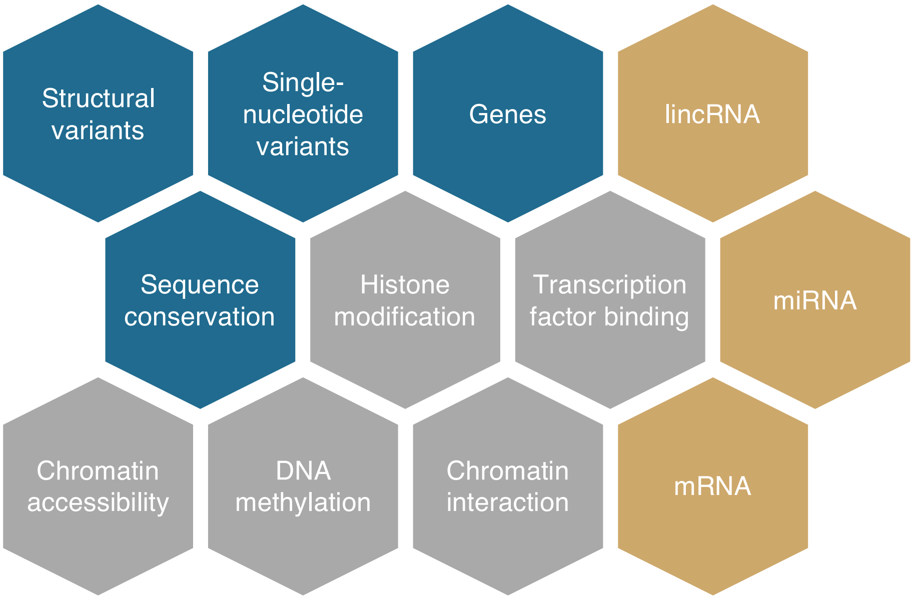


Figure 1-5. Integration of multiple data sets for annotating GWA loci. Functional annotation is based on genomic data (blue), i.e. gene annotation, sequence variation and conservation. Both epigenomic (grey) and transcriptomic (gold) data provide means to assigning function to the sequence information, 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).

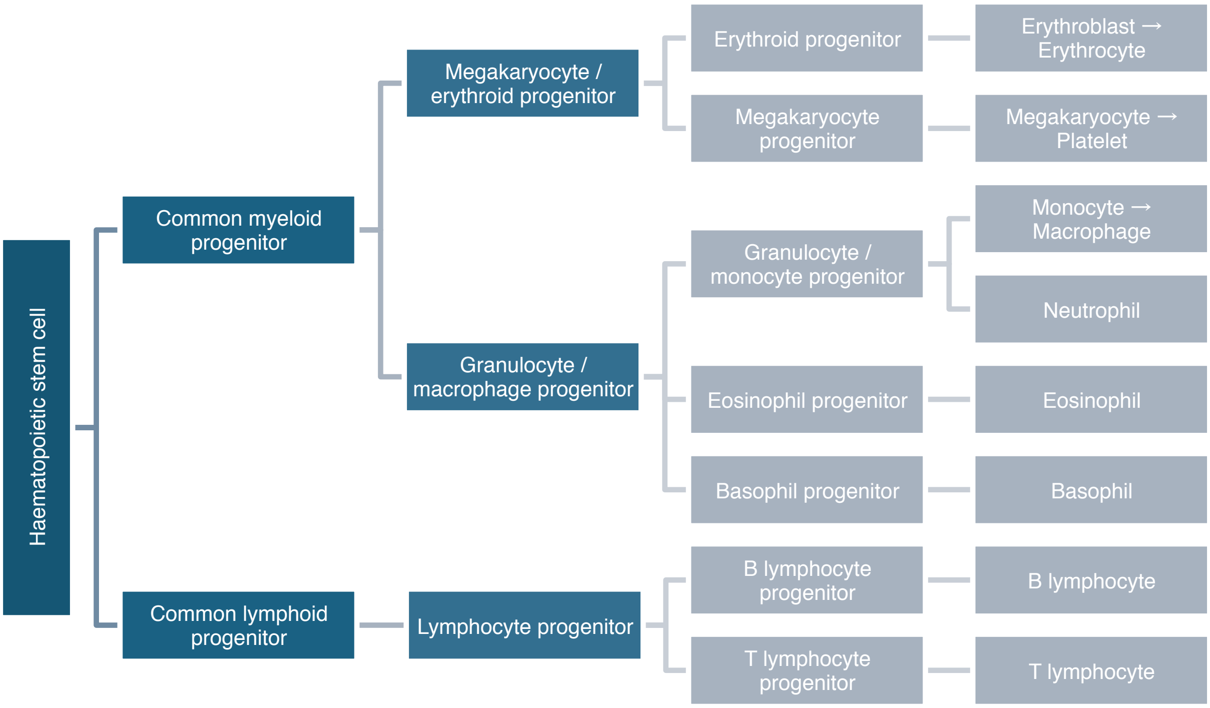


Figure 1-6. Simplified scheme of lineage determination in human haematopoietic hierarchies. In haematopoiesis, there are two major programmes, myeloid and lymphoid. HSCs, common myeloid (CMP) and lymphoid (CLP) progenitors are multipotent. From these progenitors, committed precursors for the various 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.

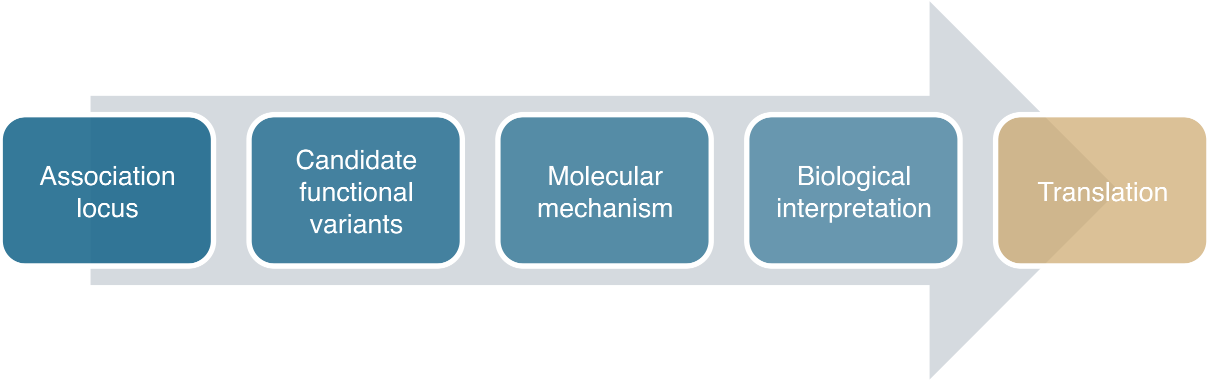


Figure 1-7. Translation of genetic signals into molecular mechanism and biological understanding. These biological insights can then be used to further aid the development of new treatments and diagnostic tools, such as reliable biomarkers.

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**).

CHAPTER 2

Materials and methods.

29-56

2.1. Culture and preparation of cell lines

The cell lines CHRF-288-11 and K562 were kindly provided by Katrin Voss (Department of Haematology, University of Cambridge; NHS Blood and Transplant, Cambridge, UK).

CHRF-288-11. The human megakaryoblastic cell line CHRF-288 was originally established from a biopsy of a metastatic solid tumour in a 17 month old infant with acute megakaryoblastic leukaemia (Witte et al., 1986). The cloned cell line (designated CHRF-288-11) exhibits markers characteristic of megakaryocytes and platelets. The cells also produce both basic fibroblast growth factor (bFGF) and transforming growth factor-β (TGF-β) (Fugman et al., 1990; Saito, 1997). CHRF-288-11 cells were maintained in RPMI-1640 medium [Sigma-Aldrich] supplemented with 20% horse serum (heat inactivated) [Invitrogen] and 1% L-glutamine-penicillin-streptomycin solution [Sigma-Aldrich].

K562. The human cell line K562 was established from a patient with chronic myeloid leukaemia (CML) in acute blast crisis (Lozzio & Lozzio, 1975). The glycoprotein pattern of K562 cells shows a striking similarity to that observed in normal erythrocytes (Andersson et al., 1979; Koeffler & Golde, 1980; Tabilio et al., 1983). K562 cells were maintained in RPMI-1640 medium [Sigma-Aldrich] supplemented with 10% foetal bovine serum (non-heat inactivated) [Biosera], 2 mM GlutaMAX-I [Invitrogen] and 1% L-glutamine-penicillin-streptomycin solution [Sigma-Aldrich].

Subculturing. Cells were taken from liquid nitrogen storage and thawed for 2 min at 37°C in a water bath. Cells were immediately removed from the vile with a sterile pipette and diluted in 20 ml of fresh growth medium. To remove dimethyl sulfoxide (DMSO), the cell suspension was centrifuged for 5 min at 72xg. Cells were resuspended in supplemented growth medium to ~$5x10^5$ cells/ml. CHRF-288-11 and K562 cells were grown at 37°C, 5% CO_2 and 100% humidified atmosphere. The cell suspension was diluted to ~$1x10^5$ cells/ml and fed or subcultured every 2–3 days. All cell culture work was performed in a class II microbiological safety cabinet.

Cell freezing. Cells were collected at early passage numbers and concentrated to $5x10^6$ cells/ml in fresh growth medium supplemented with 10% DMSO [Sigma-Aldrich]. The cell suspension was aliquoted into 1.2 ml cryopreservation vials [Nunc] (1 ml of cell suspension per tube) and placed in a freezing container [Mr. Frosty, Nalgene] at -80°C overnight. The freezing container was filled with isopropyl alcohol (IPA) [VWR BDH Prolabo], which ensures the 1°C/min-cooling rate required for successful cryopreservation of cells. Finally, vials were stored in liquid nitrogen until use.

2.2. Isolation, culture and preparation of primary cells

Ethics statement. Umbilical cord blood was obtained after informed consent under a protocol approved by the NHS Cambridgeshire Research Ethics Committee (REC 07/MRE05/44).

Monocyte isolation. Monocytes (MOs) were isolated from residual leukocytes obtained following apheresis platelet collections from NHS Blood and Transplant donors. Each sample (7.5 ml) was diluted 1:2 with PBE buffer (PBS [Sigma-Aldrich] at pH=7.2, 2 mM EDTA [Sigma-Aldrich] and 0.5% BSA [Sigma-Aldrich]) and gently layered onto the membrane of a 50 ml Leucosep tube [Greiner Bio-One]. By using these tubes, optimal separation of peripheral blood mononuclear cells (PBMCs) from human whole blood can be achieved by means of density gradient centrifugation using a porous, high-grade polyethylene barrier. Samples were centrifuged for 15 min at room temperature (RT) and 800xg. The PBMC layer was transferred into a fresh 50 ml tube. PBMCs from different Leucosep tubes were pooled, washed three times with 25 ml PBE buffer and centrifuged for 5 min at RT and 500xg. PBMCs were counted, diluted to $1x10^8$ cells/ml with PBE buffer and transferred into 5 ml polystyrene round-bottom tubes [BD Biosciences]. MO isolation was performed using the EasySep Human CD14 Positive Selection Kit [StemCell Technologies] according to the manufacturer's instructions. MOs, which strongly express the CD14 antigen, were targeted with tetrameric antibody complexes recognising CD14 and dextran-coated magnetic particles. Labelled cells were then separated using magnets, without the use of columns.

Megakaryocyte and erythroblast culture. Cord blood of newborns was collected into cord blood collection bags [MacoPharma]. $CD34^+$ haematopoietic progenitor cells (HPCs) were purified using the CD34 MicroBead Kit [Miltenyi Biotec] following the manufacturer's instructions. First, the $CD34^+$ cells were magnetically labelled with CD34 microbeads. Then, the cell suspension was loaded onto a column that is placed in a magnetic field, retaining the $CD34^+$ cells within the column. Purity (92–98%) and viability of HPCs were tested by flow cytometry. For *in vitro* differentiation of HPCs into megakaryocytes (MKs), 150,000 cells/ml/well were seeded in serum-free medium [CellGro SCGM, CellGenix] supplemented with 50 ng/ml human recombinant thrombopoietin (rhTPO) [CellGenix] and 10 ng/ml interleukin-1β (rhIL-1β) [Miltenyi Biotech]. To differentiate HPCs into erythroblasts (EBs), 5,000 cells/ml/well were seeded in serum-free medium supplemented with 6 U/ml erythropoietin (rhEPO) [R&D Systems], 10 ng/ml interleukin-3 (rhIL-3) [Miltenyi Biotech] and 100 ng/ml stem cell factor (rhSCF) [R&D Systems]. Cells were cultured for 7–10 days at 37°C and 5% CO_2. On the day of harvest, a cell aliquot was stained with 0.2% Trypan Blue [Sigma-Aldrich] and live cells counted using a haemocytometer [InCyto C-Chip, VWR International].

Cell morphology and flow cytometric analysis. For cell morphological analysis (**Figure 2-1**), aliquots of 50,000 cells were centrifuged onto a glass slide for 5 min at RT and 400xg and stained with modified Wright's stain using an automated slide stainer [HemaTek 1000, Miles Laboratories]. Stained cytospins were microscopically analysed [Axiovert 40 CFL, AxioCam HSc and AxioVision v4.5, Carl Zeiss MicroImaging]. Aliquots of 300,000 cells were used for flow cytometry. MOs were stained with human anti-CD14-PE clone TUK4 and anti-CD45-FITC clone c29/33 [Alere], as well as FITC and PE mouse monoclonal IgG1 isotype control [BD Biosciences]. After antibody incubation, 500 µl PBE buffer (PBS [Sigma-Aldrich] at pH=7.2, 2 mM EDTA [Sigma-Aldrich] and 0.5% BSA [Sigma-Aldrich]) and 5 µg/ml 7-amino actinomycin D (7-AAD) [Invitrogen] were added. Flow cytometric analysis of MKs and EBs was performed as previously described (Macaulay et al., 2007; Tijssen et al., 2011), using the following antibodies: human anti-CD41a-APC clone HIP8, anti-CD42a-FITC clone ALMA.16, anti-CD235a-FITC clone GA-R2 (HIR2) and anti-CD34-PE clone 581 [BD Biosciences]. All samples were analysed on the CyAn ADP 9-Color flow cytometer using the software Summit v4.3.02 [Beckman Coulter].

Ploidy stain of megakaryocytes. An aliquot of $1x10^6$ MKs was fixed with 70% (w/v) ethanol [Sigma-Aldrich] for 30 min at RT, washed once with PBE buffer (PBS [Sigma-Aldrich] at pH=7.2, 2 mM EDTA [Sigma-Aldrich] and 0.5% BSA [Sigma-Aldrich]) and stained with human anti-CD41a-APC clone HIP8 [BD Biosciences] or matched isotype control, as described above. After centrifugation, cells were resuspended in 500 µl staining buffer (465 µl PBE buffer, 5 µl 10% Tween-20 [Sigma-Aldrich], 5 µl of 10 mg/ml RNase A [Sigma-Aldrich] and 25 µl propidium iodide [Sigma-Aldrich]). After incubation for 30 min at 37°C, DNA content was determined using flow cytometry.

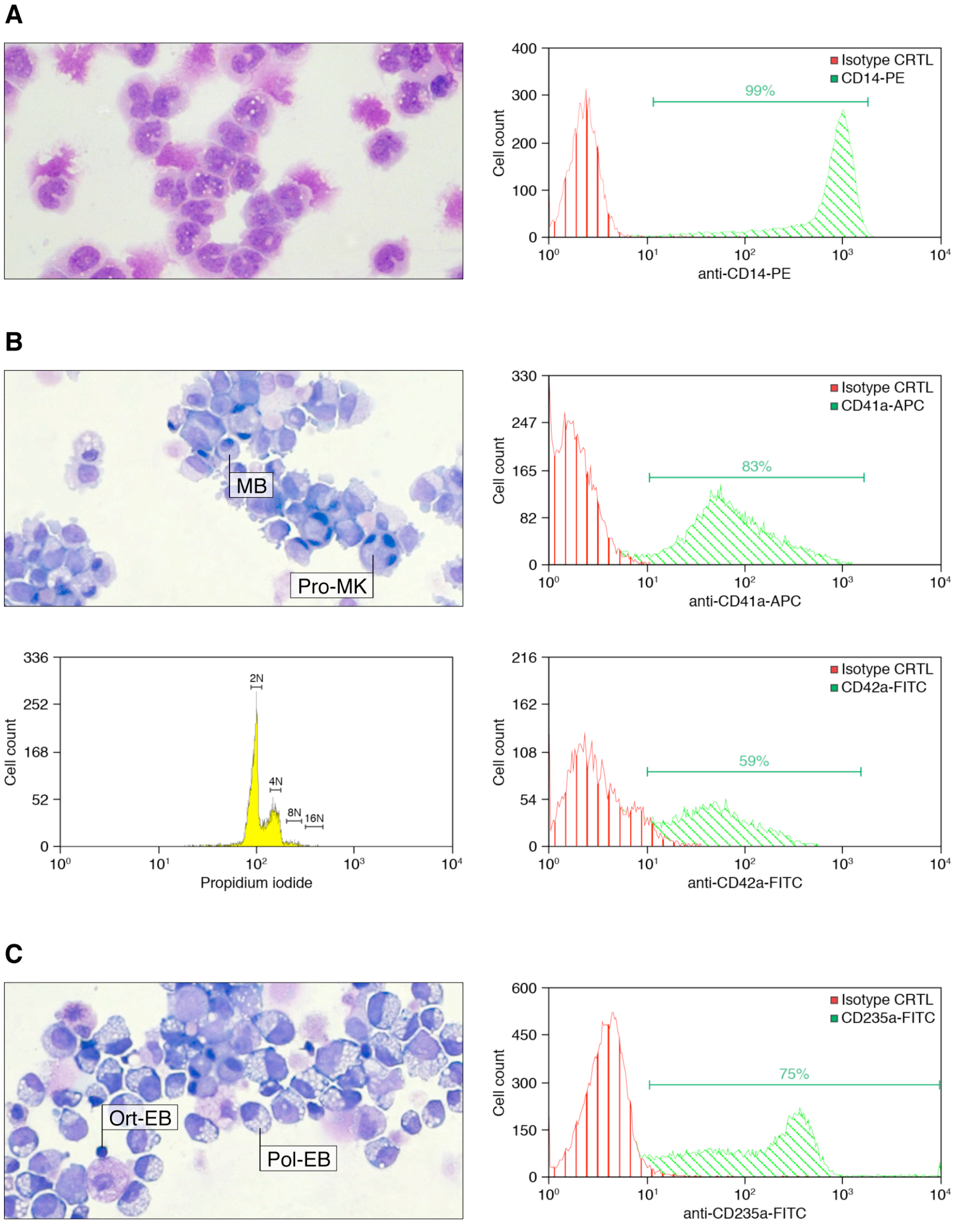


Figure 2-1. Characterisation of primary human MOs, MKs and EBs. Left panel: Representative images of stained cytospins (magnification, x40). Right panel: Gated cell populations with cell type-specific markers are shown in green, with corresponding IgG control populations in red. Indicated ranges represent average expression of gated cell populations of independent biological triplicates. For all cell cultures, more than 90% of cells tested negative for CD34 expression and corresponding isotype controls, as determined by flow cytometric analysis. Markers of other lineages were not detected. (A) Mature MOs were isolated based on morphological evaluation through microscopic analysis of stained cytospins (Goasguen et al., 2009). Isolated cells showed morphological characteristics of mature MOs, i.e. lobulated nucleus, condensed chromatin, occasional granules but no visible nucleolus. Histograms of flow cytometric analysis determined that 98% (range, 98–99%) of isolated cells expressed CD14 (and CD45, data not shown). (B) After megakaryocyte culture, analysis of stained cytospins and modified Wright's stain showed that the majority of cells were progenitor cells (megakaryoblasts, MB). MK cultures also contained pro-megakaryocytes (Pro-MK, horseshoe-shaped nucleus) and mature MKs (multinucleated) (Zeuner et al., 2007). Flow cytometric characterisation

revealed that 77% (range, 71–83%) of cells expressed CD41a and 44% (range, 28–59%) also expressed CD42a, which is only expressed by mature MKs. Ploidy analysis showed that 24.5% of MKs were 4N or higher. (**C**) In erythroblast cultures, cells were predominantly at a polychromatic (Pol-EB) and orthochromatic (Ort-EB) stage of differentiation, as determined by microscopic analysis (Panzenböck et al., 1998). Flow cytometric analysis showed that 71.6% (range, 68–75%) of cells expressed CD235a.

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

Formaldehyde cross-linking. Cells (as described in **Table 2-1**) in fresh growth medium were transferred into a 150 mm x 25 mm cell culture dish or a 50 ml tube, and 37% formaldehyde [Merck Calbiochem] was directly added to the cell suspension to a final concentration of 1%. The cells were incubated at RT with gentle shaking on an orbital shaker [SO3, Stuart]. The cross-linking time across experiments varied and is reported in **Table 2-1**. To quench the fixation, 2.5 M glycine [AppliChem] was added to a final concentration of 125 mM, and the cell suspension was shaken for 5 min. The cells were collected by centrifugation for 5 min at 340xg. Cell pellets were washed with cold 1x PBS to remove all residual media. Cells were collected for 5 min at 340xg.

Cell lysis for optimisation and FAIRE-chip experiments. The cell pellet was resuspended in lysis buffer L1 (50 mM HEPES-KOH [Sigma-Aldrich] at pH=7.5, 140 mM NaCl [VWR BDH Prolabo], 1 mM EDTA [Amresco] at pH=8.0, 0.50% Igepal CA-630 [USB Corporation], 0.25% Triton X-100 [Sigma-Aldrich] and 10% glycerol [VWR BDH Prolabo]) to a concentration of 10x10^6 cells/ml, and incubated for 10 min on ice. The cell suspension was aliquoted into 1.5 ml tubes (500 μl of lysate per tube), and cells were collected for 5 min at 4°C and 1,300xg. Next, the pellet was resuspended in 500 μl of lysis buffer L2 (200 mM NaCl [VWR BDH Prolabo], 1 mM EDTA [Amresco] at pH=8.0, 0.5 mM EGTA [Merck Calbiochem] at pH=8.0 and 10 mM Tris-HCl [Sigma-Aldrich] at pH=8.0), and incubated for 10 min at RT. The cell suspension was spun down for 5 min at 4°C and 1,300xg. The pellet was then resuspended in 300 μl of lysis buffer L3 (100 mM NaCl [VWR BDH Prolabo], 1 mM EDTA [Amresco] at pH=8.0, 0.5 mM EGTA [Merck Calbiochem] at pH=8.0, 10 mM Tris-HCl [Sigma-Aldrich] at pH=8.0, 0.1% Na-deoxycholate [Sigma-Aldrich] and 0.5% (w/v) N-lauroylsarcosine sodium salt [Sigma-Aldrich]). Before use, 2x EDTA-free Protease Inhibitor [Complete Mini, Roche] were added to each 25 ml of L3, and immediately sonicated.

Cell lysis for FAIRE-seq experiments. The cell pellet was resuspended in 5 ml of ice-cold PBS supplemented with 1x EDTA-free Protease Inhibitor [Complete Mini, Roche]. The sample was spun for 6 min at 4°C and 249xg. The cell pellet was then resuspended again in 2 ml of lysis buffer (10 mM Tris

[Thermo Fisher Scientific] at pH=8.0, 10 mM NaCl [VWR BDH Prolabo], 1x EDTA-free Protease Inhibitor [Complete Mini, Roche] and 0.2% Tergitol solution [Type NP-40, Sigma-Aldrich]. The sample was incubated for 10 min on ice. To recover the nuclei, the sample was spun for 5 min at 4°C and 1,083xg. The supernatant was removed and the nuclei resuspended in 2 ml of nuclei lysis buffer (50 mM Tris [Thermo Fisher Scientific] at pH=8.1, 10 mM EDTA [Thermo Fisher Scientific], 1x EDTA-free Protease Inhibitor [Complete Mini, Roche] and 1% SDS [VWR BDH Prolabo]). The sample was then incubated for 10 min on ice. Finally, 2 ml of dilution buffer (20 mM Tris [Thermo Fisher Scientific] at pH=8.1, 2 mM EDTA [Thermo Fisher Scientific], 150 mM NaCl [VWR BDH Prolabo], 1x EDTA-free Protease Inhibitor [Complete Mini, Roche], 1% Triton X-100 [Sigma-Aldrich] and 0.01% SDS [VWR BDH Prolabo]) was added, and immediately subjected to sonication.

Sonication. The samples were aliquoted to a final volume of 300 µl (~2–5x10^6 cells) per 1.5 ml tube, i.e. four and six aliquots for FAIRE-chip and FAIRE-seq experiments, respectively. Chromatin was subjected to sonication cycles of 30 sec at high pulse (200 W) followed by 30 sec of rest using the Bioruptor UCD-200 [Diagenode]. The sonication time varied across experiments and is reported in **Table 2-1**. A temperature of ~4°C was maintained. For optimisation experiments (**Table 2-1**; discussed in **Section 3.2**), aliquots of 2 µl were taken after respective sonication cycles to monitor sonication efficiency. Prior to analysis with a 2100 Bioanalyzer [Agilent Technologies], DNA-protein cross-links were reversed (see next paragraph). Finally, the lysate was cleared of cellular debris by spinning for 5 min at 4°C and 15,000xg, and the supernatant transferred into a new tube.

Reverse cross-linking. For optimisation experiments, samples were incubated for 6 hr at 65°C in a thermocycler [Thermomixer 5436, Eppendorf] prior to analysis using a 2100 Bioanalyzer [Agilent Technologies].

Analysis of DNA fragment length. For optimisation experiments, 1 µl of sample from each time point was applied to a 2100 Bioanalyzer [Agilent Technologies], using a DNA 1000 Chip or DNA 7500 Chip [Agilent Technologies] according to the manufacturer's protocol. Data was analysed using the software 2100 Expert [Agilent Technologies]. Experiments were performed with biological and technical replicates.

Phenol-chloroform extraction. An equal volume (300 µl) of phenol-chloroform-isoamyl alcohol (25:24:1) [Sigma-Aldrich] saturated with 10 mM Tris at pH=8.0 and 1 mM EDTA was added to the lysate. The mixture was vortexed, centrifuged for 5 min at 4°C and 12,000xg, and the aqueous phase transferred to a new tube. Then, 300 µl of TE buffer (10 mM Tris-HCl [Sigma-Aldrich] at pH=7.5 and

1 mM EDTA [Sigma-Aldrich]) was added to the organic phase, vortexed and centrifuged for 5 min at 4°C and 12,000xg. The aqueous phase was extracted and combined with the first extraction. To remove residual protein, an additional round of extraction was performed by adding 300 µl of phenol-chloroform-isoamyl alcohol to the combined aqueous fraction, followed by thorough mixing, centrifugation and retention of the aqueous phase. Then, 400 µl of chloroform-isoamyl alcohol (24:1) [Sigma-Aldrich] was added. The tube was vortexed, the two phases separated by centrifugation for 5 min at 4°C and 12,000xg, and the aqueous phase (400 µl) retained.

DNA precipitation. One-tenth volume (40 µl) of 3 M sodium acetate [VWR BDH Prolabo] at pH=5.2 and 1 µl of 20 mg/ml glycogen [Roche] were added to the mixture, and the tube was mixed by inverting. Two volumes (800 µl) of 95% ethanol [VWR BDH Prolabo] were added to the mix by inverting, and the reaction was incubated at 4°C overnight. Precipitated DNA was collected for 30 min at 4°C and 15,000xg, and the pellet washed with 500 µl of 70% ice-cold ethanol [VWR BDH Prolabo] by centrifugation for 10 min at 4°C and 15,000xg. The pellet was dried for ~10 min at RT in a SpeedVac [Concentrator 5301, Eppendorf], and resuspended in 25 µl of 10 mM Tris-HCl [Sigma-Aldrich] at pH=7.5. The aliquots were combined to form aliquots of 50 µl each. Next, 1 µl of 200 µg/ml RNase A [ICN Biomedicals] was added to the mix and incubated for 1 hr at 37°C. Finally, DNA was purified using the MinElute PCR Kit [Qiagen] according to the manufacturer's protocol. DNA was eluted in 2x10 µl of EB Buffer.

Table 2-1. Overview of experimental parameters applied in FAIRE experiments.

Parameter	Optimisation			FAIRE-chip		FAIRE-seq
	FAIRE	*FAIRE*	*Reference*	*FAIRE*	*Reference*	
Cell number	20x10^6	20x10^6	5x10^6	20x10^6	20x10^6	15x10^6
Fixation	5 min	5, 8, 12 min	n/a	8, 12 min	n/a	12 min
Sonication	9, 12, 14 min	12 min	4, 6, 8, 9 min	12 min	9 min	12 min
Cell type(s)	CHRF-288-11			CHRF-288-11, K562		CHRF-288-11, MKs, EBs, MOs

2.4. Detection and analysis using DNA tiling microarrays

Sample labelling. Precipitated DNA recovered from cross-linked cells ('FAIRE sample') and uncross-linked cells ('reference sample') was labelled with Cy5 and Cy3 dye, respectively. Sample labelling was performed using the Dual-Color DNA Labeling Kit [all components, Roche NimbleGen] according to the manufacturer's protocol (Roche NimbleGen Arrays User's Guide, ChIP-chip Analysis v4.1). Both Cy3- and Cy5-Random Nonamer Primers were diluted in 462 µl of Random Primer Buffer supplemented with β-Mercaptoethanol [Sigma-Aldrich]. The FAIRE and reference samples were placed in separate 200 µl thin-walled PCR tubes [Ambion]. For FAIRE and reference samples, each 80 µl reaction contained 40 µl of purified DNA (**Table 2-2 A**) and 40 µl of diluted Cy5- and Cy3-Random Nonamers, respectively. The samples were heat denatured for 10 min at 98°C in a thermocycler [Thermomixer 5436, Eppendorf], immediately quick-chilled in an ice-water bath and incubated for 2 min. Next, 20 µl of the dNTP/Klenow Master Mix was added on ice to each of the denatured samples to a final volume of 100 µl. The solution was mixed well by pipetting up and down 10 times. Subsequently, the mix was collected by centrifugation and incubated for 2 hr at 37°C in a thermocycler with heated lid, protected from light. The reaction was stopped by addition of 10 µl of 0.5 M EDTA. Then, 11.5 µl of 5 M NaCl was added to each sample. The mix was briefly vortexed, spun down and the entire contents transferred to a 1.5 ml tube containing 110 µl of 100% isopropanol. After vortexing, the mix was incubated for 10 min at RT, protected from light. The precipitate was collected by centrifugation for 10 min at 12,000xg. The supernatant was removed and the pellet rinsed with 500 µl of 80% ice-cold ethanol by centrifugation for 2 min at 12,000xg. The supernatant was removed and the pellet dried for ~10 min at 30°C in a SpeedVac [Concentrator 5301, Eppendorf], protected from light. The dried pellet was rehydrated in 25 µl of nuclease-free water. Finally, the sample was vortexed several times until the pellet was completely rehydrated. After labelling, the FAIRE and reference samples were quantitated using a NanoDrop spectrophotometer [ND-1000, Labtech]. Samples were analysed in the Nucleic Acid module, DNA-50 mode using the software ND-1000 v3.5.2 (**Table 2-2 B**).

Table 2-2. DNA quantity of FAIRE and reference samples before and after labelling with cyanine dyes.

Cell line	(A) Before labelling			(B) After labelling		
	Reference	*8 min*	*12 min*	*Reference*	*8 min*	*12 min*
CHRF-288-11	11.25 µg	0.79 µg	0.38 µg	21.72 µg	8.31 µg	6.20 µg
K562	11.75 µg	0.84 µg	0.59 µg	33.11 µg	11.62 µg	15.55 µg

For the array hybridisation reaction, 6 µg of both FAIRE and reference DNA were required (**Table 2-2 B**). Based on the determined concentration, the respective volumes of the FAIRE and reference samples were calculated and combined in a 1.5 ml tube. The content was dried for ~15 min at 30°C in a SpeedVac [Concentrator 5301, Eppendorf], protected from light.

Array hybridisation. The MAUI Hybridization System [BioMicro Systems] was set to 42°C and the temperature was allowed to stabilise for 3 hr. Hybridisation on 385K arrays was performed using the Hybridization Kit, Mixer X1 and Precision Mixer Alignment Tool (PMAT) [Roche NimbleGen] according to the manufacturer's protocol. The dried sample pellet was resuspended in 5 µl of DNase- and RNase-free water [Gibco]. The mix was vortexed well and spun down. Next, 13 µl of the Hybridization Solution Master Mix [Roche NimbleGen] was added to 5 µl of resuspended sample. The solution was vortexed, spun down and incubated for 5 min at 95°C, protected from light. The sample was incubated for 5 min at 42°C in the hybridisation system. The Mixer X1 was assembled with the 385K-feature slide using the PMAT. The assembly was placed in the slide bay of the hybridisation system, and 16 µl of the hybridisation mix was loaded into the fill port of the slide/mixer assembly. Finally, the sample was hybridised to the array for 20 hr at 42°C in mixing mode B.

Array washing. Arrays were washed using the Wash Buffer Kit [Roche NimbleGen] according to the manufacturer's protocol. To process the protocol without interruption and ensure high quality data, only one slide was washed at a time. After hybridisation, the mixer-slide assembly was removed from the hybridisation system and loaded in the Mixer Disassembly Tool that was immersed in 250 ml of warm Wash I (42°C). With the mixer-slide assembly submerged, the mixer was carefully peeled off the slide. The mixer was discarded and the slide quickly removed from the Mixer Disassembly Tool. The slide was gently agitated for 15 sec to quickly remove the hybridisation buffer. Subsequently, the slide was transferred into a slide container containing Wash I and agitated vigorously for 15 sec. During all wash steps, the microarray area of the slide was submerged at all times and not allowed to dry between wash steps. The slide was washed for an additional 2 min in Wash I with vigorous constant agitation. The slide was transferred to Wash II and washed for 1 min and subsequently to Wash III for 15 sec with vigorous, constant agitation. The slide was removed from Wash III and immediately dried for 2 min in a microarray centrifuge [Spectrafuge Mini, Labnet]. Residual moisture was removed by blow-drying the edges. Immediately after washing, the slide was scanned.

Array scanning. Arrays were scanned using a DNA Microarray Scanner [Agilent Technologies] and the software Scan Control v8.3.1 [Agilent Technologies] according to the manufacturer's protocol. Arrays were scanned with wavelengths of 532 nm and 635 nm for Cy3 and Cy5, respectively, PMT power of 100% and a pixel size of 5 µm.

Data processing. Experimental data were analysed using the software NimbleScan v2.5 [Roche NimbleGen]. The two-channel raw signal intensities were scaled between channels by subtracting the Tukey bi-weight mean for the $\log_2$-ratio values for all features from each $\log_2$-ratio value. This scaling procedure accounts for differences in the signal intensities of the dyes by centring the data on zero. Since the experimental setup provided a two-colour array with the reference sample on the array, normalisation of data was not performed.

Peak calling. In order to find peaks in the scaled $\log_2$-ratio data, a sliding window was moved across each chromosome probe by probe. Within this window, each probe was tested if its $\log_2$-ratio was above a certain cut-off value. A peak was registered when the number of qualifying probes was above a set probe number within the sliding window. The genetic position of the identified peak was set from the start position of the first qualifying probe to the end position of the last qualifying probe. For each chromosome, the $\log_2$-ratio cut-off value was calculated as the percentage of a hypothetical maximum (P_{max}=arithmetic mean + 6x standard deviation). The peak finding process was repeated using a series of $\log_2$-ratio cut-off values from P_{start} to P_{end}. By using a hypothetical maximum rather than the overall maximum of the $\log_2$-ratios, the effects of outliers can be minimised (Lucas et al., 2007). The following settings for the peak finding analysis were applied: sliding window: 300 bp; min. probes>cut-off in peak=4; all probes in peak>cut-off=2; P_{start}=90%, P_{end}=15%, P_{step}=0.5, number of steps: 100. In all array experiments, the signal $\log_2$-ratio between the FAIRE and the reference sample showed a normal distribution, with an enrichment of FAIRE signal at the right end of the distribution.

Data visualisation. Log_2-ratio and peak data sets were displayed as UCSC Genome Browser (http://genome.ucsc.edu/) custom tracks.

Data availability. The FAIRE microarray data sets are available online in the Gene Expression Omnibus (GEO) database under accession number GSE25716.

2.5. Detection and analysis using high-throughput next-generation sequencing technology

Library preparation and sequencing. FAIRE DNA was processed following the Illumina paired-end library generation protocol. Genomic libraries derived from MO extractions and CHRF-288-11 cells were sequenced on Illumina HiSeq 2000 with 50 bp and 75 bp paired-end reads, respectively. Libraries derived from EB and MK cultures were sequenced on Illumina GAIIx with 54 bp paired-end reads. All FAIRE sample libraries were prepared and sequenced at the Wellcome Trust Sanger Institute by the library-making and sequencing core groups, respectively.

Sequence data processing. Raw sequence reads were aligned to the human reference sequence (NCBI build 37) using the algorithm Stampy (Lunter & Goodson, 2011). Reads were realigned around known insertions and deletions (The 1000 Genomes Project Consortium, 2010), followed by base quality recalibration using the Genome Analysis Toolkit (GATK) (McKenna et al., 2010). Duplicates were flagged using the software Picard (http://picard.sourceforge.net/) and excluded from subsequent analyses. The raw sequence files from two independent FAIRE experiments in K562 cells were obtained from the ENCODE Project (GEO accession number GSM864361), and remapped as described above. An overview of the sequencing statistics is provided in **Table 2-3**.

Table 2-3. Overview of sequencing statistics. Summary of the total number of DNA sequence reads mapped to the human reference sequence (NCBI build 37), listed for each sample. The following exclusion criteria were applied to sequence reads: mapped with quality below 30; duplicated; and mapped to mitochondrial DNA and unplaced chromosomes. For paired-end data sets, reads not properly paired and paired farther than 1 kb were also excluded.

Cell type	Samples	Number of mapped reads		Percentage of total reads
		Before filtering	*After filtering*	
EB	Indiv. A	40,823,840	29,202,677	71.5%
	Indiv. B	48,350,482	40,542,641	83.9%
MK	Indiv. B	43,292,704	36,376,461	84.0%
	Indiv. C	48,365,225	35,697,504	73.8%
MO	Indiv. D	250,497,524	201,993,051	80.6%
	Indiv. E	225,495,458	184,262,652	81.7%
K562	Repl. 1	59,913,440	43,401,021	72.4%
	Repl. 2	59,741,112	40,694,874	68.1%
CHRF	Repl. 1	58,867,716	52,438,908	89.1%
	Repl. 2	60,088,050	53,769,761	89.5%

Peak calling and normalisation. Regions of enrichment (peaks) were determined using the software F-Seq v1.84 (Boyle, Guinney, et al., 2008). A feature length of L=600 bp and two different standard deviation thresholds of T=6.0 (‘moderate’) and T=8.0 (‘stringent’) over the mean across a local background were applied. In order to reduce false positive peak calls, regions of collapsed repeats were removed, as described in Pickrell et al., 2011, applying a threshold of 0.1% (http://eqtl.uchicago.edu/Masking). Four equally spaced bins (in $\log_{10}$-transformed peak score units) were defined between the 1st and 99th percentile of the peak score distribution. For comparison of cell type-specific chromatin profiles, all read fragments were merged into one data set for each cell type. Then, peaks were called as described. For the K562 single-end sequencing data set, the mode of the peak width distribution was adjusted to the mean of the modes across all non-K562 cell types. **Tables 2-4** and **2-5** give an overview of the final peak data sets. ChIP-seq data sets in primary MKs were obtained from Tijssen et al., 2011 (GEO accession number GSE24674). The peak coordinates were remapped to hg19 (minimum ratio of bases that must remap: 0.95) using the Lift-Over tool v1.0.3 of the web-based platform Galaxy (http://main.g2.bx.psu.edu/).

Table 2-4. Overview of FAIRE peak statistics.

Cell type	Samples	(A) Number of FAIRE peaks		(B) Number of merged FAIRE peaks	
		Moderate (T=6.0)	*Stringent (T=8.0)*	*Moderate (T=6.0)*	*Stringent (T=8.0)*
EB	Indiv. A	67,395	22,754	96,741	37,252
	Indiv. B	82,761	31,553		
MK	Indiv. B	86,270	28,543	148,124	49,364
	Indiv. C	81,622	26,845		
MO	Indiv. D	55,512	16,399	101,034	34,135
	Indiv. E	43,664	17,669		
K562	Repl. 1	84,356	41,638	122,463	67,832
	Repl. 2	84,991	43,981		
CHRF	Repl. 1	128,184	60,970	222,424	109,610
	Repl. 2	123,053	63,538		

Table 2-5. Overview of the number of FAIRE peaks for each intensity bin. Shown are the merged peak data sets, called with the stringent F-Seq threshold (T=8.0). Peaks were filtered based on the criteria described above, and did not count to the total number of peaks.

	EB	MK	MO	K562	CHRF
Bin 1	28,415	42,206	21,465	33,008	80,482
Bin 2	5,235	4,407	7,745	20,612	17,039
Bin 3	2,563	1,860	3,565	9,974	8,357
Bin 4	1,039	891	1,360	4,238	3,732
Total number	37,252	49,364	34,135	67,832	109,610
Total filtered	*762*	*1,008*	*698*	*1,694*	*2,238*

Data visualisation. The coverage profile on the combined data was created using the R packages ShortRead (Morgan et al., 2009) and rtracklayer (Lawrence, Gentleman, et al., 2009). Coverage and peak data sets were displayed as UCSC Genome Browser custom tracks.

Data availability. All FAIRE sequencing data sets are available online in the GEO database under accession number GSE37916.

2.6. Annotation of NDRs and statistical analyses

Hierarchical cluster analysis and bootstrapping. First, a union set of all peaks across all samples was created. Next, a vector of binary values for each sample *s* was defined, whereby the length of the vector is given by the total number of peaks in the union set and is therefore the same for all samples. Position *i* in the vector for sample *s* was set to a value of one, if the peak *i* in the union peak set overlaps with a peak in sample *s*. If there was no overlap with a peak in sample *s*, position *i* was set to zero. From these vectors, bin-specific vectors were constructed based on the binning scheme described in **Section 2.5**. For each bin and sample, a vector was defined where all entries with a peak score not between the lower and upper peak scores defined for that bin, were set to zero. Then, the R package Pvclust (Suzuki & Shimodaira, 2006) was used to perform a bootstrapped hierarchical cluster analysis of the samples based on these binary vectors, using the 'binary' distance measure and the 'complete' method for defining the clusters (Suzuki & Shimodaira, 2006). Here, 1,000 bootstrap samples were applied. All analyses were carried out in the R/Bioconductor environment.

Annotation of NDRs using GREAT. The ontology of genes flanking FAIRE peaks was analysed using the Genomic Regions Enrichment of Annotations Tool (GREAT) v1.8.2 (McLean et al., 2010) with the following parameters: association rule: single nearest gene; 1 Mb maximal extension; curated regulatory domains included. The genomic distances between FAIRE peaks and transcription start sites were exported from GREAT.

Enrichment analysis using bootstrapped quantile distributions. The association analysis for the eight quantitative traits was performed by imputation from the Phase II HapMap panel (The International HapMap Consortium, 2007). To improve the coverage, for each Phase II HapMap SNP it was determined which 1000 Genomes SNPs (interim phase I release of June 2011) within a distance of 50 kb had an r^2 of at least 0.95 with the imputed Phase II HapMap SNPs. For each trait, the *P*-value of the Phase II HapMap SNP from the meta-analysis to the 1000 Genomes SNP was assigned. To prevent chance inflation from LD and to obtain confidence estimates, 500 bootstrap samples were created from this set of 1000 Genomes SNPs by randomly removing SNPs until the genomic distance between remaining SNPs was at least 50 kb. For each trait, the mean genomic inflation at the 0.005 quantile was inferred from these bootstrap samples. This genomic inflation factor provides a baseline genomic inflation factor. It should be noted that corrections for population stratification and covariates have already been performed in the original meta-analyses (Gieger et al., 2011; van der Harst et al., 2012). Using the same pruning approach, for each of the three peak sets ('merged', 'intersected' and 'cell type-specific'), 500 bootstrap samples were generated using only those 1000 Genomes SNPs that are located in a peak from the respective peak set. Finally, the relative genomic inflation factors (as reported in **Figure 4-7**) were calculated as the ratio between the baseline genomic inflation factor and the genomic inflation factor calculated for SNPs located in peaks from a given peak set. The 0.005 quantile provides a trade-off between highlighting differences in enrichment across different cell types and reducing uncertainty in the estimates of the relative genomic inflation factors. All analyses were carried out in the R/Bioconductor environment.

Canonical pathway analysis. Genes were subjected to the core analysis module of the software Ingenuity IPA v12402621 (http://www.ingenuity.com), and analysed using the following parameters: reference set: Ingenuity Knowledge Base (genes only); relationship to include: direct and indirect; filter: only molecules and/or relationships where species=human and confidence=experimentally observed. Benjamini-Hochberg multiple test-corrected *P*-values are reported.

Overlap of NDRs with association loci and significance analysis. For each association locus, candidate functional SNPs were selected by identifying all biallelic SNPs with an r^2>0.8 and within 1 Mb of the sentinel SNP in the European samples of the 1000 Genomes Project data set (interim phase I release of June 2011). To establish whether the association of a locus could potentially be of regulatory origin, it was determined if at least one candidate functional SNP overlapped with a FAIRE peak. Since this analysis is sensitive to the number of peaks, the overlap was carried out for successively increasing number of peaks by considering peaks with decreasing peak height. As more peaks are considered, the chance of finding an overlap increases. Therefore, the significance of the findings was estimated by resampling. Of both 68 and 75 loci associated with platelet and erythrocyte phenotypes, respectively, 100,000 sets were drawn from the same SNPs onto which the GWA data was imputed (~$2.5x10^6$ SNPs from Phase II HapMap), while preserving the distribution of allele frequencies and the number of loci that overlapped with FAIRE peaks. This was repeated (100,000 permutations) for each successive increase in the number of FAIRE peaks. All analyses were carried out in the R/Bioconductor environment.

2.7. Gene expression analysis during *in vitro* differentiation of cord blood-derived HPCs

Experiments and statistical analyses were performed as described in Gieger et al., 2011. Briefly, MKs and EBs were differentiated from cord blood-derived HPCs as described in **Section 2.2**. Time points were taken at days 3, 5, 7, 9, 10 and 12. Whole-genome gene expression levels were measured using Illumina HumanWG-6 v3 Expression BeadChips. Expressed probes were selected based on stringent thresholds and the slope of expression was determined using standard linear regression. The single closest ENSEMBL transcript (release 64) with an HGNC symbol was assigned to every FAIRE peak.

2.8. H3K4me1 and H3K4me3 ChIP-seq

MKs and EBs were differentiated from cord blood-derived HPCs as described in **Section 2.2**. ChIP assays were performed as described in Forsberg et al., 2000, using rabbit polyclonal antibodies against H3K4me1 [ab8895, Abcam] and H3K4me3 [07-473, Millipore]. Chromatin-immunoprecipitated DNA was sequenced on Illumina GAII with 54 bp single-end reads. Sequence reads were aligned using the algorithm BWA (Li & Durbin, 2009). Areas of enrichment were determined using the slice function of the R package IRanges (http://bioconductor.org/packages/2.10/bioc/html/IRanges.html).

2.9. Sanger sequencing of selected NDRs

Ethics statement. All human subjects were recruited with appropriate informed consent in Cambridgeshire and enrolled in the Cambridge BioResource (http://www.cambridge-bioresource.org.uk/).

Capillary sequencing. DNA samples from a total of 643 individuals of Northern European ancestry were subjected to capillary DNA sequencing of the targeted locus at chromosome 7q22.3. Sequencing primer pairs for two sequence-tagged sites were designed using the web-based tool Primer3 (Rozen & Skaletsky, 2000), and are reported in **Table 2-6**.

Table 2-6. Sanger sequencing primer pairs for two sequence-tagged sites at chr7q22.3. Genomic coordinates of the PCR products were based on the human reference genome, build hg18.

Forward primer sequence	Reverse primer sequence	Genomic position	Amplicon
TGGAAAATTACAAAAGTCCCAAA	GAGAAAGGATCATGAGGGAGAA	chr7:106,159,328–106,159,998	671 bp
ACAAAAGTCCCAAAATTTCACA	GAGAAAGGATCATGAGGGAGA	chr7:106,159,337–106,159,998	662 bp

PCR products were applied to bi-directional sequencing using Big Dye chemistry on 3730 DNA sequencers [Applied Biosystems]. DNA amplification by PCR and capillary sequencing were performed at the Wellcome Trust Sanger Institute by members of the ExoSeq/ExoCan facility. Details of the sequencing protocol are described online (http://www.sanger.ac.uk/resources/downloads/human/exoseq.html).

Analysis. Pre-processed sequence traces were analysed using the semi-automated analysis software ExoTrace (http://www.sanger.ac.uk/resources/downloads/human/exoseq.html), developed at the Wellcome Trust Sanger Institute. Results of the SNP calling were displayed in a specific implementation of GAP4, which is part of the Staden Sequence Analysis Package (http://staden.sourceforge.net/). Potential SNP positions were indicated by the software and then reviewed manually.

2.10. Transcription factor binding site prediction

TRAP. Transcription factor binding sites were predicted using the transcription factor affinity prediction (TRAP) method (Thomas-Chollier et al., 2011) (sTRAP tool) with the following parameters: matrix: TRANSFAC v2010.1 (vertebrates); background: human promoters; multiple test correction: Benjamini-Hochberg.

MathInspector. In **Section 5.2**, transcription factor binding sites were predicted using the software MatInspector v8.01 (Cartharius et al., 2005). The following parameters were applied: matrix group: vertebrates; core=1.00; matrix=optimised+0.02; tissue: haematopoietic system. In **Section 6.2**, an updated MathInspector library version was used (v8.3). The same parameters were applied, except for the matrix group: general core promoter elements and vertebrates. In addition, a restriction on tissue type was omitted.

2.11. Electrophoretic mobility shift assay (EMSA)

Extraction of nuclear protein. Non-denatured, active nuclear proteins were purified from 5x10^6 CHRF-288-11 cells with the NE-PER Nuclear and Cytoplasmic Extraction Reagents [Thermo Fisher Scientific] according to the manufacturer's protocol. The cell pellets were removed from -80°C storage and incubated for 5 min on ice. Cytoplasmic Extraction Reagent I (CER I) at a volume of 200 µl was added to the ~20 µl of packed cell volume. The sample was vigorously vortexed for 15 sec on the highest setting to fully suspend the cell pellet, and subsequently incubated for 10 min on ice. Next, 11 µl of Cytoplasmic Extraction Reagent II (CER II) was added, and the tube vortexed for 5 sec on the highest setting. The sample was incubated for 1 min on ice. The sample was then vortexed for 5 sec on the highest setting, and centrifuged for 5 min at 4°C and 16,000xg. After cell membrane disruption and release of cytoplasmic contents, the supernatant (cytoplasmic extract) was immediately transferred to a clean pre-chilled tube. The pellet, which contains intact nuclei, was resuspended in 100 µl of ice-cold Nuclear Extraction Reagent (NER). The sample was vortexed for 15 sec on the highest setting. The sample was placed on ice whilst continuing to vortex for 15 sec every 10 min, for a total of 40 min. Then, the sample was centrifuged for 10 min at 4°C and 16,000xg. The supernatant (nuclear extract) was immediately transferred to a clean pre-chilled tube. Extracts obtained with this protocol generally have less than 10% contamination between nuclear and cytoplasmic fractions. For every EMSA, fresh nuclear protein was prepared.

Probe design. Oligonucleotides were designed based on the genomic sequence surrounding each candidate functional SNP (**Table 2-7**). Oligonucleotides were prepared with a biotin tag at the 5'-end and without modification ('competitor'), for both alleles of the candidate SNP. In addition, unlabelled complementary strands were prepared for both alleles. All oligonucleotides (desalting purification) were provided by Sigma-Aldrich.

Table 2-7. EMSA probes.

Candidate SNP	Sequence of probe 5'→3'
rs342293	AGCCCTGTGGTTTTAATTAT[C/G]TTGAGGTTCAGGCTCA
chr1:145,507,646 (5'-UTR *RBM8A*)	AGTGTCTGAGCGGCACAGAC[G/A]AGATCTCGATCGAAGG
chr1:145,507,765 (intron *RBM8A*)	AGACGGCTGGTGGGAAGC[G/C]GGGAAGGTGCGAGAGAAGG
rs1006409	TTCCTTCTTTTCCTTTT[A/G]TGGTATGCATAGATATCA
rs1107479	CTGCCAAGGACGTCA[C/T]AGGCAGATGGAAGGAAGCTT
rs11731274	TGGCACACGCTGGTGGC[T/G]TTCCCCGGGCTCTCTGCT
rs11734099	GAGCTCCCTCCCTGGCCT[G/A]CCTGGCACACGCTGGTGGCT

Candidate SNP	Sequence of probe 5’→3’
rs17192586	AGATTCTTAGGAGTAAC[G/A]GCTGACATTCACCATATT
rs2015599	AATGAATTCTAACTCACT[G/A]CAAGTACTACAGTGTTCCT
rs2038479	ACTGCTATTTTCATTTTAT[C/A]GATGGAATACTTTGAAG
rs2038480	CAAGCGTGTGTTAAGAATA[A/T]GTATATAAAATGTGTTTT
rs214060	AGACAACCGGCAGCTCTAA[C/T]GAAAATATTGGAGACACT
rs2735816	TTTGCCCTGCACTGAGCA[G/C]AGAGCATCTGAAATGTGGA
rs3214051	CGGGGGTGGTGACAAG[G/A]ACTAAAGGGTAAGAATTTA
rs3804749	GCTGCAGGCTGCAAACAGG[C/T]GAAACAGGAAGAGAGA
rs4148450	AACAGGGAACTTGACATC[C/T]GCCCAGACCATCAGTCAAT
rs55905547	AATCTCAGTGTTGTGGGCC[A/G]TAGCGTCCTCACCACA
rs6771416	AACTTCCAGAGACAGCTA[G/A]ATGGGGCAGTGAGTCCAGT
rs7618405	CTTTTGGGAGGCCAC[C/A]ATGAGTTAGCACTCTTTTCT

The complementary strands were annealed using a standard protocol (http://www.piercenet.com/files/TR0045-Anneal-oligos.pdf), consisting of denaturation of the complementary strands to remove any secondary structure and hybridisation of the strands. Annealing occurs most efficiently when the temperature is slowly decreased after denaturation. Complementary oligonucleotides of 100 μM were combined at an equal molar ratio to a final concentration of 1 μM in Buffer EB [QIAGEN]. To anneal the strands, the sample was incubated in a thermocyler [PTC-225 Peltier Thermal Cycler, MJ Research] using the following programme: 5 min at 95°C; 1 min at 94–25°C, with a decrease of 1°C per cycle (70 cycles); hold at 4°C. Double-stranded probes were quantitated in triplicates using a NanoDrop spectrophotometer [ND-1000, Labtech], and analysed in the Nucleic Acid module, DNA-50 mode using the software ND-1000 v3.5.2.

Binding reaction. Gel mobility shift assays were performed with the LightShift Chemiluminescent EMSA Kit [all components, Thermo Fisher Scientific] according to the manufacturer’s instructions. Each 20 μl binding reaction contained 1x binding buffer, 75 ng/μl poly(dI/dC), 2.5% glycerol, 0.05% NP-40, 87.5 mM KCl and 6.25 mM $MgCl_2$. For each reaction, 2.5 μl of freshly prepared nuclear protein was used. Biotin-labelled DNA containing the candidate SNP of interest was added in a final amount of 20 fmol. For competition assays, unlabelled probes were added in final amounts of 2 or 4 pmol, representing a 100- or 200-fold molar excess over the labelled probes, respectively (**Table 2-8**). An overview of the incubation times is provided in **Table 2-8**.

Electrophoresis. The binding reaction was then subjected to gel electrophoresis on a native polyacrylamide gel [Novex 6% DNA Retardation Gel, Invitrogen] according to the manufacturer’s protocol. The wells of the gel were flushed, and the gel was pre-electrophoresed for 30 min in ice-cold

0.5x Novex TBE Running Buffer [Invitrogen], applying 100 V using the XCell SureLock Electrophoresis Mini-Cell [Invitrogen]. Then, the wells were flushed and loaded with 19 µl of each sample. Samples were electrophoresed for 75 min at 100 V.

Electrophoretic transfer to nylon membrane. After electrophoresis, the gel was extracted from the cassette and carefully transferred onto 0.8 mm blotting paper [Whatman] or filter paper [Mini Trans-Blot, Bio-Rad]. The nylon membrane [Biodyne B, Thermo Fisher Scientific] was soaked in 0.5x Novex TBE Running Buffer [Invitrogen] for at least 10 min. Both blotting/filter paper and blotting pads were briefly soaked prior to use. The blot module was assembled according to the manufacturer's protocol. In brief, the order of assembly was as followed: cathode core, blotting pad, filter paper, gel, nylon membrane, filter paper and blotting pad. The transfer was performed for 1 hr at 30 V (360–270 mA) using a Mini Trans-Blot Cell [Bio-Rad]. Immediately after transfer, the membrane was exposed for 45–60 sec to UV-light (254 nm) [Stratalinker UV Crosslinker 2400, Stratagene], applying 120 mJ/cm^2 in the auto cross-link mode, to cross-link the transferred DNA to the nylon membrane.

Detection of biotin-labelled DNA by chemiluminescence. The biotin-labelled DNA was detected using the Chemiluminescent Nucleic Acid Detection Module [all components, Thermo Fisher Scientific]. The Blocking Buffer and 4x Wash Buffer were gently warmed to 37–50°C in a water bath until all particulate was dissolved. After UV-cross-linking, the membrane was immediately blocked by submerging in 20 ml of Blocking Buffer. The membrane was incubated for 15 min with gentle shaking. The buffer was decanted from the membrane and replaced with 20 ml of Blocking Buffer supplemented with 66.7 µl of Stabilized Streptavidin-Horseradish Peroxidase Conjugate (1:300 dilution). The membrane was incubated for 15 min with gentle shaking, subsequently transferred to a new container and washed four times for 5 min each in 20 ml of 1x Wash Buffer with gentle shaking. The membrane was transferred to a new container and 30 ml of ice-cold Substrate Equilibration Buffer was added. The membrane was incubated for 5 min with gentle shaking. Next, the Chemiluminescent Substrate Working Solution was prepared by adding 5 ml of Luminol/Enhancer Solution to 5 ml of Stable Peroxide Solution, protected from light. The membrane was removed from the Substrate Equilibration Buffer, excess buffer removed, and then placed onto a clean sheet of plastic wrap. The Substrate Working Solution was poured onto the membrane so that it completely covered the membrane. Then, the membrane was incubated for 5 min, protected from light. After incubation, the membrane was removed from excess buffer and covered with plastic foil. Finally, the membrane was placed in a film cassette and exposed to X-ray film [CL-XPosure Film, Thermo Fisher Scientific] for 5–10 min. The film was developed using an X-ray film processor [Compact X4, Xograph] according to manufacturer's instructions.

Quantification of signal density. The blots were quantified by measuring the signal density of the probes competed with the respective unspecific competitor using the software ImageJ v1.45 (http://rsbweb.nih.gov/ij/). The mean density ratio represents the ratio of the measured density of the stronger and the weaker band (0.15 x 1.50 rectangular area centred on each band).

Supershift. For supershift experiments, the antibody was added to the reaction mix at the end, prior to incubation. **Table 2-8** gives an overview of the conditions used in the experiments.

Table 2-8. Overview of the experimental setup for EMSA and supershift experiments. Changes in experimental parameters were based on optimisation experiments.

Candidate SNP	Incubation of binding reaction	Additional agent	Molar excess of competitor	Antibody for supershift experiments
rs342293	120 min at RT	n/a	200-fold	4 µl EVI1 [sc-8707 X, Santa Cruz Biotechnology (SCB)]
	60 min at RT	n/a	200-fold	2 µl GATA1 [ab28839, Abcam] 4 µl RUNX1 [sc-28679 X, SCB]
chr1:145,507,646 (5'-UTR *RBM8A*)	45 min at RT	n/a	100-fold	2 µl EVI1 [sc-8707 X, SCB]
chr1:145,507,765 (intron *RBM8A*)	120 min at RT	0.1 mM EDTA	100-fold	2 µl MZF1 [sc-46179 X, sc-66991 X, SCB] 2 µl RBPJ [sc-28713 X, SCB]
All other (Table 2-7)	45 min at RT	n/a	100- or 200-fold (Figure 4-11)	n/a

2.12. Expression QTL analysis

Data sets. Published gene expression profiling and genotypic data sets were obtained from different sources depending on the cell type studied. Details regarding experimental protocols and data processing can be found in the respective references (**Table 2-9**).

Table 2-9. Genotyping and gene expression platforms used for eQTL analyses. Even though different versions of Illumina platforms were used, the probe for *PIK3CG* was the same across all chips (probe-ID: ILMN_1770433).

Cell type/ tissue	Genotyping	Gene expression profiling	Reference
Platelets	Applied Biosystems TaqMan	Illumina HumanWG-6 v2	Sivapalaratnam et al., *in preparation*
Macrophages	Illumina Human 1.2M-Custom / Illumina Human 670-Quad-Custom	Illumina HumanRef-8 v3	Rotival et al., 2011
Monocytes			
LCLs	Illumina Human 1M-Duo	Illumina HumanHT-12 v3	Nica et al., 2011
Adipose			
Skin			

Analysis. Expression QTL analyses with rs342293 and its proxy SNPs were performed with the software Genevar (Gene Expression Variation) using a window of ±1 Mb centred on the SNP (Yang et al., 2010). The strength of the relationship between alleles and gene expression intensities was estimated using Spearman's rank correlation and reported as nominal *P*-values.

2.13. Whole-genome gene expression profiling of *Pik3cg*$^{-/-}$ mice

Ethics statement. The study had ethical approval from the NHS Cambridgeshire Research Ethics Committee. The care and use of all mice in this study was carried out in accordance with the UK Home Office regulations under the Animals (Scientific Procedures) Act 1986.

***Pik3cg* knockout mice.** *Pik3cg* knockout mice were obtained from sources described in Sasaki et al., 2000, backcrossed onto the C57BL/6J Jax genetic background for eight generations (B6J;129-Pik3cg^{tm1Pngr}) and then maintained as a closed colony by intercrossing from within the colony (C57BL/6J Jax contribution: 99.6%). PCR genotyping was performed with the following primer pairs: 5'-TCA GGC TCG GAG ATT AGG TA, 5'-GCC CAA TCG GTG GTA GAA CT (wild type); 5'-GGA CAC GGC TTT GAT TAC AAT C, 5'-GGG GTG GGA TTA GAT AAA TG (mutant), as described in Sasaki et al., 2000.

Blood collection. Approximately 0.5 ml of whole blood from three *Pik3cg*$^{-/-}$ and three C57BL/6J Jax wild type mice (all females, age: 13–16 weeks, Mouse Breeders Diet [Lab Diets 5021-3]) was collected from terminally anesthetised mice via the retro-orbital sinus.

RNA extraction. Total RNA was extracted using the Mouse RiboPure-Blood RNA Isolation Kit [Ambion] according to the manufacturer's protocol. Total RNA was quantitated and quality-checked using a NanoDrop spectrophotometer [ND-1000, Labtech]. **Table 2-10 A** gives an overview of the RNA extraction yield.

Gene expression profiling. After extraction, 500 ng of total RNA was transformed into biotinylated cRNA using the TotalPrep RNA Amplification Kit [Ambion] according to the manufacturer's protocol. The protocol comprised three main steps: (1) reverse transcription of total RNA to synthesise full-length, first-strand cDNA with an oligo(dT) primer bearing a T7 promoter; (2) second-strand synthesis to convert the single-stranded cDNA into a double-stranded DNA template for *in vitro* transcription, followed by cDNA purification; and (3) *in vitro* transcription with T7 RNA Polymerase and biotin-UTP to generate multiple copies of biotinylated antisense RNA (cRNA) from the double-stranded cDNA templates. Following cRNA purification, the amplified and labelled cRNA was directly used for array hybridisation. **Table 2-10 B** gives an overview of the yield and quality of biotin-labelled cRNA.

Table 2-10. Assessment of quantity and quality of total RNA and biotin-labelled cRNA.

Mouse	(A) RNA extraction	(B) Labelled cRNA preparation			
	Yield	*Repl.*	*Yield*	*260/280*	*260/230*
Pik3cg$^{-/-}$-A	20.55 µg	1	16.43 µg	2.04	2.02
		2	16.00 µg	2.07	2.05
Pik3cg$^{-/-}$-B	17.73 µg	1	15.41 µg	2.07	1.99
		2	21.51 µg	2.06	2.03
Pik3cg$^{-/-}$-C	20.63 µg	1	18.38 µg	2.08	2.05
		2	16.42 µg	2.08	2.04
WT-A	11.65 µg	1	13.81 µg	2.07	1.96
		2	15.90 µg	2.05	1.99
WT-B	25.08 µg	1	13.07 µg	2.06	2.00
		2	17.39 µg	2.05	1.86
WT-C	13.10 µg	1	13.38 µg	2.05	1.97
		2	15.35 µg	2.05	2.02

For each sample, 750 ng of biotinylated cRNA was hybridised to Illumina MouseWG-6 v2 Expression BeadChips. The BeadChip contains 45,281 unique probes that target the NCBI Reference Sequence (RefSeq) database v22 (ftp://ftp.ncbi.nih.gov/refseq/release/), Mouse Exonic Evidence-Based Oligonucleotide (MEEBO) set (http://www.arrays.ucsf.edu/archive/meebo.html) and the exemplar protein-coding sequences described in the RIKEN FANTOM2 database (http://fantom2.gsc.riken.jp/).

After hybridisation, arrays were washed, detected and scanned on a BeadArray Reader [Illumina] according to the manufacturer's protocol.

Data processing. On the raw expression data, background subtraction, variance-stabilising transformation and quantile normalisation were performed across all samples with the R package lumi (Du et al., 2008). Technical replicates were averaged and the differentially expressed transcripts between wild type and knockout mice were identified by calculating the $\log_2$-fold changes of the averaged expression values. *P*-values were calculated by 1-way analysis of variance (ANOVA). All analyses were carried out in the R/Bioconductor environment.

Gene Ontology. Gene Ontology term enrichment analysis was performed using the web-based tool AmiGO v1.7 (Carbon et al., 2009) with the following parameters: gene expression fold-change cut-off: ±1.5; background: MGI; *P*-value cut-off: $1x10^{-5}$; minimum number of gene products: 10.

Data availability. The whole-genome gene expression data sets of *Pik3cg*$^{-/-}$ and wild type mice are available online in the GEO database under accession number GSE26111.

2.14. Protein-protein interaction network

Of the 220 differentially expressed genes between *Pik3cg*$^{-/-}$ and wild type mice (**Section 2.13**), 191 orthologous human genes were retrieved using BioMart (http://www.ensembl.org/biomart/martview/), and their respective proteins using UniProt (http://www.uniprot.org/). These 'core' proteins were used as primary seeds to develop the protein-protein interaction network. First-order interactors of core proteins were determined using Reactome v36 (http://www.reactome.org/). Only clustered non-redundant first-level interactions between human proteins that were connected to the largest connected component were considered. Based on the HaemAtlas data (Watkins et al., 2009), interactors that are not expressed in MKs (*P*>0.01) were excluded. Further details about the approach are described in Gieger et al., 2011.

2.15. Exome sequencing of individuals with TAR syndrome

Ethics statement. Informed consent was obtained from all study subjects with approval from the ethics committees of the following institutions: University Hospital Bristol (MREC/00/6/72), Universitair

Ziekenhuis Leuven (ML-3580), University of Cambridge (REC 10/H0304/66, REC 10/H0304/65), INSERM (RBM 1-14) and Charité Universitätsmedizin Berlin (EA2/170/05).

Exome sequencing and analysis. The 'baits' to capture the complete human exonic sequence were designed using annotations from the GENCODE Consortium, comprising a total of 740,000 exons in 79,000 transcripts from a highly redundant set of 34,642 genes, and covering 39.3 Mb of genomic sequence (Coffey et al., 2011). Exonic sequence was captured and enriched using the SureSelect Human All Exon Kit [Agilent Technologies] and sequenced on the Illumina GAII platform. Sequence analysis was performed as described in Albers et al., 2011. Variants that did not overlap the targeted regions ±25 bp were filtered and not considered for further analyses. The functional consequence of SNPs and small indels were predicted using the Ensembl Variation API (http://www.ensembl.org/info/docs/api/variation/). Each variant was annotated for presence in databases of genetic variation. Sequence variants with allele frequencies of up to 5% were considered, as inferred from variation data from dbSNP v131 (http://www.ncbi.nlm.nih.gov/SNP/), the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) and 354 exomes from the CoLaus Cohort (Firmann et al., 2008). Target enrichment for exome sequencing was performed at the Wellcome Trust Sanger Institute by the sequencing core group.

Data availability. Sequencing data are available online in the European Genome-phenome Archive (EGA) under accession number EGAD00001000018.

2.16. Sanger sequencing of the *RBM8A* locus

Primers for capillary sequencing were designed using Primer3 as described in **Section 2.9**, and are reported in **Table 2-11**. PCR products were amplified from genomic DNA using the ThermoStart Taq DNA Polymerase Kit [Thermo Scientific], and cleaned up using the ExoSAP-IT PCR Product Clean-Up Kit [USB]. Capillary sequencing was performed by Source Bioscience.

Table 2-11. Primer pairs used for Sanger sequencing of the *RBM8A* locus. Genomic coordinates were based on the human reference genome, build hg19.

Forward primer sequence	Reverse primer sequence	Genomic position	Description
CACGCCAGCCTCTGAGTT	CCCTAATCTCAAACCACTTCCT	chr1:145,501,836–145,502,420	upstream reg. element
GGACGAGCAGGAGACAGATG	CGCACCTGGCCTAAAAATTC	chr1:145,502,151–145,502,744	upstream reg. element
CAAAGCACACCCTGCACA	CACCTCCTGGGTTCAGGTAA	chr1:145,502,469–145,502,949	upstream reg. element

Forward primer sequence	Reverse primer sequence	Genomic position	Description
GGCCAGCCTGGGTAGTATAA	CCAGTCTGGGGGACAAGAG	chr1:145,506,316–145,506,883	promoter
TGCACCACTGCACTCTTAGC	TTTAGGCAGCGTGGTGTATG	chr1:145,506,650–145,507,248	promoter
GTCTCCCGGGTTCAACTG	TCTAAATCCCTCCCTCTGCAC	chr1:145,506,920–145,507,559	promoter
GCCCAGCTAATCAGCTTCC	TCCCTCTGCACGGTAAAAAC	chr1:145,506,991–145,507,549	promoter
TTTCCCAGTTTGGGATGAAG	GGGCGGAATCTCTAATCCAC	chr1:145,507,301–145,507,871	genic
GCCGGGCCTCACTGTTAAT	TCAGTTTGTGAATGCTCTCTGG	chr1:145,507,354–145,508,033	genic
GCCGCGGTTAAGAGGAAG	TTGTGAATGCTCTCTGGAACC	chr1:145,507,395–145,508,028	genic
ATGGCCACAGAAACACTTCC	CACCGCCTCCAGTCTTAGTG	chr1:145,507,474–145,507,924	genic
AGTTAGCCTTTGATTGGTCAGC	ACCCGTAGCTCCTGCCCTA	chr1:145,507,474–145,508,174	genic
ATGGCCACAGAAACACTTCC	TCCTCCTTTCTCCCATTGTTC	chr1:145,507,474–145,508,174	genic
ATGGCCACAGAAACACTTCC	CCACAGACACGGATACCTCA	chr1:145,507,474–145,508,324	genic
CGGGTCTTGGGTGGATTA	CCACAGACACGGATACCTCA	chr1:145,507,842–145,508,324	genic
GGGTCTTGGGTGGATTAGAGA	TTTAAGCAGGCTCACAGGAA	chr1:145,507,843–145,508,430	genic
GGGTTCCAGAGAGCATTCAC	GATATCCTGTTCGCCTGTCG	chr1:145,508,007–145,508,606	genic
CCTAGTAGGGCAGGAGCTACG	CCAACCACAGCAAACACAGA	chr1:145,508,105–145,508,692	genic
CGCAGTAGGAATGGGTTCAG	CCTGGGCTTCCTTGTATGTT	chr1:145,508,355–145,508,958	genic
GGCCAAGAGCAAAGTTGAAA	CCCAGTCCTATTTGTCCAAGG	chr1:145,508,691–145,509,284	genic
TTGTCAGACACGCCAAAGAG	CAATGATCCATACAGCCTTGC	chr1:145,508,736–145,509,438	genic
TGGGTGAAGGGAATACGAAC	ATGGTGGCATGTGCCTGTA	chr1:145,509,079–145,509,626	genic
GTGTTACCCAGGGTGGATTG	CATGCCTTTAGACAGCTGGA	chr1:145,509,508–145,509,946	genic
GGGAGGGACTTCAGTTAGCA	CCTGTTGCCTCTAGCATCATT	chr1:145,510,103–145,510,687	genic
TGATAGAAATATGAAGCCACCAAG	AAGGATGAATTGGGAGGAGAC	chr1:145,510,458–145,511,028	genic
AAGAGGCAGCAGAAGGTGAA	CAGCCCAATAGCATTTGGAA	chr1:145,510,814–145,511,453	genic
GGCTTGAATATGATGCTGAACA	GCCTGATCGTAACTCCAAACA	chr1:145,511,127–145,511,721	genic

2.17. Genotyping of the 5'-UTR and intronic SNPs at the *RBM8A* locus

The *RBM8A* 5'-UTR and intronic SNPs were genotyped in 7,504 individuals from the Cambridge BioResource with custom TaqMan SNP Genotyping Assays [Applied Biosystems] according to the manufacturer's protocols. All genotyping data were scored twice by different operators.

2.18. Sequencing of megakaryocyte RNA

Megakaryocyte RNA was prepared and sequenced as described in Albers et al., 2011. Reads were aligned to the *Homo sapiens* high-coverage assembly (build hg19) using the software GSNAP v2011-03-28 (Wu & Nacu, 2010). Read trimming was disabled and up to five mismatches were allowed. Newly identified splicing sites had to be at most 100 kb apart. Aligned reads were visualised using the Integrative Genomics Viewer (Robinson et al., 2011).

2.19. *RBM8A* promoter activity by luciferase reporter assay

Co-transfection experiments in different cell lines, i.e. EAHY926, HEK296, MC3T3, CHRF-288-11 and DAMI, were performed with pEGFP vectors [Clontech], and *RBM8A* reporter plasmids (wild type or with the 5'-UTR/intronic SNP) were constructed from pGL3-Basic luciferase vectors [Promega]. The *RBM8A* promoter region, starting at -303 nt upstream of the transcription start site and including exon 1 and the first 142 nt of intron 1, was cloned 5' to the luciferase gene. For each co-transfection assay, cells were transfected using Lipofectamine [Life Technologies] with 2 µg of pEGFP and 4 µg of *RBM8A*-pGL3 plasmid for HEK293, EAHY926 and MC3T3 cells. DAMI and CHRF-288-11 cells were transfected using the Amaxa electroporation system [Lonza] according to the manufacturer's instructions. Luciferase activity was determined as described in Freson et al., 2007. Each plasmid was assayed in triplicates in six separate transfection experiments. Firefly luciferase activity was normalised to EGFP expression. Statistical analysis was performed using the software InStat v3.01 [GraphPad].

2.20. Y14 protein expression analysis in platelet extracts

Blood (20 ml) anticoagulated with 3.8% trisodium citrate was centrifuged at 200xg to obtain platelet-rich plasma (PRP). The platelet pellet was obtained by centrifugation at 700xg after addition of 0.1 volume of ACD buffer at pH=4.5 (2.5% trisodium citrate, 1.5% citric acid and 2.0% D-glucose). The platelet pellet was lysed in ice-cold lysis buffer (1.0% Igepal CA-630 [Sigma-Aldrich], 1 mM EDTA, 2 mM DTE and 1x EDTA-free Protease Inhibitor [Complete Mini, Roche] per 50 ml of PBS) and subjected to three freeze-thaw cycles. Lysates were cleared of insoluble debris by centrifugation for 10 min at 4°C and 16,100xg. Protein fractions were mixed with 5% SDS reducing sample buffer, separated by SDS-PAGE and transferred to Hybond ECL-nitro-cellulose membranes [GE Healthcare]. The blots were blocked for 1 hr at RT in Tris-buffered saline with Tween-20 supplemented with 5% non-fat dry milk. The blots were first incubated with primary antibody overnight at 4°C, and then with horseradish peroxidase (HRP)-conjugated secondary antibody for 2 hr at RT. The following primary antibodies were used: rabbit polyclonal antibody against Y14 (Q-24), mouse monoclonal antibody against Y14 (4C4) [Santa Cruz Biotechnology], mouse monoclonal antibody against Gsα (Freson et al., 2001) and mouse monoclonal antibody against β-actin [A5441, Sigma-Alrich]. Both Y14 antibodies were tested for their specificity using recombinant Y14-GST purified by sepharose beads as described in Freson et al., 2001. Signal was detected with the ECL Western Blotting Substrate [Thermo Fisher Scientific] according to the manufacturer's protocol. Densitometry analysis was carried out using the software ImageJ64.

CHAPTER 3

Maps of open chromatin guide the functional follow-up of genome-wide association signals at haematological trait loci.

This chapter is in parts based on the following publication:

Paul, D.S., et al. (2011). Maps of open chromatin guide the functional follow-up of genome-wide association signals: application to hematological traits. PLoS Genet. *7*, e1002139.

3.1. Introduction

The first objective of this thesis was to generate maps of open chromatin in myeloid cell types using the formaldehyde-assisted isolation of regulatory elements (FAIRE) technique. This method provides an effective means of discovering active gene regulatory elements through the identification of nucleosome-depleted regions (NDRs). The second objective was to intersect the identified NDRs with GWA signals of haematological and cardiovascular-related traits. I hypothesised that this may identify sequence variants that play a role in regulation of gene expression.

In this chapter, I first describe the implementation and optimisation of the FAIRE assay in our laboratory. Then, I apply FAIRE to test the above hypothesis by overlapping FAIRE-generated NDRs identified in a megakaryocytic and an erythroblastoid cell line with sequence variants associated with haematological and cardiovascular-related quantitative traits, as well as coronary artery disease and myocardial infarction. NDRs are mapped at selected GWA loci on high-density oligonucleotide tiling microarrays.

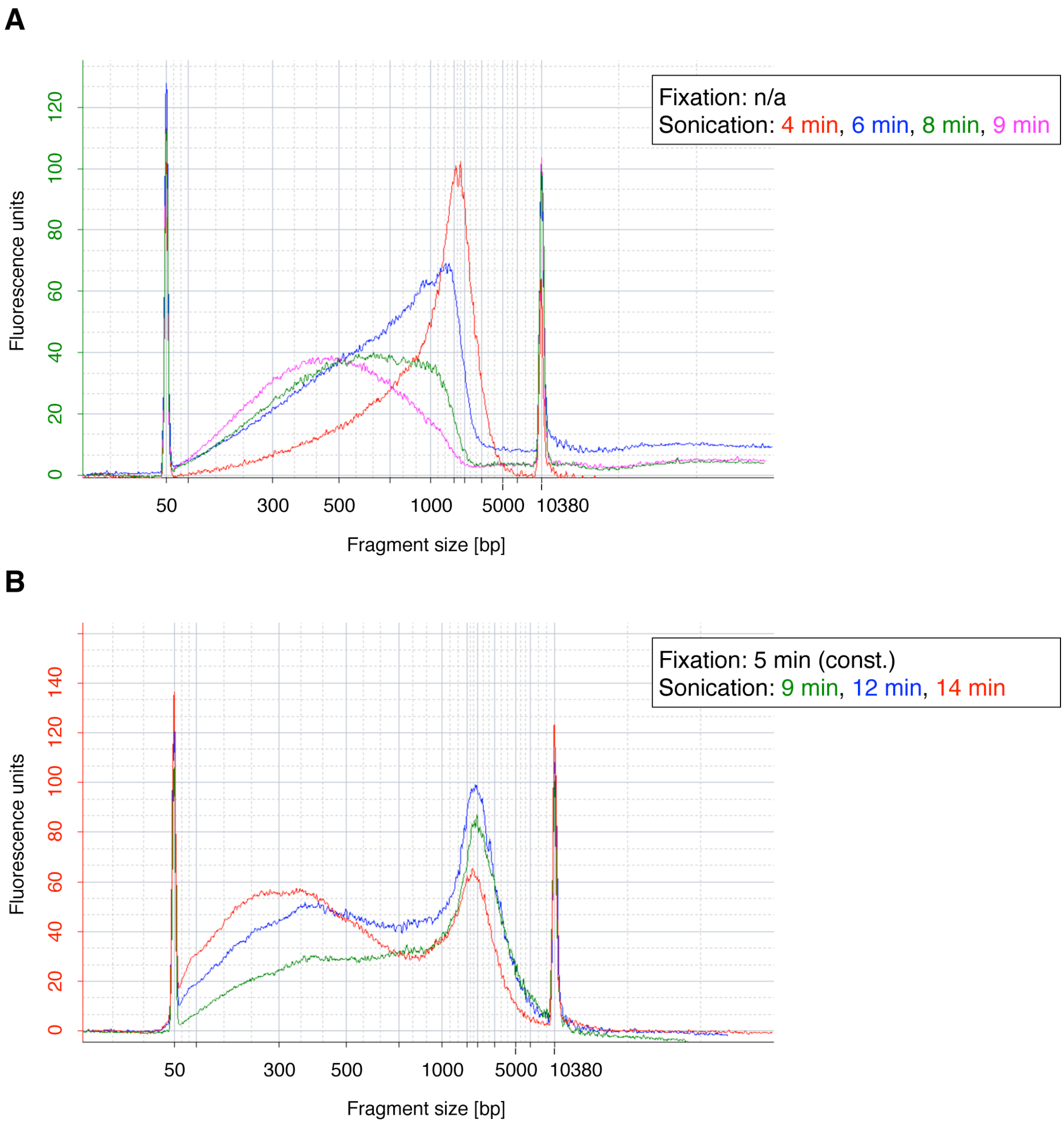
3.2. Optimisation of formaldehyde-assisted isolation of regulatory elements (FAIRE)

The experimental procedure of FAIRE is well-documented in the literature (Giresi et al., 2007; Giresi & Lieb, 2009). However, the efficiency of formaldehyde fixation may vary between the cell lines and primary tissues studied, due to factors including differences in cellularity, permeability, purity and surface area. In addition, the efficiency of chromatin shearing may vary with respect to the laboratory equipment used. To ensure consistent high-quality FAIRE data, I therefore optimised both fixation and sonication using the megakaryocytic cell line CHRF-288-11 (**Table 2-1**).

Sonication. I aimed to shear chromatin to a fragment range of 100–1,000 bp with an average of 500 bp, resulting in a final range of 75–200 bp after phenol-chloroform extraction and DNA precipitation, as described by Giresi & Lieb, 2009. DNA fragments within this range are suitable for optimal hybridisation to oligonucleotide arrays and next-generation sequencing. Both cross-linked and uncross-linked chromatin samples were sonicated over a time course to monitor the distribution of DNA fragment lengths (**Figure 3-1**). For cross-linked chromatin, DNA-protein cross-links were reversed prior to the analysis with an Agilent Bioanalyzer. The sonication time for an optimal distribution of DNA fragment lengths was 9 and 12 min for the uncross-linked (**Figure 3-1 A**) and cross-linked sample

(**Figure 3-1 B**), respectively. Both experiments were repeated, and confirmed the initial findings (data not shown).

Fixation. Independent of the applied formaldehyde fixation times, the resulting fragment range of 50–250 bp after phenol-chloroform extraction and DNA precipitation was within the optimal range, as suggested by Giresi & Lieb, 2009 (**Figure 3-1 C**). However, the quantity of nucleosome-depleted DNA fragments decreased with longer formaldehyde fixation times, as a smaller number of open chromatin regions were recovered.



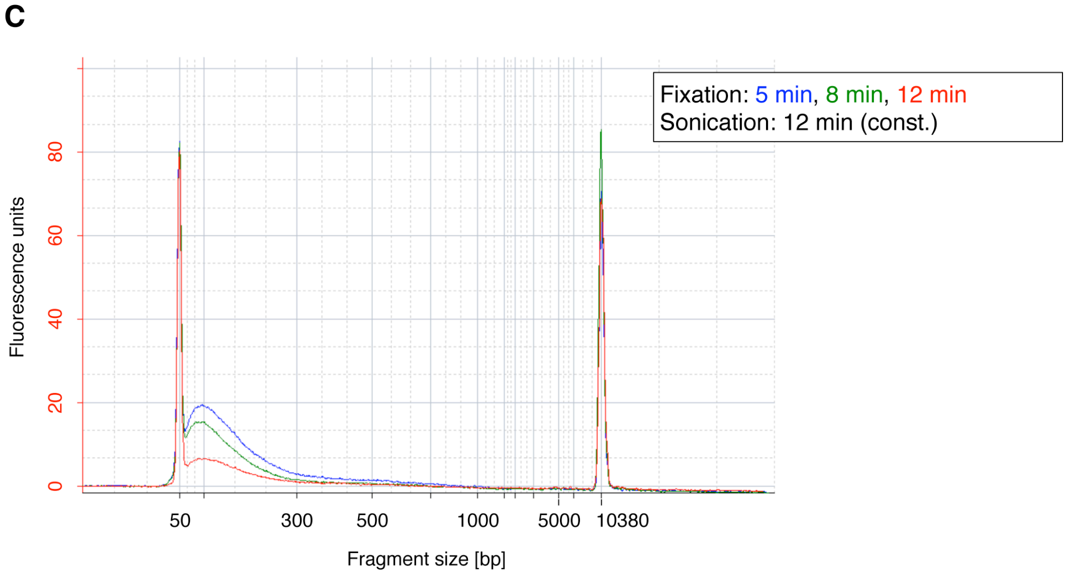


Figure 3-1. Electropherograms of time course experiments to monitor the distribution of DNA fragment lengths after sonication and formaldehyde fixation of chromatin. Shown are the electropherograms of (A) uncross-linked chromatin and (B) chromatin cross-linked with 1% formaldehyde for 5 min after applying different sonication times. After formaldehyde fixation, DNA-protein interactions were reversed prior to analysis with an Agilent Bioanalyzer. This resulted in a proportion of large DNA fragments of ~2,000 bp, which represents inaccessible genomic regions, i.e. regions that are occupied with histones. Panel (C) shows the effect of formaldehyde fixation on DNA fragment length and quantity whilst applying a constant sonication time. Here, chromatin samples were extracted by phenol-chloroform and purified prior to analysis. The colour of the y-axis refers to the sonication time to which data was referenced.

3.3. Design of an oligonucleotide tiling microarray for NDR mapping

I designed a 385,000-oligonucleotide tiling array using 72 known genetic loci associated with haematological and cardiovascular-related traits (**Table 3-1**, based on the National Human Genome Research Institute catalogue of published GWA studies, http://www.genome.gov/gwastudies/, as of November 2009). The array design has to be considered a 'snapshot' of the published GWA studies at this time, as subsequent GWA studies may have identified additional genetic loci. However, I considered the design appropriate given the proof-of-principle nature of this study.

Genetic loci were only considered if they reached genome-wide significance with the threshold of $P<5\text{x}10^{-8}$ (or as otherwise indicated in **Appendix, Table 8-1 A**) in a GWA study conducted with individuals of Northern and Western European ancestry (CEU population). In addition, I selected genetic loci based on biological evidence, where there was suggestive evidence of association. For each locus, the entire genetic region of the index SNP was included, as defined by recombination hotspots based on Phase II HapMap (Myers et al., 2008). If a recombination interval exceeded 500 kb, I included the closest target gene ±10 kb. In addition, I included eight lineage-specific reference genes on the array for each of megakaryocytes, erythroblasts and monocytes, in order to assess patterns of cell type specificity (**Appendix, Table 8-1 B**). I selected these transcripts on the basis of their expression profiles (**Figure 3-6 A**) according to the HaemAtlas, a systematic analysis of expression profiles in differentiated human blood cells (Watkins et al., 2009).

The oligonucleotide (50–75-mer probes) tiling array [Roche NimbleGen] provided a mean probe span of 23 bp and harboured only probes unique to the human genome (build: hg18, coverage: 79%). In summary, a total of 62 unique complex trait loci fulfilling the above criteria and 24 reference gene loci representing 9.59 Mb and 1.77 Mb of genomic DNA, respectively, were selected for the array design.

Table 3-1. Summary of the genetic loci included on the custom DNA tiling array. A detailed list, including genomic coordinates of the intervals and references, is presented in Appendix, Table 8-1 A.

Complex trait	Number of genetic loci	Genomic footprint
Coronary artery disease (CAD) and (early-onset) myocardial infarction (MI)	18	3,605.2 kb
Mean platelet volume (MPV)	12	1,789.7 kb
Platelet count (PLT)	4	712.2 kb
Platelet signalling (PLS)	15	1,968.4 kb
White blood cell count (WBC)	1	73.2 kb
Red blood cell count (RBC)	1	50.0 kb
Mean corpuscular volume (MCV)	4	436.0 kb
Mean corpuscular haemoglobin (MCH)	1	100.0 kb
Systolic blood pressure (SBP)	6	940.0 kb
Diastolic blood pressure (DBP)	8	1,475.0 kb
Hypertension (HYP)	2	240.0 kb
Total	72	11,149.7 kb
Total unique	62	9,593.5 kb

3.4. Open chromatin profiles in a megakaryocytic and an erythroblastoid cell line

I profiled chromatin accessibility at 62 non-redundant genetic loci, representing all associations known in November 2009 with 11 haematological and cardiovascular-related traits (**Table 3-1**) in the megakaryocytic cell line CHRF-288-11 ('MK cells') and the erythroblastoid cell line K562 ('EB cells').

Maps of open chromatin were created with FAIRE applied under optimised experimental conditions (**Section 3.2**). For each cell line, I prepared an uncross-linked sample with a sonication time of 9 min and two cross-linked samples with formaldehyde fixation times of 8 and 12 min and a sonication time of 12 min. DNA derived from cross-linked cells ('FAIRE DNA') and uncross-linked cells ('reference DNA') were labelled with the fluorescence dyes Cy5 and Cy3, respectively. Subsequent hybridisation to the custom oligonucleotide array spanning the selected loci, and analysis on a dual-channel array platform revealed enriched regions (NDRs) in the FAIRE DNA sample compared to the reference DNA sample. Enriched regions (peaks) were called using the software NimbleScan v2.5 [Roche NimbleGen].

Open chromatin regions showed high concordance across cross-linking conditions, with 95.3% and 93.1% overlapping regions in MK and EB cells, respectively. In order to reduce experimental error and achieve higher stringency, only concordant open chromatin regions retained from each cell type were subject to further analysis (**Figure 3-2**).

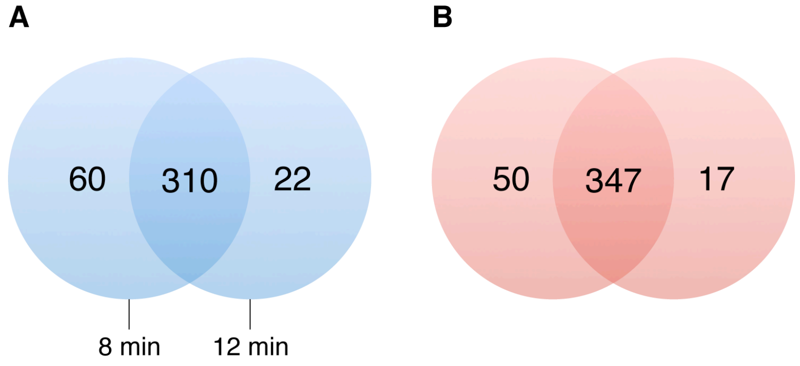


Figure 3-2. Number of peaks in FAIRE-chip data sets. FAIRE was performed with different formaldehyde cross-linking times (8 and 12 min), and enriched regions identified using the software NimbleScan v2.5 [Roche NimbleGen]. Shown is the number of peaks across the entire array content, either present in both data sets using different formaldehyde fixation times or unique to one of the data sets, identified in (A) EB and (B) MK cells. The protocol using 8 min formaldehyde fixation retained a larger number of open chromatin regions. Replicated and overlapping (at least 1 bp) peaks were merged for each cell type and subjected to downstream analysis.

At the 62 selected GWA loci, I identified 254 and 251 NDRs in MK and EB cells, respectively, of which 147 (57.9% and 58.6%, respectively) were common to both cell types. A substantial overlap is expected between these two cell types, as they share a bi-potent progenitor cell – the megakaryocyte-erythrocyte progenitor (MEP) cell (McDonald & Sullivan, 1993; Pang et al., 2005; Miranda-Saavedra & Göttgens, 2008; **Section 1.9**).

To evaluate the parameters and performance of the peak finding algorithm, the peak data sets were compared with an existing data set provided by the ENCODE Pilot Project. The aim of the ENCODE Pilot was to establish experimental and analytical methods to generate a catalogue of functional DNA elements across different human cell lines (ENCODE Project Consortium, 2007). The 44 ENCODE regions comprised a total of ~30 Mb of genomic space (1% of the human genome) and were divided into regions for which there was already substantial biological knowledge, and randomly chosen regions. I assumed that the ENCODE regions had a similar gene density as the association regions selected for the tiling array.

FAIRE peak density (per megabase) in my data set was roughly consistent with that in foreskin fibroblast cells reported by the ENCODE Project. However, a slightly higher number of peaks were found in my peak data sets (**Table 3-2**). This may be due to the choice of certain loci and in particular, the incorporation of the 24 reference gene loci to the array content. This presumably resulted in an enrichment of promoter and other regulatory regions in comparison to the GWA loci, where accessibility of regulatory factors to chromatin is expected.

Table 3-2. Comparison of the FAIRE peak density between the ENCODE data set and the data sets presented here.

Human cell line	Formaldehyde cross-linking	Total sequence coverage	Total number of peaks	Number of peaks per Mb
CCD-1070Sk (Foreskin fibroblasts)	7 min	29,998 kb	1,008	33.6
CHRF-288-11 (Megakaryocytes)	8 min	9,651 kb	397	41.1
	12 min		364	37.7
K562 (Erythroblasts)	8 min	9,651 kb	376	39.0
	12 min		333	34.5

3.5. Characterisation of open chromatin regions in relation to gene annotations and cell types

I then analysed the 254 and 251 NDRs at the GWA loci in MK and EB cells, respectively, with regard to their genomic location, i.e. intergenic, intronic, overlap 5'-untranslated region (UTR), overlap 3'-UTR or exonic (**Table 3-3**). It is important to note that the observations were based on a selected set of loci and therefore cannot be extrapolated to the whole genome. For the genomic characterisation of FAIRE peaks, I retrieved all annotation from the Ensembl database v54 (build: hg18). NDRs were most frequently located at non-coding segments (98.2% and 92.5% of peaks found only in MK and EB cells, respectively, and 98.1% of peaks common to both cell types). Promoter/5'-UTR regions were enriched in NDRs common to both cell types (28.4%) compared to open chromatin specific to either cell type, i.e. 4.6% (6.2-fold) and 5.6% (5.1-fold) for MK and EB cells, respectively.

Table 3-3. Characterisation of open chromatin regions in relation to gene annotations.

Genomic location	MK cells only	EB cells only	Both cell types
Intergenic	20.2%	39.3%	26.5%
Intronic	72.5%	45.8%	41.4%
Overlap 3'-UTR	0.9%	1.9%	1.9%
Overlap 5'-UTR	4.6%	5.6%	28.4%
Exonic	1.8%	7.5%	1.9%

At the GWA loci, NDRs clustered around transcription start sites (TSS). In MK and EB cells, respectively 70.5% and 76.1% of all FAIRE peaks were located within 20 kb of a TSS (**Figure 3-3**). However, accessible chromatin regions as far as 264 kb upstream of a TSS were also detected (*TBX3* gene locus). These may represent distal regulatory elements or regulatory elements of yet unannotated genes. Open chromatin observed in MK but not EB cells was located on average 2.80 kb upstream of a TSS. NDRs found in EB but not MK cells were located on average 1.77 kb upstream of a TSS, whereas NDRs common to both cell types were on average 0.98 kb upstream of a TSS.

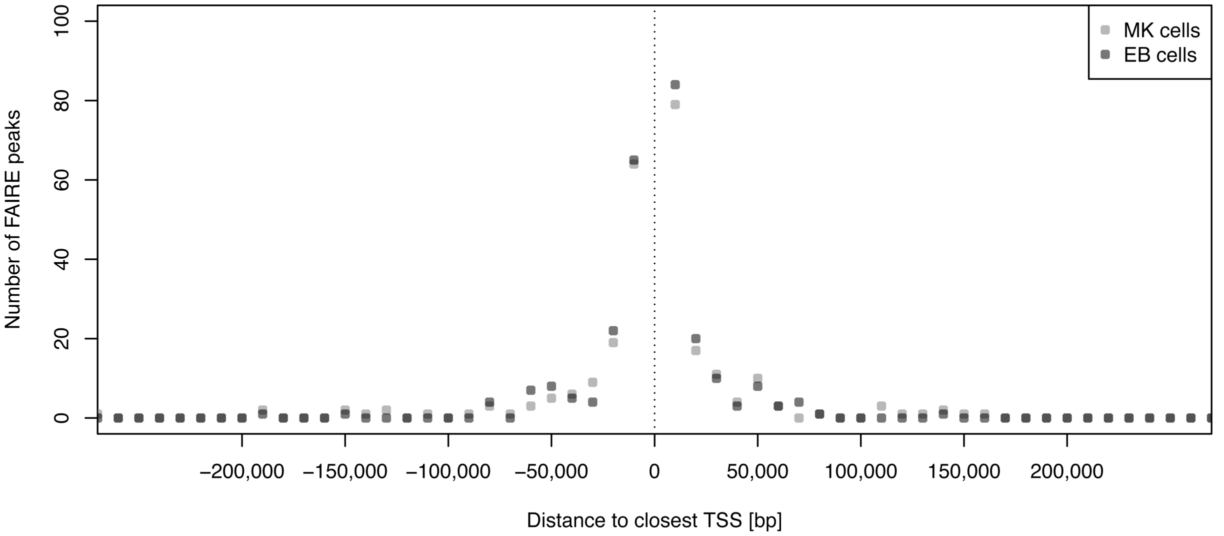


Figure 3-3. Location of open chromatin sites with respect to the closest TSS at the selected GWA loci.

To assess the cell type specificity of NDRs marked by FAIRE, I determined the number of peaks in lineage-specific genes for MK and EB cells present on the array (**Figure 3-4**). A significant enrichment of FAIRE peaks at MK lineage-specific genes was observed in MK cells, when compared to the number of peaks in EB cells (*P*=0.0225, Wilcoxon rank-sum test). A similar trend of enrichment was observed in EB lineage-specific genes in EB cells (*P*=0.0781). This result highlights the importance of studying chromatin architecture and gene regulatory circuits in a cell type-dependent manner.

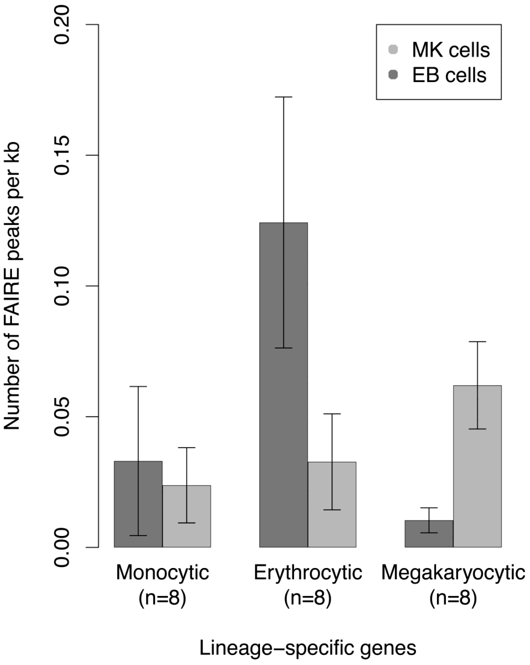


Figure 3-4. Average number of FAIRE peaks in lineage-specific genes in MK and EB cells. The number of open chromatin sites in lineage-specific genes (±2 kb) was averaged and normalised for the length of the gene. Error bars indicate standard error of the mean.

3.6. Seven sequence variants associated with haematological and cardiovascular-related quantitative traits are located in NDRs

At seven of the 62 tested GWA loci, I found SNPs in strong LD with the corresponding GWA index SNP located within an NDR (**Table 3-4**). Proxy SNPs were identified using the Genome-wide Linkage Disequilibrium Repository and Search Engine (GLIDERS) (Lawrence, Day-Williams, et al., 2009) with the following settings: Phase II HapMap v23 (CEU population); MAF limit≥0.05; r^2 limit≥0.8; no distance limits. Five out of the seven loci were associated with platelet-related quantitative traits.

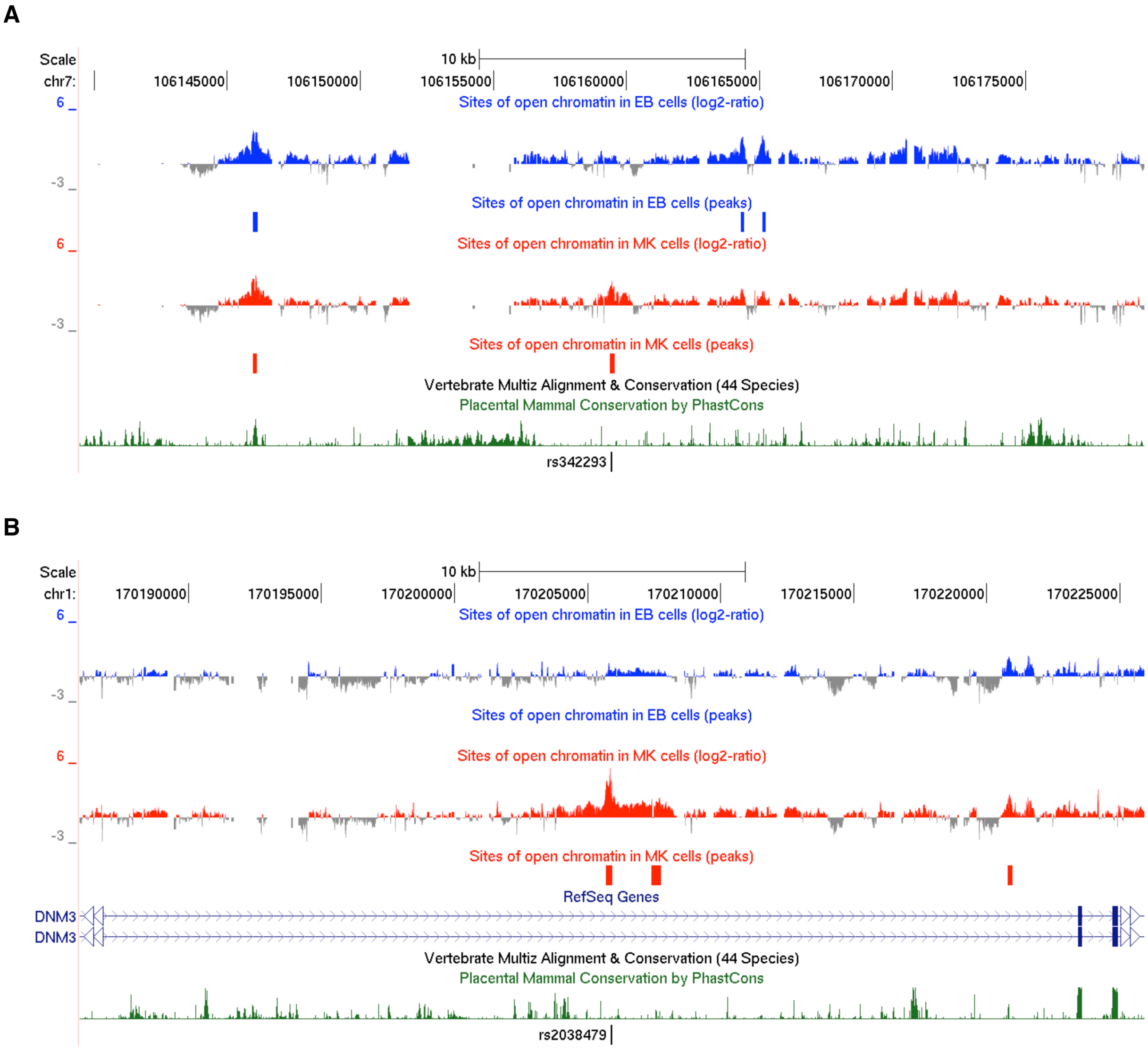
Table 3-4. Seven sequence variants associated with haematological and cardiovascular-related traits are located in NDRs. Abbreviations: MK: megakaryocytic cell line; EB: erythroblastoid cell line; MPV: mean platelet volume; MCV: mean corpuscular volume of erythrocytes; PLS: platelet signalling; SBP: systolic blood pressure; MAF: minor allele frequency.

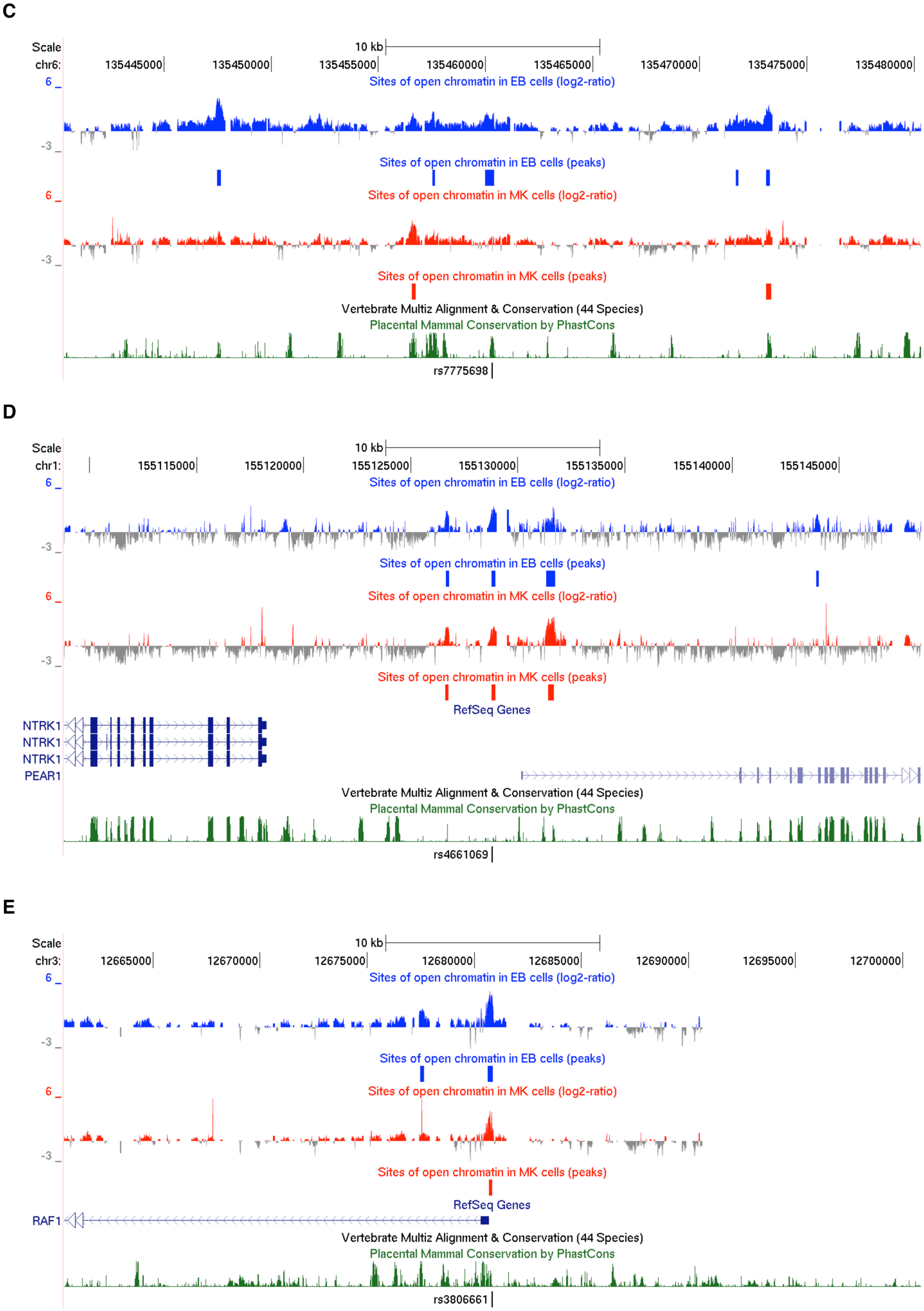
Cell type	Trait	Locus	SNP in open chromatin			GWA index SNP		
			ID	*MAF*	*Annotation*	*ID*	*r^2*	*Distance*
MK	MPV	*FLJ36031-PIK3CG*	rs342293	0.45	Intergenic	rs342293	1.00	(index SNP)
MK	MPV	*DNM3*	rs2038479	0.16	Intronic	rs10914144	0.94	10 kb
EB	MCV	*HBS1L-MYB*	rs7775698	0.22	Intergenic	rs9402686	0.85	9 kb
EB/MK	MPV	*TMCC2*	rs1172147	0.35	Intronic	rs1668873	0.89	10 kb
EB/MK	PLS	*PEAR1*	rs4661069	0.11	Promoter	rs3737224	0.83	17 kb
EB/MK	PLS	*RAF1*	rs3806661	0.30	Promoter	rs3729931	0.85	79 kb
EB/MK	SBP	*CYP17A1-C10orf32*	rs3824754	0.07	Intronic	rs1004467	1.00	19 kb

At these seven loci, NDRs were found only in MK but not EB cells ('MK-specific', n=2), in EB but not MK cells ('EB-specific', n=1), or in both cell types (n=4). The two MK-specific NDRs harbouring SNPs associated with mean platelet volume (MPV) were located at an intergenic region of the *FLJ36031-PIK3CG* gene locus (**Figure 3-5 A**) and an intronic region of *DNM3* (**Figure 3-5 B**). Both genes, *PIK3CG* and *DNM3*, were upregulated in megakaryocytes compared to erythroblasts (2.08- and 6.42-fold, respectively, according to the HaemAtlas). The EB-specific NDR was located at an intergenic region of the *HBS1L-MYB* gene cluster (**Figure 3-5 C**). Sequence variants at this locus are known to be associated with mean corpuscular volume (MCV) of erythrocytes, mean corpuscular haemoglobin (MCH) and red blood cell count (RBC). *HBS1L* and *MYB* were upregulated in EB cells (1.30- and 2.40-fold, respectively, according to the HaemAtlas). In the four NDRs common to both cell types, I found variants associated with: platelet signalling (PLS) located in the promoter regions of *PEAR1* (**Figure 3-5 D**) and *RAF1* (**Figure 3-5 E**); MPV found in an intronic region of *TMCC2* (**Figure 3-5 F**);

and systolic blood pressure (SBP) in an intronic region of *C10orf32* (*CYP17A1* gene cluster; **Figure 3-5 G**).

Expression profiles of these genes (**Table 3-4**) based on the HaemAtlas data confirmed transcription in both MK and EB cells (**Figure 3-6-B**).




C
Scale
chr6:
10 kb
135445000
135450000
135455000
135460000
135465000
135470000
135475000
135480000
6
-3
Sites of open chromatin in EB cells (log2-ratio)
Sites of open chromatin in EB cells (peaks)
Sites of open chromatin in MK cells (log2-ratio)
Sites of open chromatin in MK cells (peaks)
Vertebrate Multiz Alignment & Conservation (44 Species)
Placental Mammal Conservation by PhastCons
rs7775698
D
Scale
chr1:
10 kb
155115000
155120000
155125000
155130000
155135000
155140000
155145000
6
-3
Sites of open chromatin in EB cells (log2-ratio)
Sites of open chromatin in EB cells (peaks)
Sites of open chromatin in MK cells (log2-ratio)
Sites of open chromatin in MK cells (peaks)
RefSeq Genes
NTRK1
NTRK1
NTRK1
PEAR1
Vertebrate Multiz Alignment & Conservation (44 Species)
Placental Mammal Conservation by PhastCons
rs4661069
E
Scale
chr3:
10 kb
12665000
12670000
12675000
12680000
12685000
12690000
12695000
12700000
6
-3
Sites of open chromatin in EB cells (log2-ratio)
Sites of open chromatin in EB cells (peaks)
Sites of open chromatin in MK cells (log2-ratio)
Sites of open chromatin in MK cells (peaks)
RefSeq Genes
RAF1
Vertebrate Multiz Alignment & Conservation (44 Species)
Placental Mammal Conservation by PhastCons
rs3806661

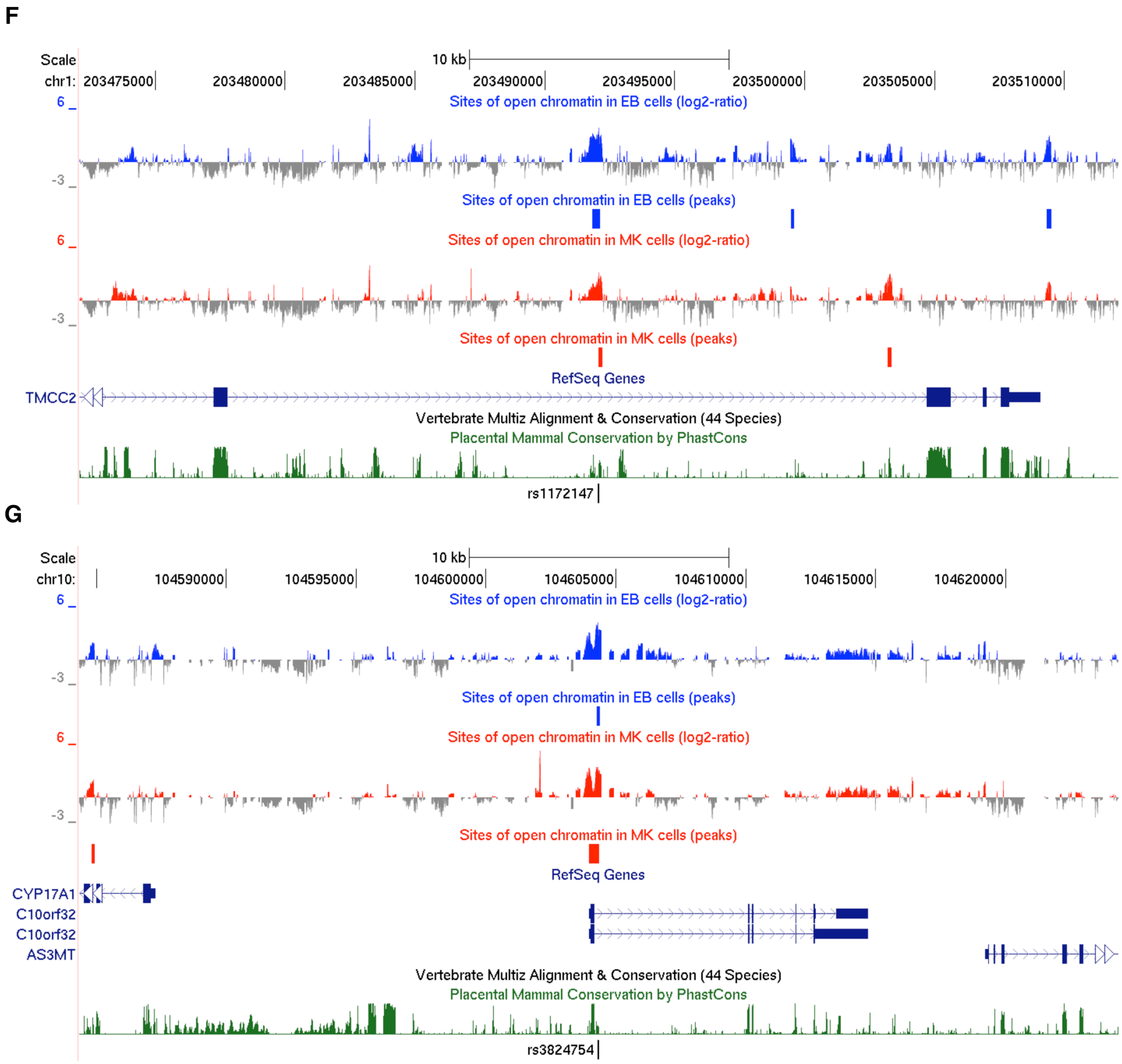


Figure 3-5. Open chromatin profiles at selected genetic loci in MK and EB cells displayed as UCSC Genome Browser custom tracks. (A) *FLJ36031-PIK3CG*; (B) *DNM3*; (C) *HBS1L-MYB*; (D) *PEAR1*; (E) *RAF1*; (F) *TMCC2*; (G) *CYP17A1-C10orf32*. Shown are the scaled $\log_2$-ratio and the called peaks from FAIRE experiments in an erythroblastoid (blue) and a megakaryocytic cell line (red). Only the data sets using a formaldehyde fixation time of 12 min are shown for both cell types. The putative regulatory SNP located within a site of open chromatin is shown below each track.

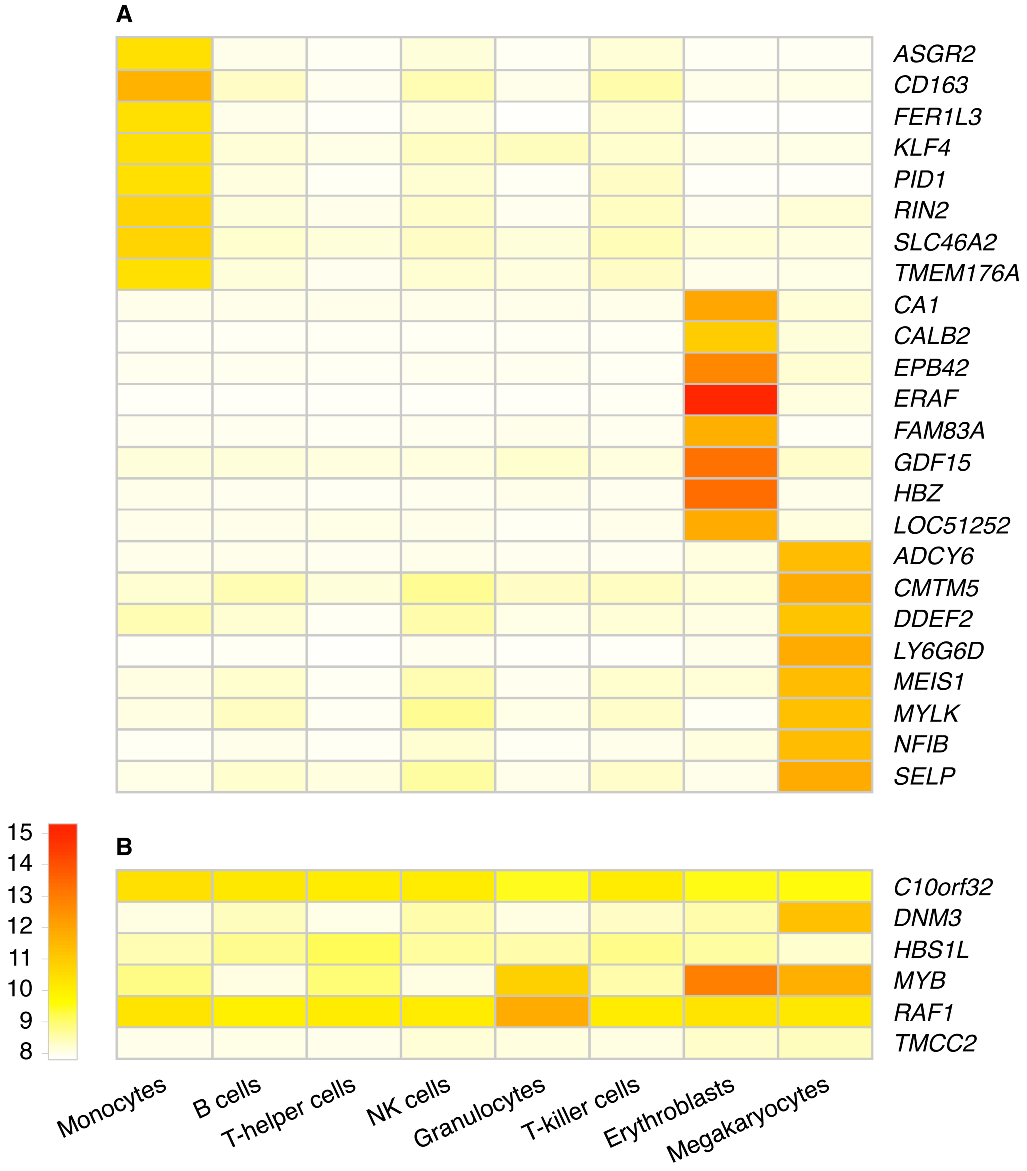


Figure 3-6. Gene expression profiles in differentiated human blood cells. Based on the HaemAtlas data (Watkins et al., 2009), the heat map shows normalised gene expression profiles of (A) lineage-specific reference genes selected for the custom DNA tiling array, and (B) genes that harbour cell type-specific open chromatin and putative regulatory sequence variants. Expression profiles of genes at the *FLJ36031-PIK3CG* locus are reported in Figure 5-5 A. The analysis to define lineage-specific genes was performed by Nicholas Watkins and Augusto Rendon. The following parameters were applied: detection of marker *P*<0.0001, with signal intensity *I*>10, whereas other markers must have *I*≤8.7.

3.7. Discussion

I applied the FAIRE assay to generate a catalogue of NDRs in a megakaryocytic and an erythroblastoid cell line at 62 selected genetic loci associated with haematological and cardiovascular-related traits. I provided initial evidence that open chromatin profiles exhibit distinct patterns among different cell types, and that cell type-specific NDRs may be useful in prioritising regions for further functional analysis (**Table 3-4**). Thus, the intersection of maps of open chromatin with variants identified through GWA studies may facilitate the search for underlying functional variants.

Seven putative functional variants associated with haematological and cardiovascular-related traits were located in sites of open chromatin (**Table 3-4**). Correlation of signatures of open chromatin with experimentally determined transcription factor binding sites in different cell types could systematically and rapidly translate GWA signals into functional components and biological mechanisms.

Access to cell types relevant to the studied trait is not always feasible and can be a limitation for functional studies. For instance, the majority of the identified candidate functional SNPs are associated with platelet quantitative traits (**Table 3-4**), suggesting that MKs/megakaryocytic cells may be an effector cell type. Conversely, I did not observe any intersection of FAIRE peaks with variants associated with CAD/MI or hypertension, indicating that cell types other than the megakaryocytic and erythroblastoid cell lines analysed here may be more suitable. Other possibilities are that the studied cells may have to be exposed to certain stimuli that influence chromatin structure and binding of regulatory factors. In addition, the immortalised cell lines may have become altered during serial passaging. In subsequent studies, primary cell types and tissues were used to overcome this caveat (**Chapter 4**).

By converting the read-out system from microarrays to high-throughput next-generation sequencing, genome-wide open chromatin profiles can be interrogated (**Chapter 4**). This would scale up analysis to the whole genome, generating a catalogue of open chromatin profiles to annotate association loci identified in past and future GWA studies.

CHAPTER 4

Maps of open chromatin highlight cell type-specific patterns of regulatory sequence variation at haematological trait loci.

Collaboration note:

Section 4.2: Katrin Voss[1,2] and I conceived the experiments. Primary myeloid cells were prepared by Katrin Voss and Jonathan Stephens[1,2]. Cornelis A. Albers[1–3] and Augusto Rendon[1,2,4,5] helped with raw sequencing data analyses. Cornelis A. Albers performed peak normalisation and hierarchical clustering. Augusto Rendon analysed the overlap of open chromatin and histone mark peak data sets, as well as gene expression profiles during *in vitro* differentiation of cord blood-derived haematopoietic stem cells. I performed FAIRE assays, sequencing data analysis, ontology and pathway analyses, and interpreted the results.

Section 4.3: Cornelis A. Albers and Augusto Rendon performed enrichment analyses. On behalf of the HaemGen Consortium, Pim van der Harst[6,7], John C. Chambers[8–11] and Nicole Soranzo[3] provided genome-wide association data sets of haematological traits. I performed ontology and pathway analyses, and contributed towards the interpretation of the results.

[1]Department of Haematology, University of Cambridge, Cambridge, UK; [2]National Health Service (NHS) Blood and Transplant, Cambridge, UK; [3]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; [4]Biostatistics Unit, Medical Research Council, Cambridge, UK; [5]NIHR Biomedical Research Centre, Cambridge, UK; [6]Department of Cardiology, University of Groningen, University Medical Center Groningen, The Netherlands; [7]Department of Genetics, University of Groningen, University Medical Center Groningen, The Netherlands; [8]Department of Epidemiology and Biostatistics, Imperial College London, London, UK; [9]Imperial College Healthcare NHS Trust, Hammersmith Hospital, London, UK; [10]Royal Brompton and Harefield Hospitals NHS Trust, London, UK; [11]Ealing Hospital NHS Trust, Southall, Middlesex, UK.

4.1. Introduction

In **Chapter 3**, I provided initial evidence that FAIRE is a valuable tool in mapping NDRs at selected genetic loci associated with haematological traits, and in identifying candidate functional variants for experimental validation.

To generalise this approach to the whole genome, the work in this chapter creates genome-wide maps of open chromatin in primary human myeloid cells using FAIRE combined with next-generation sequencing (FAIRE-seq). NDRs are mapped in megakaryocytes (MKs) and erythroblasts (EBs), the precursor cells of platelets and erythrocytes, respectively, as well as monocytes (MOs). We then define global cell type-specific patterns of gene regulatory variation at genetic loci associated with platelet and erythrocyte phenotypes.

4.2. Functional characterisation of open chromatin profiles in human myeloid cells

Cord blood-derived CD34$^+$ haematopoietic progenitor cells (HPCs) from two unrelated individuals were differentiated *in vitro* into either MKs in the presence of thrombopoietin (TPO) and interleukin-1β (IL-1β), or EBs in the presence of erythropoietin (EPO), interleukin-3 (IL-3) and stem cell factor (SCF). MOs were purified from the peripheral blood of another two individuals (**Section 2.2**). In addition, we generated FAIRE-seq data in the megakaryocytic cell line CHRF-288-11 (two biological replicates), and obtained FAIRE-seq ENCODE data (The ENCODE Project Consortium, 2011) for the erythroblastoid cell line K562. **Figure 4-1** gives an overview of the study design.

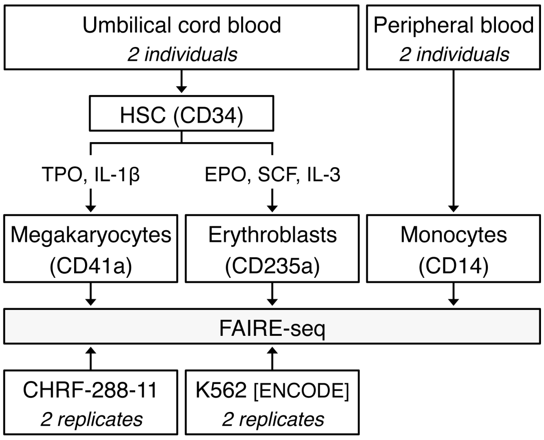
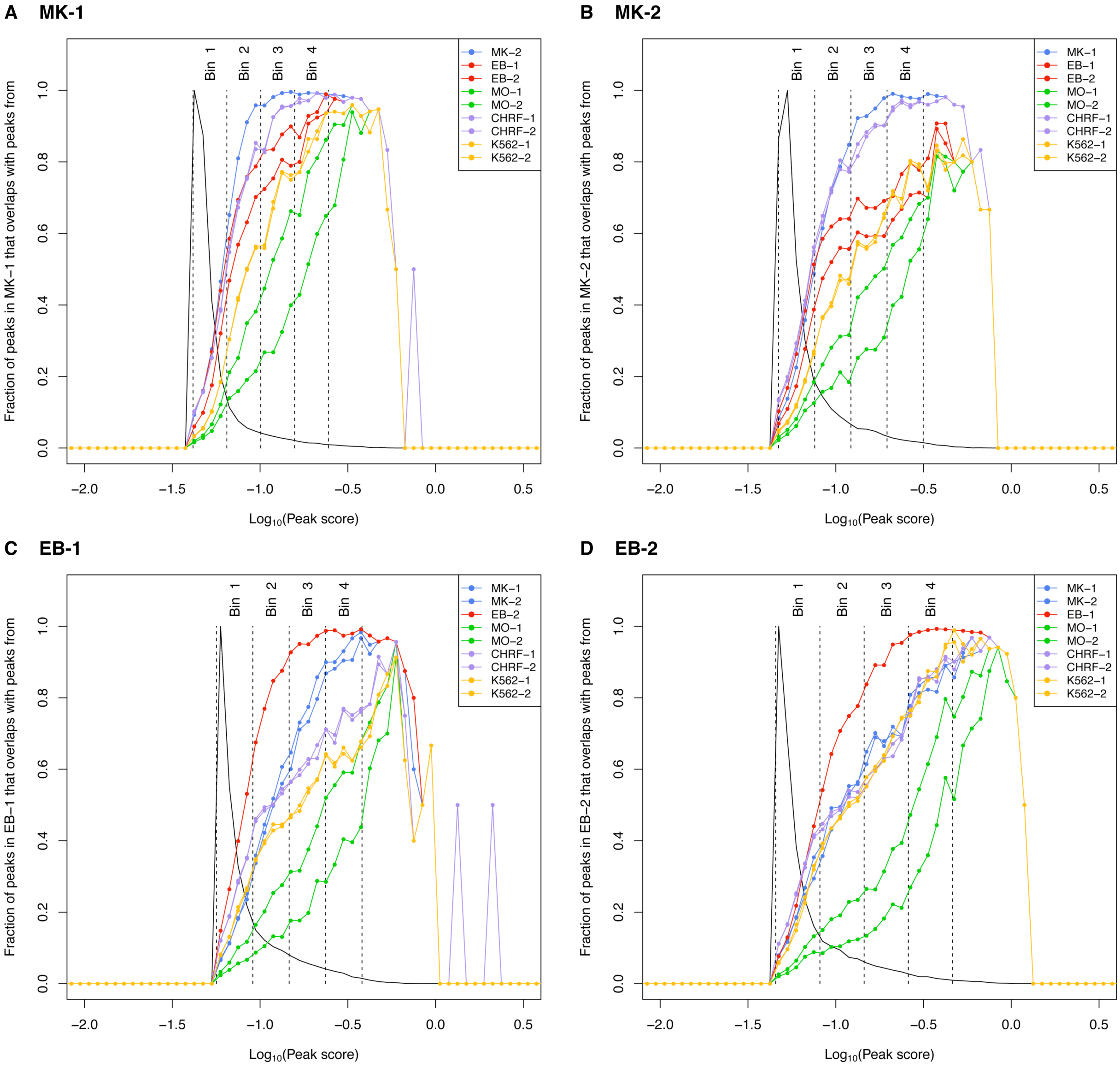
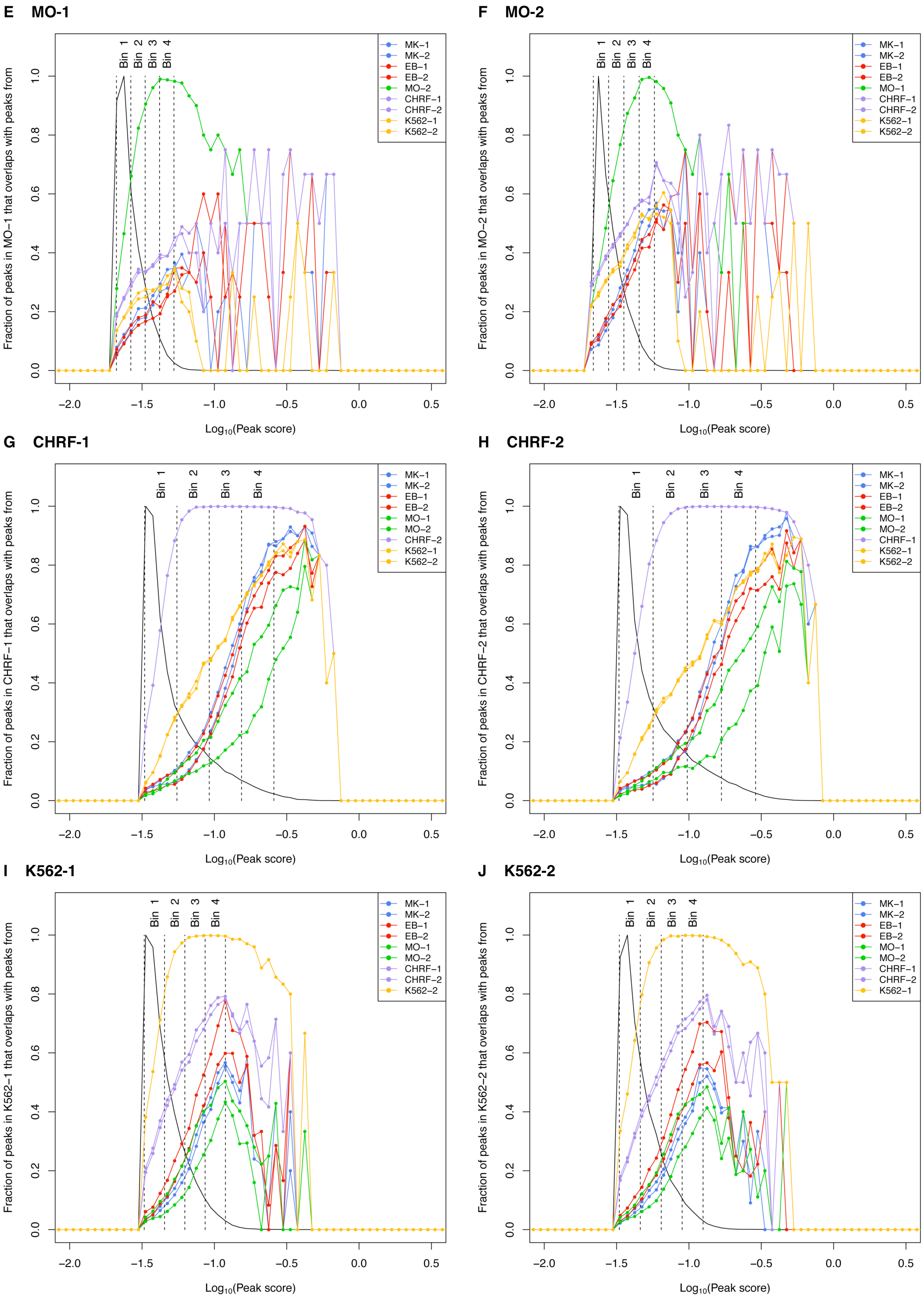


Figure 4-1. Overview of the study design.

I determined FAIRE peaks marking NDRs using a Gaussian kernel density estimator implemented in the software F-Seq (Boyle, Guinney, et al., 2008). In order to reduce false positive peak calls, regions of collapsed repeats were removed, as described in Pickrell et al., 2011 (**Section 2.5**). For each data set, I excluded the top and bottom one percentile of the peak score distribution (in $\log_{10}$-transformed peak score units). Peaks at the extremes of the peak score distribution may contain outliers due to sequencing errors that could bias downstream analyses. We then stratified the remaining peaks into four equally sized 'intensity bins' according to their score (**Figure 4-2**). The number of intensity bins was chosen arbitrarily, but assigning four resulted in a large number of peaks per bin. This is important for the power in subsequent statistical analyses. In addition, we expected that the FAIRE signal would be reasonably stratified across the four bins. An overview of the number of peaks per intensity bin for each cell type is reported in **Table 2-5**.




E MO-1
F MO-2
G CHRF-1
H CHRF-2
I K562-1
J K562-2
Bin 1
Bin 2
Bin 3
Bin 4
Fraction of peaks in MO−1 that overlaps with peaks from
Fraction of peaks in MO−2 that overlaps with peaks from
Fraction of peaks in CHRF−1 that overlaps with peaks from
Fraction of peaks in CHRF−2 that overlaps with peaks from
Fraction of peaks in K562−1 that overlaps with peaks from
Fraction of peaks in K562−2 that overlaps with peaks from
0.0
0.2
0.4
0.6
0.8
1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
Log_{10}(Peak score)
MK−1
MK−2
EB−1
EB−2
MO−1
MO−2
CHRF−1
CHRF−2
K562−1
K562−2

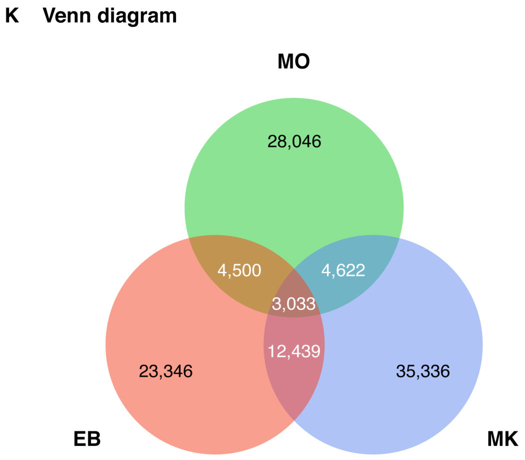


Figure 4-2. Overlap of NDRs across primary cells and their representative cell lines. The peak score distribution and 'intensity bins' for all studied cell types are shown (A–J). The top and bottom one percentile of the peak score distribution was excluded. The remaining peaks were stratified according to their normalised score into four equally sized intensity bins. As the peaks in Bin 1 showed limited overlap with the peak set of the same cell type obtained in the other individual/preparation, we focused on the top three bins (Bins 2–4) for biological interpretation. (K) Venn diagram of the overlap of FAIRE peaks (all bins considered) across the primary cell types using the 'stringent' peak calling cut-off.

We first investigated to what extent individual myeloid cell types have distinct open chromatin signatures, and whether NDRs of different peak score have different functional properties. As the next-generation DNA sequencing platform provides a large dynamic range and high sensitivity using discrete, digital sequencing read counts, we hypothesised that a sub-classification of FAIRE peaks based on peak score may allow more precise downstream functional analyses.

Pearson's correlation coefficients were calculated for peak scores between biological replicates/individuals in independent FAIRE experiments (**Figure 4-3**). Note that only overlapping peaks were considered for correlation analyses. The correlation coefficients between biological replicates were 0.66 and 0.89 in K562 and CHRF-288-11 cells, respectively. Between individuals, the correlation coefficients were 0.50, 0.61 and 0.70 in MOs, EBs and MKs, respectively. These analyses indicated that peak score is reproducible across samples of the same cell type.

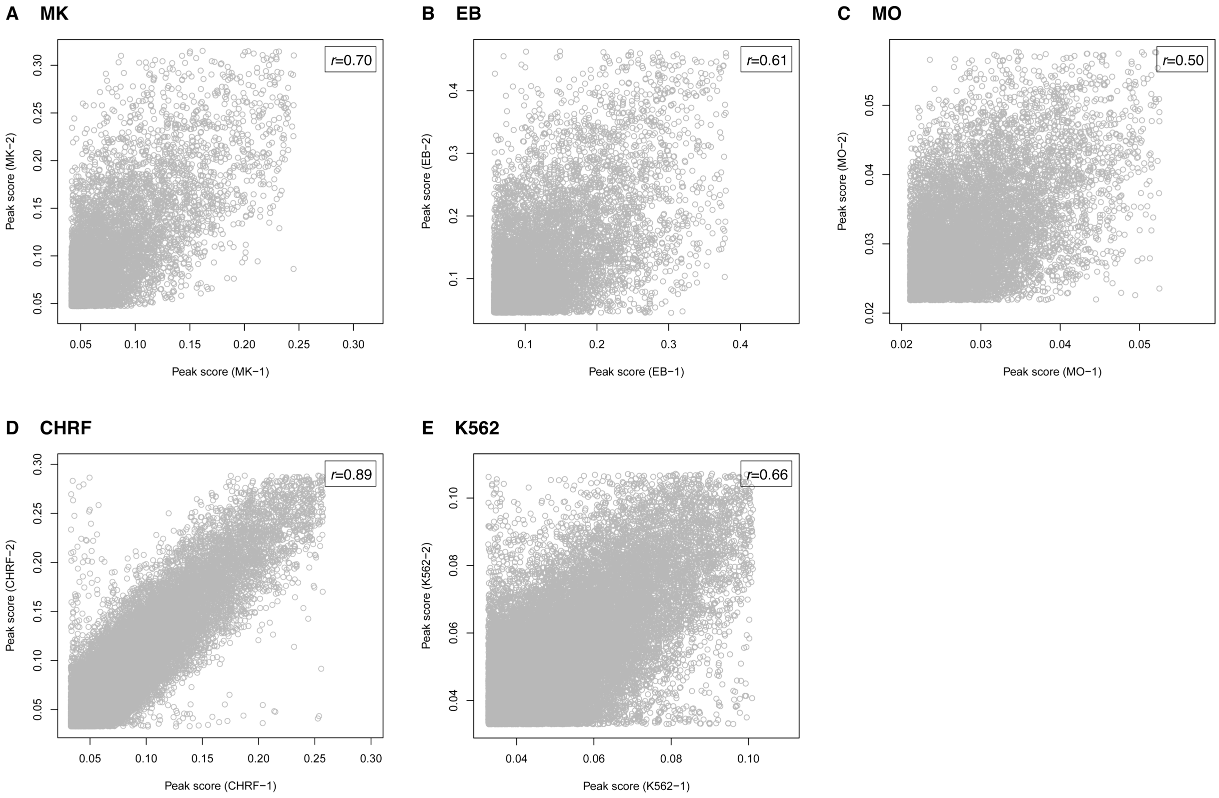


Figure 4-3. Pearson's correlation of peak score between independent FAIRE experiments. Correlation coefficients were calculated for scores of overlapping peaks between biological replicates (cell lines) or individuals (primary cells). Peaks unique to replicates/individuals were not included in these analyses.

We constructed distance matrices based on the overlap of peaks across cell types (**Figure 4-2**), and assessed the uncertainty of the clustering using bootstrap resampling (Suzuki & Shimodaira, 2006). We differentiated between peaks of higher peak score (represented in Bin 4; **Figure 4-4 A**) and lower score (represented in Bin 2; **Figure 4-4 B**). We did not consider the NDRs in Bin 1, as these displayed limited overlap between replicates/individuals and may be enriched for noise (**Figure 4-2**). Irrespective of the peak score, we found that the clustering of the primary cells based on the open chromatin profiles is dominated by cell type rather than individual. This reflects the corresponding haematopoietic lineage: MKs and EBs share a common cell progenitor, which in turn derives from a common myeloid progenitor that also gives rise to MOs (**Section 1.9**). We compared the open chromatin profile of MKs and EBs with the cell lines CHRF-288-11 and K562, respectively, which are commonly used as models for these primary cells. We found no co-clustering indicating that open chromatin structure of immortalised lines does not fully reflect that of primary cells (discussed in **Section 4.5**). This lack of co-clustering was more prominent when we only considered the high-intensity peaks (Bin 4) compared to low-intensity peaks (Bin 2). Many of the NDRs found in MKs are present in CHRF-288-11 cells; however, we identified a large number of additional NDRs in the CHRF-288-11 cell line that overlapped with K562 cells but not with MKs (**Figure 4-2**).

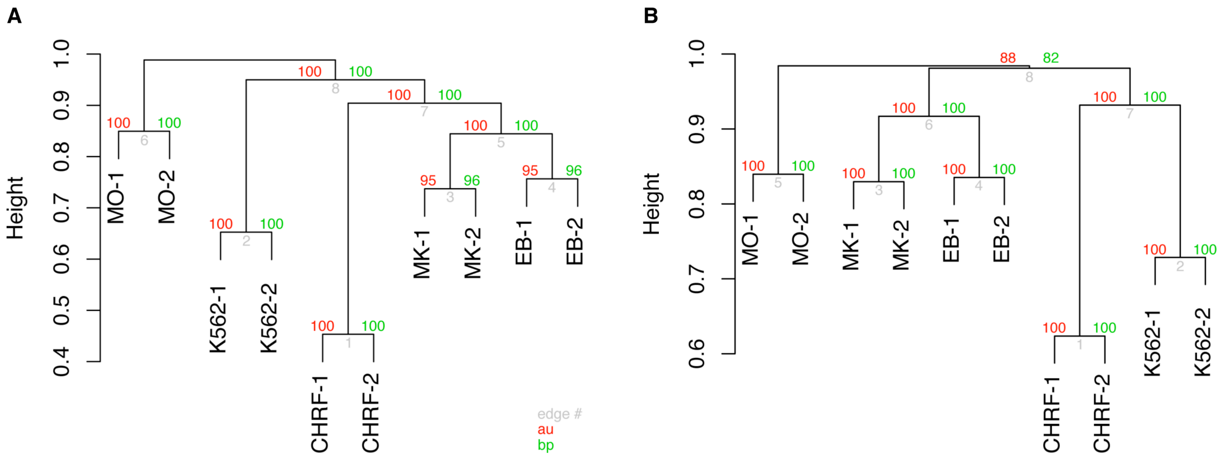


Figure 4-4. Characterisation of open chromatin profiles of primary human megakaryocytes (MKs), erythroblasts (EBs) and monocytes (MOs). Dendrogram of the hierarchical clustering of the overlap of FAIRE-derived NDRs across primary cells and immortalised cell lines. For assessing the relationship of the cell types, we used (A) peaks of higher peak score (represented in Bin 4) and (B) peaks of lower score (represented in Bin 2). The hierarchical cluster analysis was performed using the R package Pvclust (Suzuki & Shimodaira, 2006) (distance: binary; cluster method: complete). The uncertainty of the clustering was assessed using bootstrap resampling. Abbreviations: au: approximately unbiased *P*-value; bp: bootstrap probability value.

Next, for each cell type I pooled the sequence fragments of the two replicates/individuals, and processed the data as described above. I applied the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010) to aid the functional interpretation of NDRs of different scores by analysing the annotations of the single closest flanking gene (**Appendix, Table 8-2**). In MKs and EBs, NDRs in Bin 2 and 4 were enriched in cell type-specific and housekeeping genes, respectively. In contrast, in MOs I observed an enrichment of cell type-specific gene sets proximal to NDRs of both Bins 2 and 4 (**Appendix, Table 8-2**). One possible explanation for this difference could be that mature MOs, as studied here, do not proliferate (Geissmann et al., 2010).

I examined the location of NDRs of different bins relative to promoter regions, i.e. within 5 kb upstream of TSSs. I observed that for both MKs and EBs, but not MOs, the NDRs in Bin 4 were more often located at promoter regions than NDRs in Bin 2 (**Figure 4-5**).

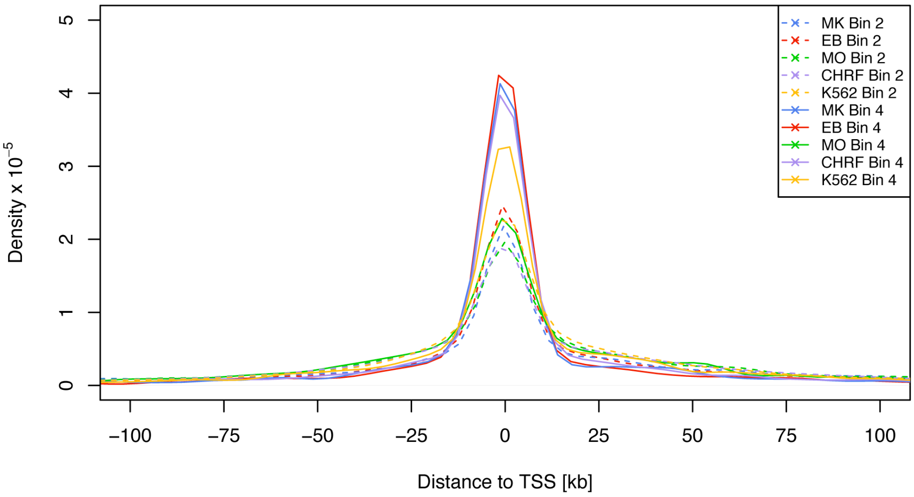


Figure 4-5. Distance of NDRs to the closest TSSs. The genomic distances between FAIRE peaks and transcription start sites were exported from GREAT (McLean et al., 2010). The density graph was plotted with a bandwidth of 5,000.

In MKs and EBs, we further investigated these observations by performing ChIP combined with high-throughput next-generation sequencing of the histone modifications H3K4me3 and H3K4me1, marking active promoters and enhancers, respectively (**Figure 4-6**). In both cell types, NDRs of higher score overlapped with gene promoters proximal to TSSs, whereas NDRs of lower score overlapped with enhancer elements distal to the closest TSS. NDRs that did not overlap with histone marks were more likely to be low-scoring and far from promoters.

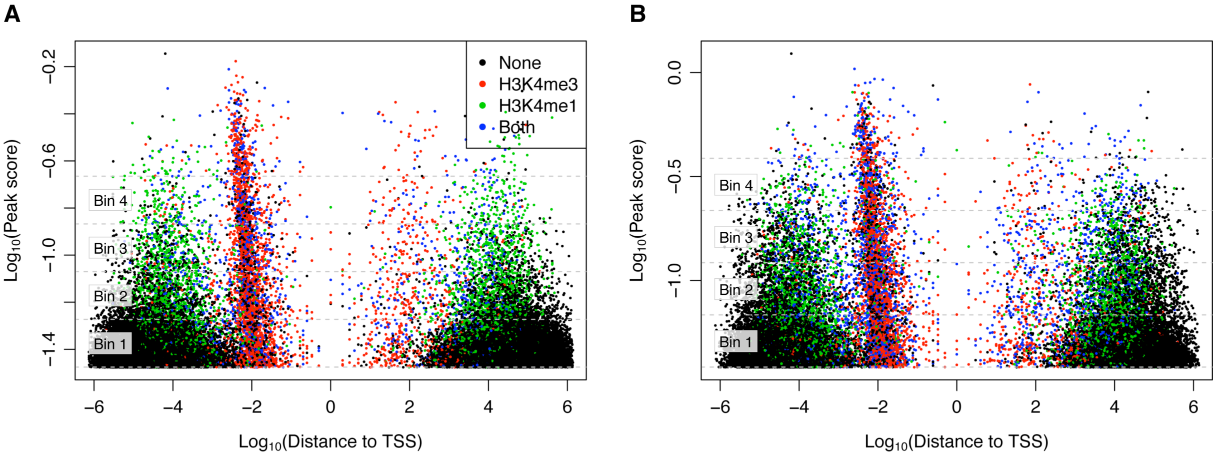


Figure 4-6. Overlap of H3K4me3 (promoter) and H3K4me1 (enhancer) histone marks with NDRs identified in (A) MKs and (B) EBs with respect to NDR score and distance to the closest TSS. The peak bins are indicated with a dashed grey line. In MKs, we identified 79,049 and 17,402 regions of enrichment of H3K4me1 and H3K4me3, respectively. In EBs, 66,410 and 16,871 H3K4me1 and H3K4me3 peaks were identified, respectively.

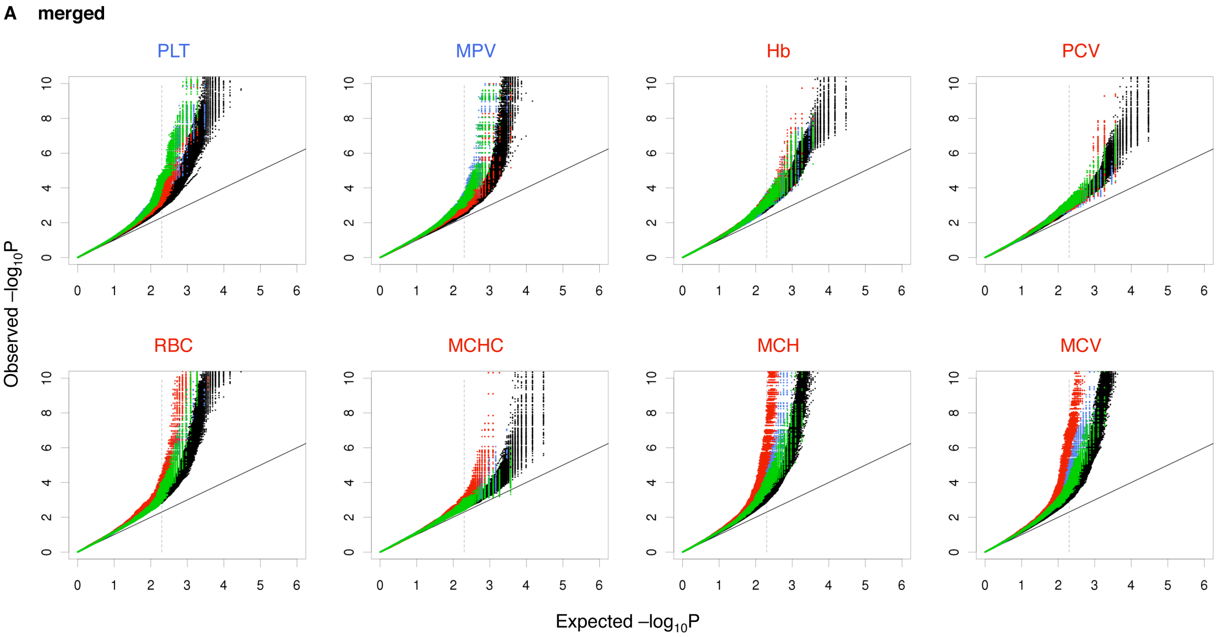
We then examined if cell type-restricted NDRs mark lineage-specific elements involved in regulation of expression of genes relevant to blood cell lineage commitment. We assessed the expression levels of the single closest gene to each lineage-specific NDR, interrogated over several time points during *in vitro* differentiation of HPCs into MKs and EBs (**Table 4-1**) (Gieger et al., 2011). In MKs, these transcripts were more likely to be upregulated during MK differentiation relative to all expressed transcripts (fold enrichment: 1.13; P=2.01x10^{-30}, two-tailed chi-squared test). In EBs, we observed the same effect directionality for transcripts during EB differentiation (1.20; P=1.37x10^{-61}). Transcripts close to MO-specific NDRs were downregulated during both MK and EB differentiation with a fold enrichment of 0.92 and 0.88, respectively. Interestingly, transcripts close to NDRs shared between MKs and MOs were also upregulated during MK differentiation (1.16; P=2.65x10^{-6}); I annotated the corresponding genes (n=382) using the Ingenuity Knowledge Base and found an enrichment of genes in the canonical pathway 'Fcγ receptor-mediated phagocytosis in macrophages and monocytes' (P=2.6x10^{-3}, Benjamini-Hochberg corrected for multiple testing; n=11 genes).

Table 4-1. Trends of up- or downregulation of genes close to FAIRE peaks during haematopoietic differentiation. Different classes of NDRs are defined based on their presence in different cell types. We report fold enrichment as the fraction of upregulated genes near NDRs of a given class versus the total fraction irrespective of NDR class. Gene expression was measured during haematopoietic differentiation into mature (A) MKs and (B) EBs.

Cell type(s) containing NDRs	Number genes ↓	Number genes ↑	P-val (two-tailed chi-squared test)	Fold enrichment
(A) Differentiation into megakaryocytes				
EB	3,490	3,909	7.64x10^{-02}	0.98
MK	2,777	4,282	2.01x10^{-30}	1.13
MK/EB	1,294	1,631	3.90x10^{-02}	1.04
MK/MO	274	457	2.65x10^{-06}	1.16
MK/MO/EB	1,093	1,242	5.17x10^{-01}	0.99
MO	5,611	5,553	2.59x10^{-18}	0.92
MO/EB	368	326	2.75x10^{-04}	0.87
(B) Differentiation into erythroblasts				
EB	3,209	4,190	1.37x10^{-61}	1.20
MK	3,928	3,131	7.19x10^{-06}	0.94
MK/EB	1,487	1,438	2.03x10^{-02}	1.05
MK/MO	435	296	4.06x10^{-04}	0.86
MK/MO/EB	1,180	1,155	1.80x10^{-02}	1.05
MO	6,520	4,644	1.67x10^{-30}	0.88
MO/EB	357	337	4.17x10^{-01}	1.03

4.3. Cell type-specific enrichment of haematological trait-associated SNPs in NDRs

We assessed the enrichment of genetic associations with platelet and erythrocyte phenotypes in NDRs in a cell type-specific context, using data from two meta-analyses of platelet count and volume (Gieger et al., 2011), and red blood cell parameters (van der Harst et al., 2012). These GWA meta-analyses are the largest conducted so far for these traits. Notably, nearly three-quarters of the 143 identified GWA signals are located at non-coding genomic regions (Gieger et al., 2011; van der Harst et al., 2012). We quantified the enrichment of *P*-values below 0.005 in cell type-specific NDRs using the genome-wide distribution of *P*-values as the baseline (**Section 2.6**). This analysis therefore explicitly considers the contribution of potential regulatory sequence variants in NDRs that do not reach the threshold of genome-wide significance (P=5x10^{-8}). As before, we excluded the NDRs in Bin 1. We calculated bootstrapped *P*-value distributions of sequence variants imputed from the 1000 Genomes Project data set (The 1000 Genomes Project Consortium, 2010) (**Figure 4**-7). We defined three classes of NDRs: merged (determined from pooled sequence fragments of two individuals), intersected (called independently in two individuals) and cell type-specific (based on the merged NDR set). The definition of these different NDR classes allows for stratification of the different properties of each NDR class.



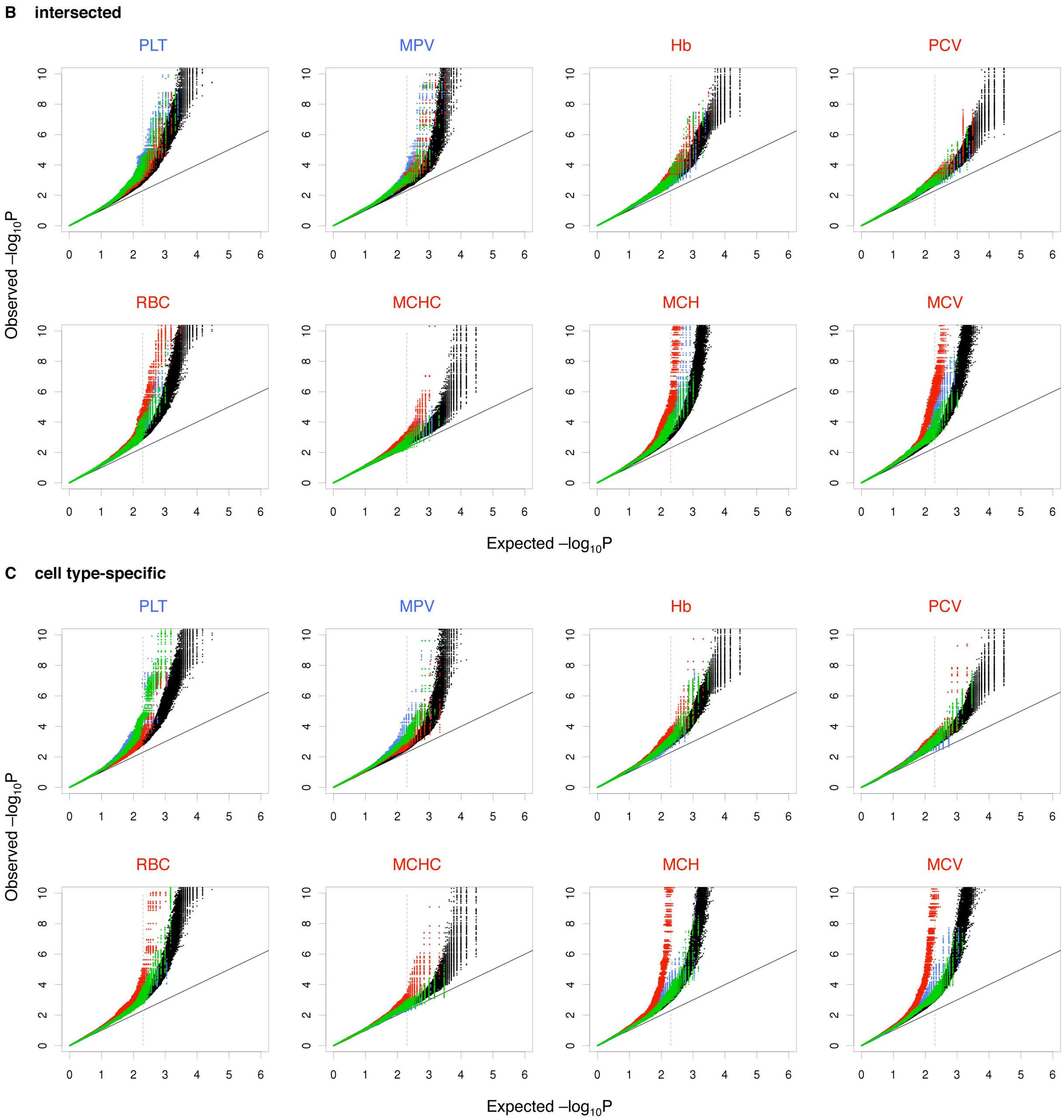


Figure 4-7. Bootstrapped quantile-quantile distributions. Data points in black represent the distribution of *P*-values for all 2.6x10^6 imputed SNPs. Enrichment values in the primary cell types shown in **Figure 4-8** are relative to this distribution. Enrichments are shown for three classes of NDRs: (**A**) merged (determined from pooled sequence fragments of two individuals), (**B**) intersected (called independently in two individuals) and (**C**) cell type-specific (based on the merged NDR set). Data points: red: EB; blue: MK; green: MO. Abbreviations: PLT: platelet count; MPV: mean platelet volume; Hb: haemoglobin; PCV: packed cell volume; RBC: red blood cell count; MCHC: mean cell haemoglobin concentration; MCH: mean cell haemoglobin; MCV: mean cell volume.

For the merged NDR set (**Figure 4-8 A**), we found that SNPs associated with erythrocyte traits consistently showed the strongest enrichment in NDRs in EBs, although weak enrichment was also seen in NDRs in MKs and MOs. Platelet trait associations were enriched in NDRs in all three studied cell types. For both platelet count and volume, the enrichment in NDRs in MKs was stronger than in EBs, but surprisingly was also marked in NDRs in MOs. For the intersected NDR set (**Figure 4-8 B**), a set of high-confidence NDRs, the strongest enrichment for platelet trait associations was found in NDRs in MKs, and for erythrocyte traits in NDRs in EBs. This trend was not observed in the merged and cell type-specific NDR sets. For cell type-specific NDRs (**Figure 4-8 C**), we found strong enrichment of SNPs associated with all of the six erythrocyte traits in EB-specific NDRs.

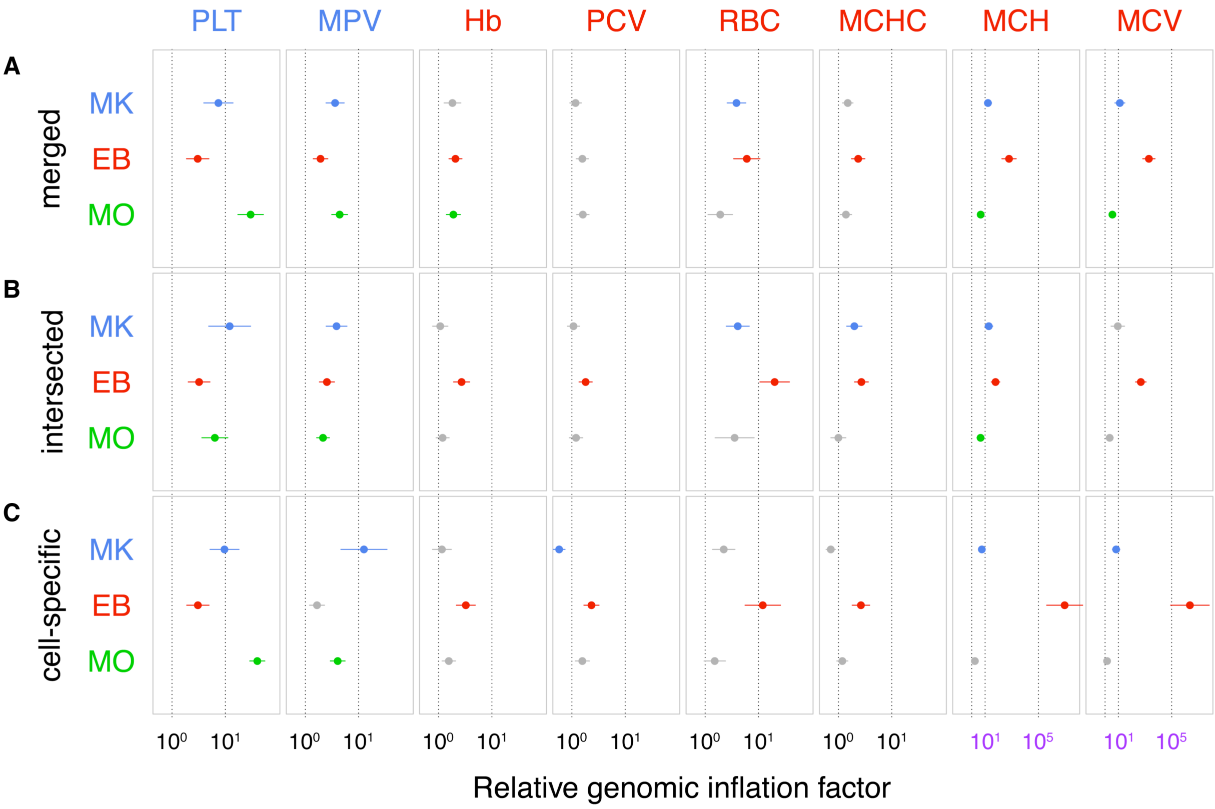


Figure 4-8. **Enrichment of associations with platelet and erythrocyte phenotypes in NDRs across quantitative haematological traits, cell types and NDR classes.** For each trait, the enrichment of associations (genomic inflation) in MKs, EBs and MOs are indicated. Enrichments are shown for three classes of NDRs: (A) merged (determined from pooled sequence fragments of two individuals), (B) intersected (called independently in two individuals) and (C) cell type-specific (based on the merged NDR set). The enrichment was quantified as relative genomic inflation at the 0.005 quantile, i.e. the ratio of the *P*-value at the 0.005 quantile of the SNPs overlapping NDRs in one of the three cell types and the *P*-value at the 0.005 quantile of the full set of 2.6x10^6 imputed SNPs (Figure 4-7). This controlled for population stratification. Error bars indicate standard deviations (s.d.). Grey data points represent non-significant enrichment (the mean is within 2 s.d. of zero). The dotted vertical lines at 10^0 indicate no enrichment at the NDRs for a given trait. Abbreviations: PLT: platelet count; MPV: mean platelet volume; Hb: haemoglobin; PCV: packed cell volume; RBC: red blood cell count; MCHC: mean cell haemoglobin concentration; MCH: mean cell haemoglobin; MCV: mean cell volume.

Mean cell haemoglobin and mean red cell volume, which are highly correlated (*r*=0.91; **Table 4-2**), showed substantially stronger enrichment compared to the other four erythrocyte traits, suggesting that these traits may be governed by processes that are regulated at an intracellular level. The strong enrichment of platelet trait (particularly platelet count) associations in MO-specific NDRs may indicate a role for MOs in influencing the platelet phenotype. Alternatively, it could reflect the role of cell types not studied in this work that share these NDRs.

Table 4-2. Pearson's correlation coefficients between erythrocyte traits. Full details of these analyses are reported in van der Harst et al., 2012. Platelet count and volume are negatively correlated, with Pearson's *r*=-0.49 (gender-adjusted) (Gieger et al., 2011). Abbreviations: Hb: haemoglobin; MCH: mean cell haemoglobin; MCHC: mean cell haemoglobin concentration; MCV: mean cell volume; PCV: packed cell volume; RBC: red blood cell count.

	Hb	MCH	MCHC	MCV	PCV	RBC	Number of associated loci
Hb	1.00	0.23	0.08	0.22	0.96	0.75	11
MCH		1.00	0.46	0.91	0.09	0.47	19
MCHC			1.00	0.07	0.21	0.25	8
MCV				1.00	0.20	0.42	23
PCV					1.00	0.80	4
RBC						1.00	10
							75

To shed light on the properties of the genes closest to MO-specific NDRs that contained a platelet count-associated SNP ($P<10^{-4}$; n=61 genes), I performed canonical pathway analyses using the Ingenuity Knowledge Base. I detected a modest enrichment of genes involved in 'haematological system development and function' (range, P=4.62x10^{-2}–9.94x10^{-2}, Benjamini-Hochberg corrected for multiple testing; n=7 genes) and 'cell-to-cell signalling and interaction' (P=4.62x10^{-2}–9.94x10^{-2}; n=9). Notably, these genes included *THBS1* (encoding thrombospondin 1) and *WASL* (Wiskott-Aldrich syndrome-like), which have a role in the activation of blood platelets (Dorahy et al., 1997; Falet et al., 2002).

In order to identify candidate functional SNPs underlying platelet and erythrocyte QTLs, I intersected the composite map of open chromatin (Bins 1–4) obtained in each cell type, with the GWA lead SNPs ($P<5x10^{-8}$ from Phase II HapMap) and their proxies ($r^2>0.8$ in the 1000 Genomes Project data set) at the 68 platelet and 75 erythrocyte QTLs previously described (Gieger et al., 2011; van der Harst et al., 2012). For this analysis, Bin 1 was included in the analysis (**Table 2-5**). As any potential functional candidate SNPs were to be functionally characterised in downstream analyses, I thought to retain as many variants in NDRs as possible in this first step. Using these criteria, I retrieved 1,680 and 4,632 SNPs at platelet and erythrocyte QTLs, respectively. At 25 (37.3%) and 31 (41.3%) of the platelet and erythrocyte QTLs, respectively, I found at least one trait-associated SNP located within an NDR across the three cell types (**Figure 4-9 A,B**; **Appendix, Table 8-3**).

At platelet QTLs, significant (P<5x10^{-6}) overlap with NDRs in MKs and MOs was observed (albeit only when the top ranking peaks were considered). The extent of overlap with NDRs in EBs was not more than expected by chance when compared to 100,000 sets of SNPs. These were matched for number of loci identified in each trait and allele frequency, and augmented with proxy SNPs (hereafter termed 'random loci') (**Figure 4-9 C**). At erythrocyte QTLs, we found significant (P<5x10^{-6}) overlap with NDRs in EBs, but not with NDRs in MKs or MOs (**Figure 4-9 D**). When compared with immortalised cell lines representing MK and EB lineages, the same trends of enrichment were observed in the relevant trait.

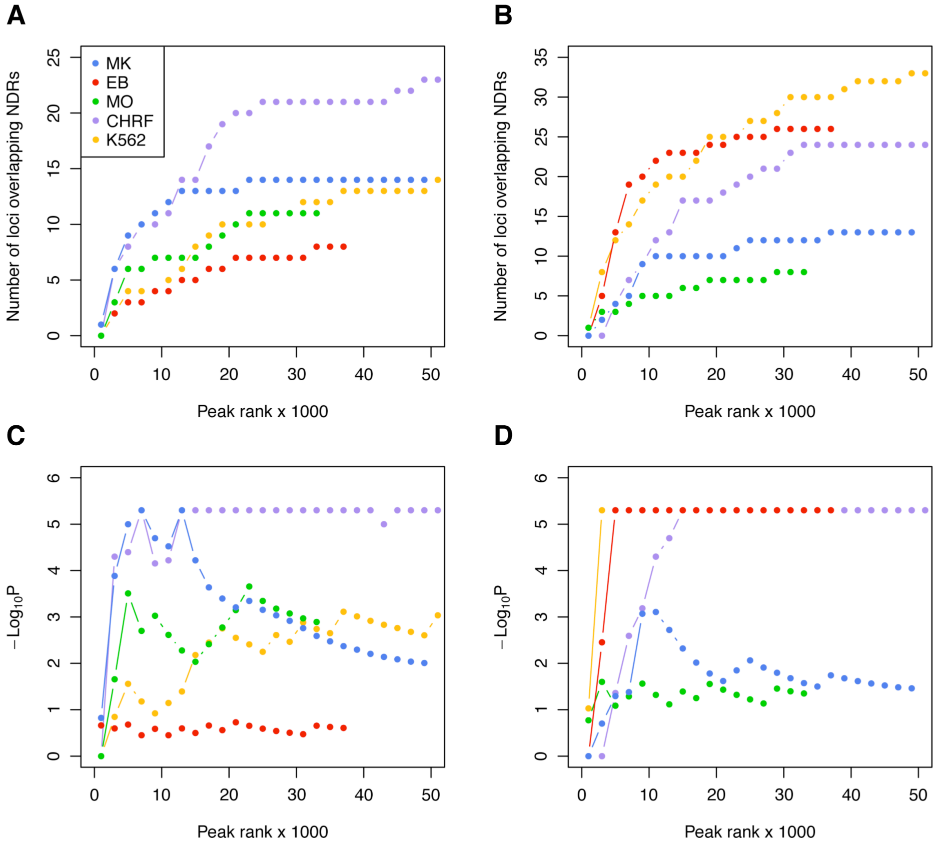


Figure 4-9. Genome-wide significant signals associated with platelet and erythrocyte phenotypes at sites of open chromatin in primary cells and immortalised cell lines. Number of GWA loci harbouring (A) platelet- and (B) erythrocyte trait-associated SNPs in NDRs in MKs, EBs, MOs, CHRF-288-11 and K562 cells. The strongest enrichment of genome-wide significant sequence variants at (C) platelet and (D) erythrocyte QTLs was found in NDRs in MKs and EBs. However, the enrichment was

equally clear in NDRs in the megakaryocytic cell line CHRF-288-11 and erythroblastoid cell line K562, respectively.

It is important to note that about half of the overlaps would be expected by chance. This was determined by calculating the enrichment of the number of loci with at least one overlap with an NDR, relative to random loci when all peaks are considered (**Figure 4-10**). However, the statistical enrichment suggests that intersection of trait-associated SNPs with NDRs is likely to provide an informative ranking for selection of candidate variants for functional follow-up experiments.

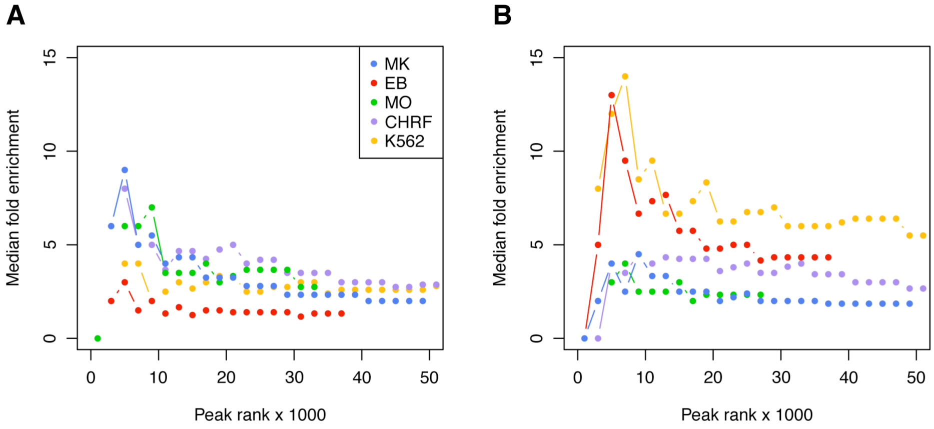


Figure 4-10. Fold enrichment of the number of loci with at least one overlap with an NDR compared to the median of 100,000 random sets of loci. The enrichment is shown at (A) platelet and (B) erythrocyte loci. The extent of overlap with NDRs in each cell type at the association loci was assessed by comparison to random loci. Increasing numbers of FAIRE peaks ranked by peak score were overlapped with the candidate SNPs, in order to examine the effect of differences in peak calling thresholds across multiple cell types. Note that the enrichment is high for high-scoring peaks and settles at about 2-fold enrichment providing an upper bound of the enrichment expected by chance when irrelevant sets of SNPs are intersected.

We investigated whether maps of open chromatin can be used to retrieve trait-associated sequence variants below the genome-wide significance threshold, without increasing the expected number of false positive associations. For these analyses, the false discovery rate (FDR) for SNPs in cell type-specific NDRs was estimated as a function of the genome-wide significance level. Specifically, the FDR was estimated as the ratio of the expected number of SNPs from the null and the observed number of SNPs that are located in a cell type-specific NDRs and have a *P*-value below the genome-wide significance threshold. The expected number of SNPs from the null was estimated by the product of the total number of SNPs in NDRs for a given cell type, regardless of the association *P*-value. Indeed, the FDR for mean red cell volume-associated SNPs in EB-specific NDRs was lower than the genome-wide average (**Figure 4-11**). This suggests that the maps of open chromatin make it possible to consider variants below the genome-wide significance threshold, without increasing the expected number of false positive associations.

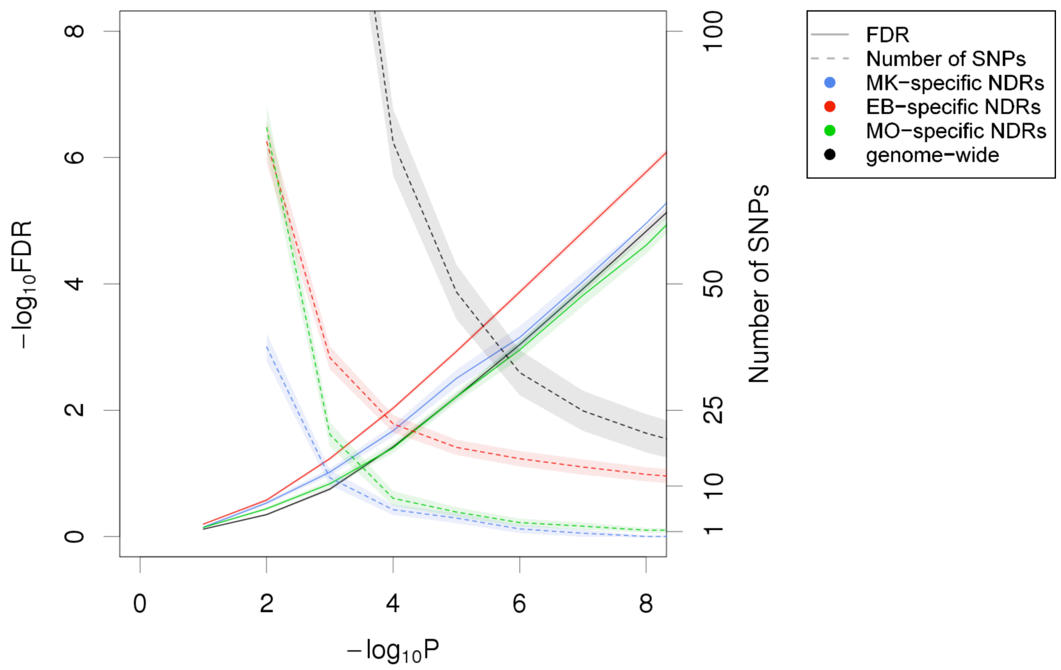
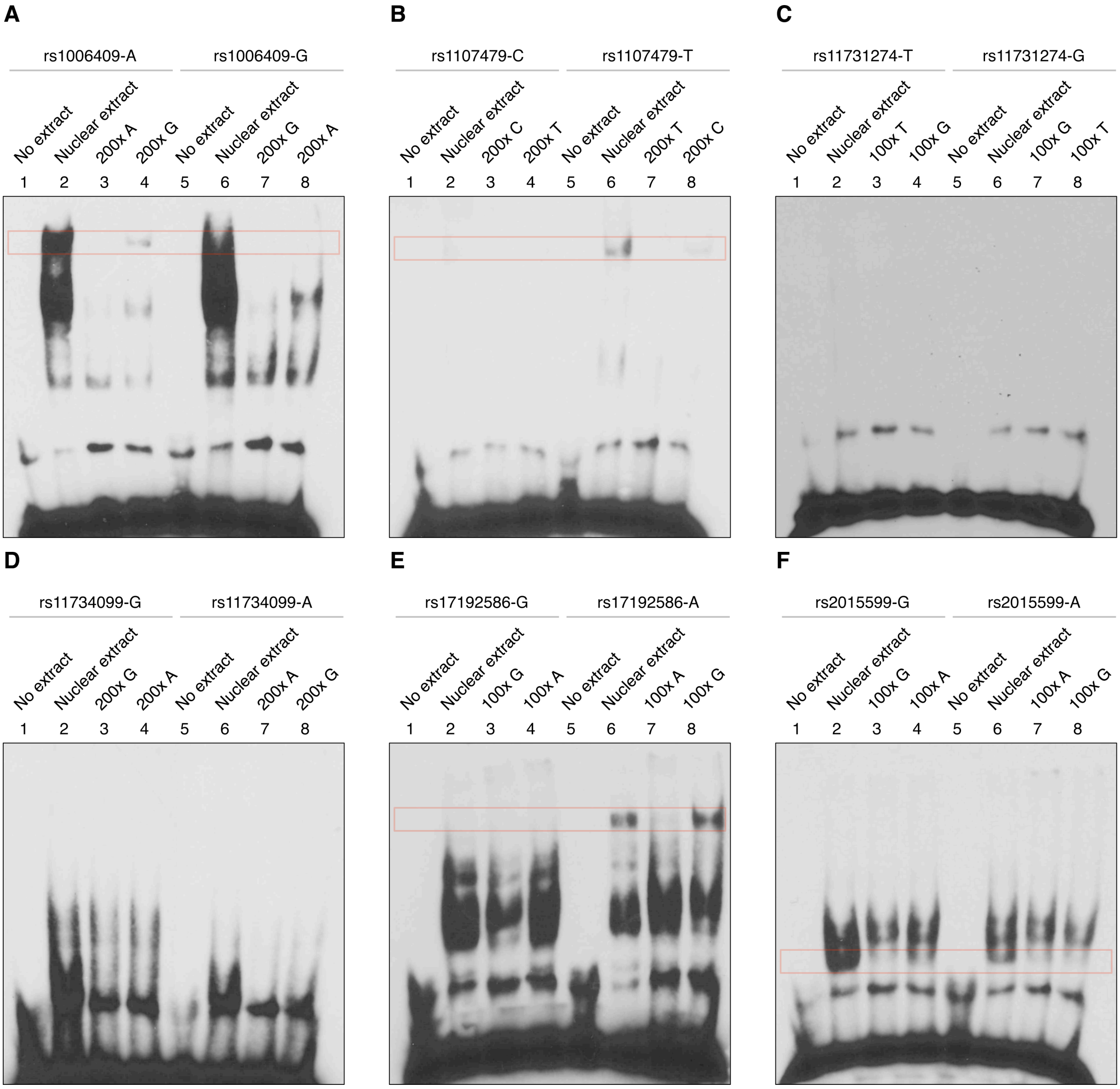


Figure 4-11. Estimated false discovery rates (FDRs) for mean red cell volume-associated SNPs in cell type-specific NDRs. The FDR for SNPs (solid lines) in cell type-specific NDRs (left y-axis) was estimated as a function of the significance level (x-axis). The number of SNPs (dashed lines) associated at a given significance level is shown on the right y-axis. The plot shows that the FDR for SNPs in EB-specific NDRs (solid red line) was lower than the genome-wide average (solid black line). For the genome-wide FDR estimate, all SNPs (both within and outside NDRs) were used. This gives the baseline FDR estimate to which the FDR estimates for SNPs in cell type-specific NDRs were compared.

4.4. Identification of candidate functional SNPs at platelet QTLs

To provide evidence that the SNPs we identified using the above approach are indeed valid functional candidates, I performed electrophoretic mobility shift assays (EMSAs) in nuclear extracts from the cell line CHRF-288-11. EMSAs are based on the principle that DNA-protein complexes migrate more slowly than non-bound DNA in a native polyacrylamide or agarose gel, resulting in a ‘shift’ in migration of the labelled DNA band. To control for differences between primary cells and cells from an immortalised line, I selected all platelet trait-associated SNPs (n=16) in NDRs found in both MKs and CHRF-288-11 cells (n=13). Importantly, 8 of these 13 NDRs also coincided with binding sites of transcription factors key in regulating megakaryopoiesis (Tijssen et al., 2011), namely FLI1, GATA1, GATA2, RUNX1 and SCL (also known as TAL1), suggesting physiologically relevant regulatory elements (**Table 4-3**). For 10 of the 16 platelet trait-associated SNPs, I observed by visual inspection of the blot, differential nuclear protein binding between alleles in EMSA studies. For the remaining 6 SNPs, I observed either comparable protein binding between allelic probes or no binding at all (**Figure 4-12**).



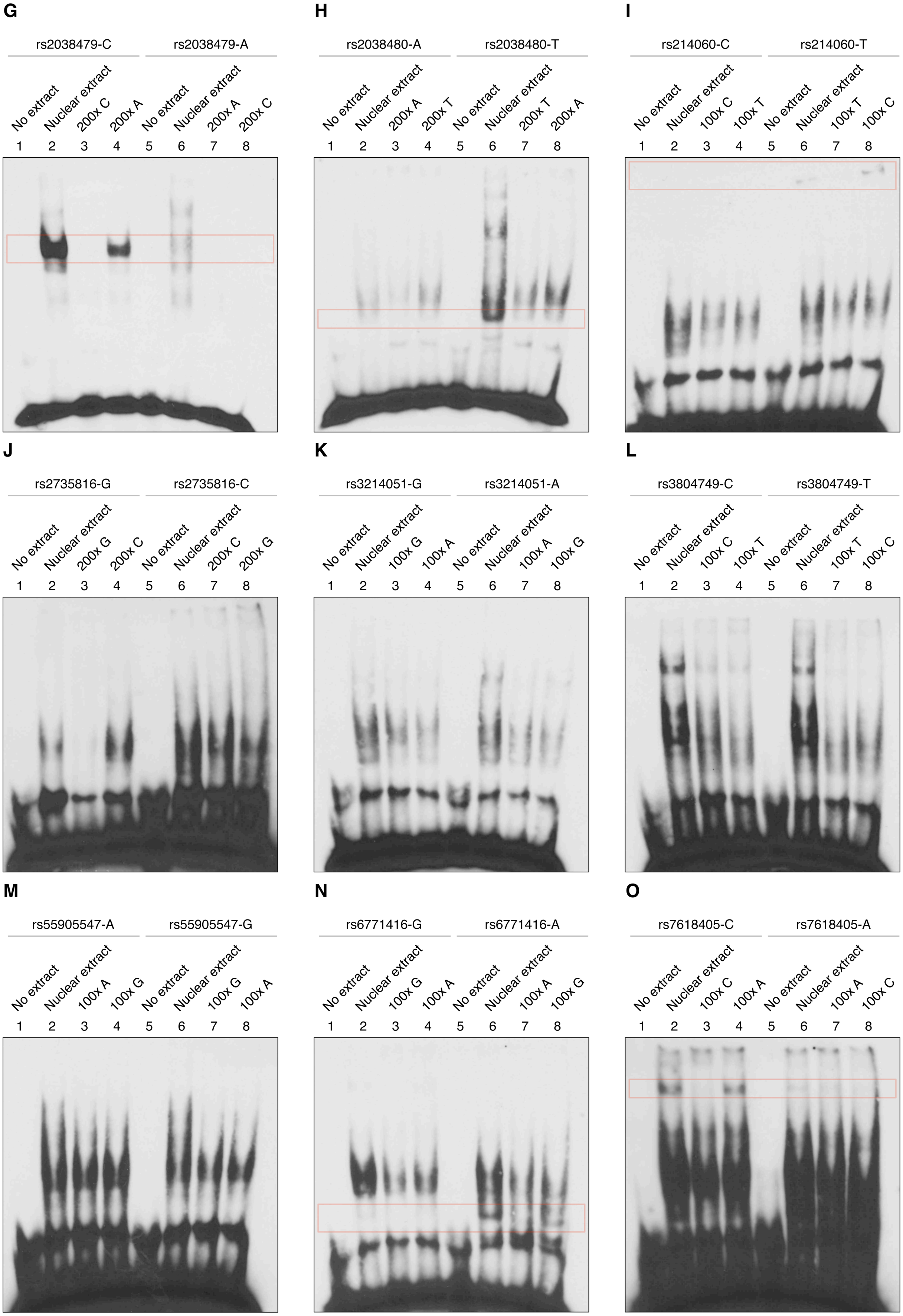

G
rs2038479-C
rs2038479-A
No extract
Nuclear extract
200x C
200x A
No extract
Nuclear extract
200x A
200x C
1 2 3 4 5 6 7 8
H
rs2038480-A
rs2038480-T
No extract
Nuclear extract
200x A
200x T
No extract
Nuclear extract
200x T
200x A
1 2 3 4 5 6 7 8
I
rs214060-C
rs214060-T
No extract
Nuclear extract
100x C
100x T
No extract
Nuclear extract
100x T
100x C
1 2 3 4 5 6 7 8
J
rs2735816-G
rs2735816-C
No extract
Nuclear extract
200x G
200x C
No extract
Nuclear extract
200x C
200x G
1 2 3 4 5 6 7 8
K
rs3214051-G
rs3214051-A
No extract
Nuclear extract
100x G
100x A
No extract
Nuclear extract
100x A
100x G
1 2 3 4 5 6 7 8
L
rs3804749-C
rs3804749-T
No extract
Nuclear extract
100x C
100x T
No extract
Nuclear extract
100x T
100x C
1 2 3 4 5 6 7 8
M
rs55905547-A
rs55905547-G
No extract
Nuclear extract
100x A
100x G
No extract
Nuclear extract
100x G
100x A
1 2 3 4 5 6 7 8
N
rs6771416-G
rs6771416-A
No extract
Nuclear extract
100x G
100x A
No extract
Nuclear extract
100x A
100x G
1 2 3 4 5 6 7 8
O
rs7618405-C
rs7618405-A
No extract
Nuclear extract
100x C
100x A
No extract
Nuclear extract
100x A
100x C
1 2 3 4 5 6 7 8

Figure 4-12. Electrophoretic mobility shift assays (EMSAs) for platelet candidate functional SNPs. The following 16 platelet candidate functional SNPs at 12 NDRs found in both CHRF-288-11 megakaryocytic cells and primary MKs were tested in EMSAs: (A) rs1006409, (B) rs1107479, (C) rs11731274, (D) rs11734099, (E) rs17192586, (F) rs2015599, (G) rs2038479, (H) rs2038480, (I) rs214060, (J) rs2735816, (K) rs3214051, (L) rs3804749, (M) rs55905547, (N) rs6771416 and (O) rs7618405. Probes harbouring either the reference or alternative allele of the candidate functional SNP are shown in lanes 1–4 and lanes 5–8, respectively. For competition assays, specific (lanes 3 and 7) and unspecific (lanes 4 and 8) unlabelled probes were added in a 100- or 200-fold molar excess over the labelled probes (lanes 2 and 6). Red rectangles indicate differential protein binding of the EMSA probes containing the reference and alterative allele of the candidate functional SNP.

I then annotated the 16 SNPs using the RegulomeDB, a database containing known and predicted regulatory elements in the human genome, and found 15 SNPs to coincide with at least one RegulomeDB feature (**Table 4-3**). Finally, I performed *in silico* transcription factor binding site analysis using the transcription factor affinity prediction (TRAP) method (Thomas-Chollier et al., 2011) and found 15 SNPs that affect a transcription factor binding motif (**Appendix, Table 8-4**). Of the 10 SNPs that showed differential protein binding, all but one overlapped with a RegulomeDB feature and affected a predicted transcription factor binding site.

Table 4-3. Summary of the functional evidence obtained for platelet candidate functional SNPs through FAIRE, ChIP and EMSA experiments, as well as the RegulomeDB. [a]The marked SNPs were located in the same NDR at the indicated GWA locus. Values reported for binding in EMSA studies represent the mean signal density ratios of the CHRF-288-11 nuclear protein binding to EMSA probes containing either allele of the candidate SNP. The allele contained in the probe with the stronger nuclear protein binding is reported in parentheses. RegulomeDB scores were retrieved from http://www.regulomedb.org/help#score; Abbreviations: Ref: reference allele; Alt: alternative allele; TF: transcription factor.

Candidate functional SNP			NDR cell type (Bin)	GATA1/2, SCL, RUNX1 or FLI1 binding site in MKs	Binding in EMSA	RegulomeDB Annotation	
ID	*Ref/alt*	*GWA locus*				*Score*	*Supporting data*
rs1006409[a]	A/G	*MLSTD1*	MK (2)	–	1.169 (Ref)	2B	TF binding + any motif + DNase footprint + DNase peak
rs2015599[a]	G/A	*MLSTD1*	MK (2)	–	1.459 (Ref)	4	TF binding + DNase peak
rs1107479	C/T	*PTGES3-BAZ2A*	MK (4)/EB (4)/MO (1)	–	1.111 (Alt)	1F	eQTL + TF binding or DNase peak
rs3214051	G/A	*PTGES3-BAZ2A*	MK (4)/EB (4)/MO (3)	FLI1	equal	4	TF binding + DNase peak
rs11731274[a]	T/G	*KIAA0232*	MK (1)	GATA1 + GATA2 + RUNX1 + SCL	none	4	TF binding + DNase peak
rs11734099[a]	G/A	*KIAA0232*	MK (1)	GATA1 + GATA2 + RUNX1 + SCL	equal	4	TF binding + DNase peak
rs17192586	G/A	*RAD51L1*	MK (3)/EB (1)	RUNX1	2.167 (Alt)	4	TF binding + DNase peak
rs2038479[a]	C/A	*DNM3*	MK (3)	–	3.353 (Ref)	5	TF binding or DNase peak
rs2038480[a]	A/T	*DNM3*	MK (3)	–	1.444 (Alt)	5	TF binding or DNase peak
rs214060	C/T	*LRRC16*	MK (3)	–	1.216 (Alt)	4	TF binding + DNase peak
rs2735816	G/C	*BRF1*	MK (1)	–	equal	5	TF binding or DNase peak
rs3804749	C/T	*PDIA5*	MK (4)	SCL	equal	5	TF binding or DNase peak
rs4148450	C/T	*ABCC4*	MK (2)	RUNX1	2.943 (Alt)	4	TF binding + DNase peak
rs55905547	A/G	*CTSZ-TUBB1*	MK (3)	GATA1 + SCL	equal	–	–
rs6771416	G/A	*KALRN*	MK (2)/EB (1)	GATA1 + SCL	2.249 (Alt)	2B	TF binding + any motif + DNase footprint + DNase peak
rs7618405	C/A	*SATB1*	MK (4)/MO (1)	GATA1 + RUNX1 + FLI1 + SCL	1.638 (Ref)	4	TF binding + DNase peak

As an example, rs4148450 (associated with platelet count) (Gieger et al., 2011) was located at an MK-specific intronic NDR of *ABCC4*. The open chromatin region coincided with a RUNX1 transcription factor binding site (**Figure 4-13**). *ABCC4* encodes the ATP-binding cassette (ABC) protein ABCC4, also known as multidrug resistance protein 4 (MRP4). Several studies indicated that ABCC4 is involved in the accumulation of the platelet-activating signalling molecule adenosine diphosphate (ADP) in platelet-dense granules (Jedlitschky et al., 2004; Jedlitschky et al., 2010). Our data suggested the non-coding SNP rs4148450 to be the functional variant at the 13q32.1 platelet count locus.

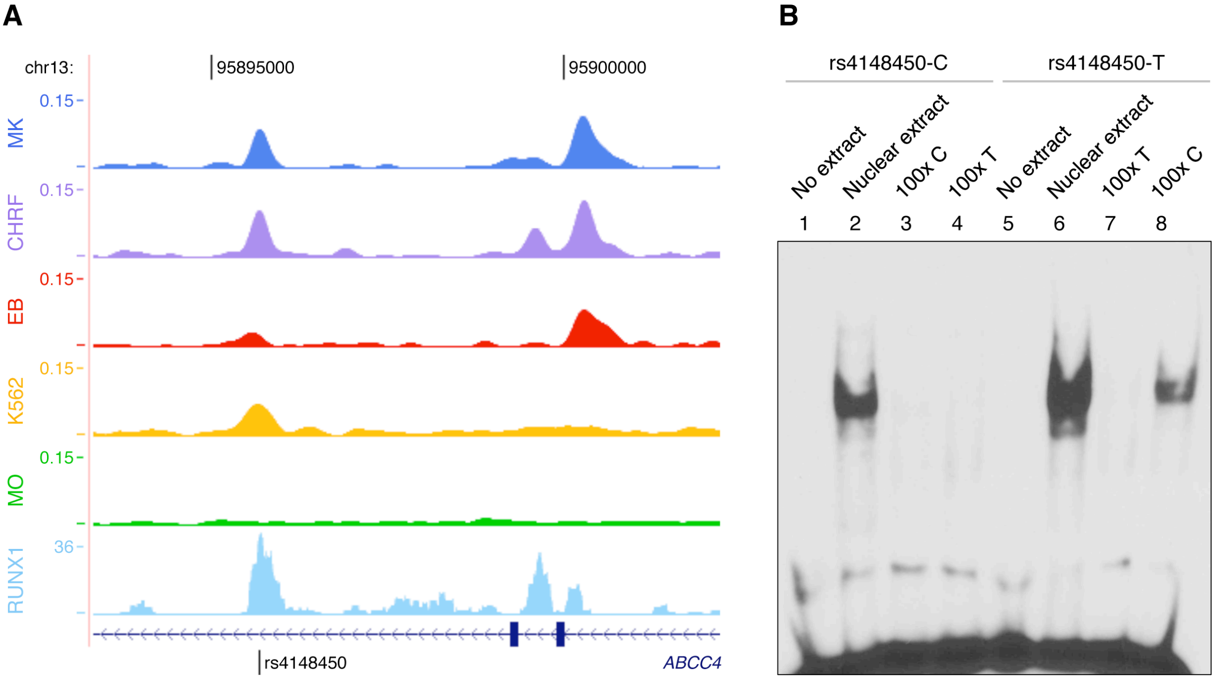


Figure 4-13. Functional follow-up of the *ABCC4* platelet count locus. (A) Coverage profiles of FAIRE-seq data in MKs, EBs, MOs, CHRF-288-11 and K562 cells. An open chromatin region found only in MKs and CHRF-288-11 cells contained the platelet count-associated SNP rs4148450 (r^2=1.0 with GWA lead SNP rs4148441). The SNP was contained within a RUNX1 transcription factor binding site in MKs (Tijssen et al., 2011). (B) Electrophoretic mobility shift assays in megakaryocytic cells showed differential nuclear protein binding to probes containing the C- and T-allele of rs4148450, where only the probes containing the T-allele were not competed by unspecific competitor probes. Further functional studies are required to elucidate the molecular mechanism underlying the *ABCC4* association locus.

4.5. Discussion

We generated genome-wide maps of open chromatin in primary human MKs, EBs and MOs and used these to define enrichment patterns of GWA signals of quantitative haematological traits in NDRs in a cell type-dependent manner. We showed that analyses in primary cells are important for biological interpretation of NDRs. Although immortalised cell lines are valuable tools for discovery of NDRs, there were clear differences in the chromatin structure compared to primary cells (**Figures 4-2**, **4-3** and **4-4**). These differences may arise through serial subculturing of immortalised cell lines, resulting in a more homogenous cell population. The primary cell cultures may be composed of a mixture of cell types, or of the same cell type but at different differentiation stages. We found that NDRs of different score (peak height) have different functional features, in particular their location relative to the TSS and overlap with different histone modification marks. Therefore, the sectioning of peaks into different intensity bins allows more precise downstream functional analyses.

The relative strength of the enrichment of association signals in NDRs highlighted distinct patterns across the phenotypic traits and cell types examined. These analyses allowed us to dissect the haematological trait associations in different potential effector cell types within the myeloid lineage. To provide further support to these findings, we are currently expanding the enrichment analyses using GWA signals of fasting glucose and FAIRE-seq data in human pancreatic islets.

There are likely to be many true association signals below the genome-wide significance threshold for well-powered studies such as the two large GWA meta-analyses we examined here (Gieger et al., 2011; van der Harst et al., 2012; **Appendix, Table 8-3**). We showed that considering trait-associated SNPs – both above and below the genome-wide significance threshold – located within NDRs, can enhance the ability to identify gene sets underlying processes relevant to the phenotype. Indeed, NDRs have the potential to reduce the false discovery rate for SNPs selected at a given threshold of significance (**Figure 4-11**). Integration of such variants in network analyses and subsequent functional studies may provide valuable biological insights. Importantly, the enrichment of associations in NDRs suggests that maps of open chromatin may be valuable for prioritising variants underlying association signals that are in high linkage disequilibrium. To quantify this, GWA signals not located in an NDR may be tested in EMSAs. GWA signals in NDRs that did not reach the genome-wide significance threshold may also be tested. As a substantial fraction of the overlaps with NDRs can be attributed to chance, integration of additional epigenetic marks will further improve the power of this approach to identifying functional variants at GWA loci.

We tested 16 candidate regulatory variants at 12 known platelet QTLs in EMSA studies, and provided evidence that the majority (62.5%) of the tested SNPs exerted their effect through disruption/introduction of protein binding sites. This suggests that the impact of trait-associated sequence variants on protein binding sites may prove to be a key molecular mechanism at non-coding regions. However, additional studies are required to establish the underlying molecular mechanisms through which these SNPs affect the platelet phenotype. In **Chapter 5**, I describe suitable strategies for establishing biological mechanism at GWA loci.

A more complete catalogue of chromatin profiles will be needed to address whether the candidate functional SNPs indeed have truly cell type-specific effects (i.e. out of all possible cell types). This can only be addressed by large collaborative efforts such as the ENCODE (The ENCODE Project Consortium, 2011), BLUEPRINT (Adams et al., 2012) and Roadmap Epigenomics Projects (Bernstein et al., 2010). Incorporation of these genome- and epigenome-wide data sets in a multitude of different primary cell types will greatly facilitate the functional interpretation of non-coding trait-associated SNPs in terms of effector cell type and underlying molecular mechanism.

CHAPTER 5

Functional follow-up of the platelet volume and function locus at chromosome 7q22.3.

This chapter is in parts based on the following publication:

Paul, D.S., et al. (2011). Maps of open chromatin guide the functional follow-up of genome-wide association signals: application to hematological traits. PLoS Genet. *7*, e1002139.

Collaboration note:

Section 5.2: DNA amplification by PCR and capillary sequencing in 643 individuals was performed at the Wellcome Trust Sanger Institute (WTSI) by the ExoSeq/ExoCan facility[1]. I designed the experiments and reviewed potential SNP positions.

Section 5.4: The platelet sample cohort of 24 healthy individuals was collected by Suthesh Sivapalaratnam[2], Hanneke Basart[2] and Mieke D. Trip[2]. Processing of raw expression and genotyping data was performed by Augusto Rendon[3–6]. Expression QTL data in macrophages and monocytes, as well as in LCLs, adipose and skin, were generated as part of the Cardiogenics and MuTHER consortia, respectively. I performed the eQTL analysis with help from Tsun-Po Yang[1].

Section 5.5: The breeding, maintaining and processing of the mice used in this study was performed by Katta Hautaviita[1], Jonna Tallila[1], Jacqui White[1], the staff from the WTSI's Research Support Facility (RSF) and Mouse Genetics Project. James P. Nisbet[1] performed total RNA extraction and whole-genome gene expression profiling. I processed and analysed the expression data, and performed functional ontology classification and canonical pathway analysis.

Section 5.6: The protein-protein interaction network of core proteins was created in collaboration with Stuart Meacham[3,4]. I retrieved the orthologous human genes and their respective proteins, and interpreted all results.

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; [2]Department of Vascular Medicine, Academic Medical Centre Amsterdam, Amsterdam, The Netherlands; [3]Department of Haematology, University of Cambridge, Cambridge, UK; [4]National Health Service (NHS) Blood and Transplant, Cambridge, UK; [5]Biostatistics Unit, Medical Research Council, Cambridge, UK; [6]NIHR Biomedical Research Centre, Cambridge, UK.

5.1. Introduction

In **Chapters 3** and **4**, I identified putative functional variants associated with haematological traits by intersecting trait-associated SNPs with maps of open chromatin in myeloid cell types. Among many, I identified an NDR harbouring rs342293 at chromosome 7q22.3 in the megakaryocytic cell line CHRF-288-11 ('MK cells') but not in the erythroblastoid cell line K562 ('EB cells') (**Chapter 3**). This common non-coding sequence variant is known to be associated with platelet volume and function. In **Chapter 4**, I demonstrated that this NDR was also present in primary megakaryocytes, and not in erythroblasts and monocytes, confirming the initial findings from cell lines.

Cell type-specific open chromatin regions are likely to play a role in regulation of cell type-specific gene expression. Therefore, sequence variants in NDRs restricted to a cell type may have the potential to impact traits linked to that specific cell type, for example megakaryocytes/platelets and platelet volume. However, functional studies at individual loci are crucial to prove that these non-coding sequence variants do indeed have a functional effect on the regulatory element, for example by disruption/introduction of transcription factor binding sites, and the phenotypic trait of interest.

The third objective of this thesis was to perform a proof-of-principle experiment to define the molecular mechanism of a GWA locus associated with platelet volume and function. In this chapter, I use *in silico* transcription factor binding site prediction, electrophoretic mobility shift assays (EMSAs), expression QTL (eQTL) analyses, transcription factor chromatin immunoprecipitation (ChIP) and mouse models, to elucidate the molecular basis of the association with the platelet phenotype at the 7q22.3 locus.

5.2. Identification of rs342293 as the only likely putative functional candidate at the 7q22.3 locus

The 65 kb recombination interval associated with platelet volume and function at chromosome 7q22.3 (Soranzo, Rendon, et al., 2009; Soranzo, Spector, et al., 2009; **Figure 5-1 A**) exhibited a total of six distinct NDRs in MK and EB cells. Of these six NDRs, two were common to both cell types and located at an evolutionary conserved element 10 kb upstream of the GWA index SNP rs342293 and at the promoter region of *FLJ36031*, three were specific to EB cells, and one was specific to MK cells (**Figure 5-1 B**).

The MK-specific NDR contained SNPs associated with mean platelet volume (MPV): rs342293 (MAF=0.48, Phase II+III HapMap, CEU population) and its best proxy rs342294 (r^2=1.0, Phase II+III HapMap, CEU). Data from the 1000 Genomes Project (Pilot 1, CEU; The 1000 Genomes Project Consortium, 2010) revealed 34 SNPs in LD ($r^2 \geq 0.8$) with rs342293, and confirmed that only rs342293 and rs342294 fall into the MK-specific NDR (**Appendix, Table 8-5**). This NDR was absent in both FAIRE-chip data sets (8 and 12 min formaldehyde cross-linking conditions) in EB cells (**Figure 5-1 C**). Irrespective of LD to rs342293, there were no sequence variants reported by the 1000 Genomes Project (Pilot 1, CEU) within sites of open chromatin in MK cells at the recombination interval.

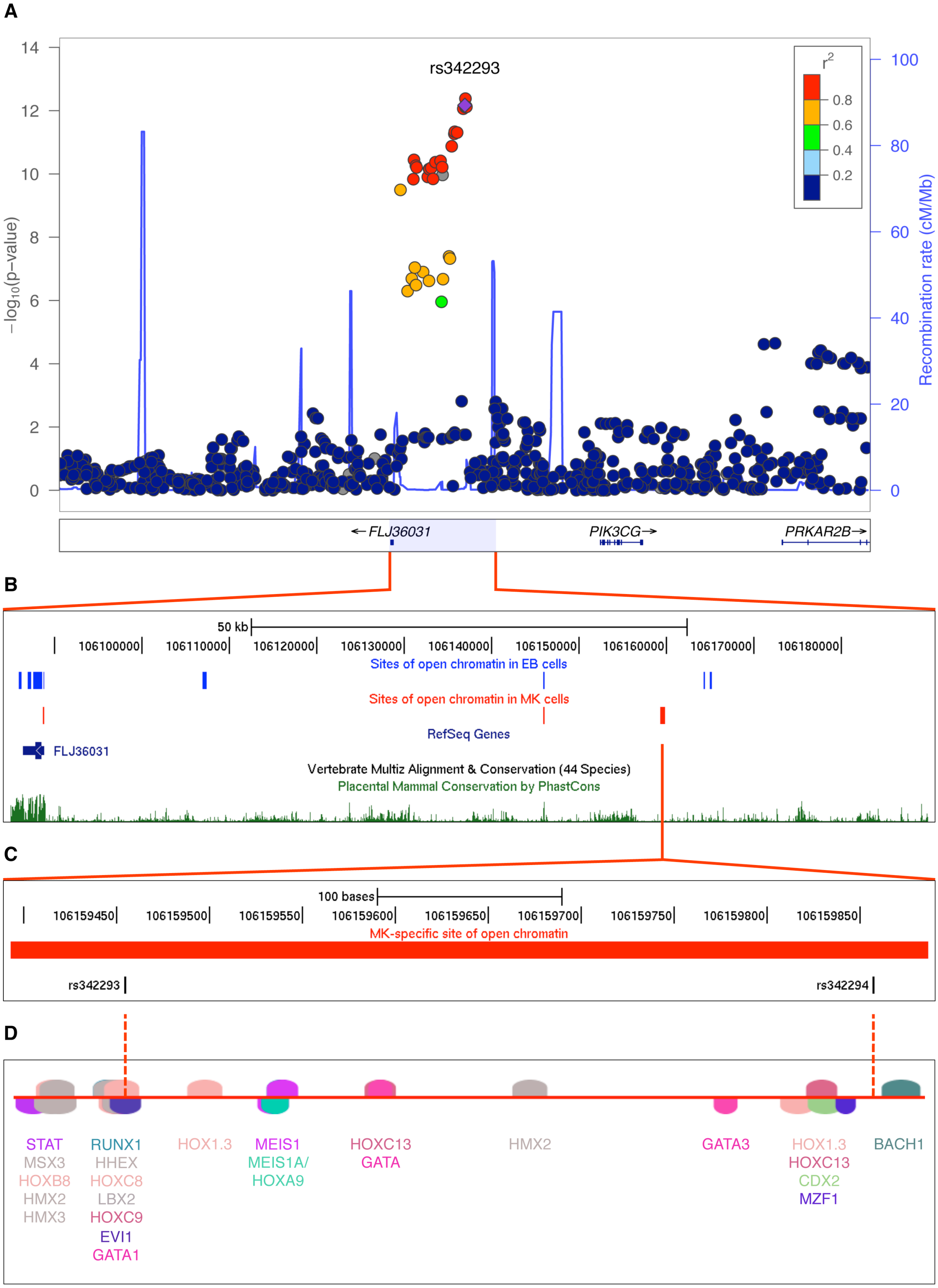


Figure 5-1. Functional follow-up of the 7q22.3 locus associated with platelet volume and function. (A) Regional plot of the 7q22.3 locus showing data from the discovery GWA meta-analysis as reported by Soranzo, Spector, et al., 2009. The SNP rs342293 is indicated in purple (*P*=6.75x10^{-13}). Values for

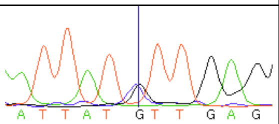
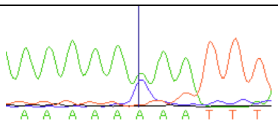
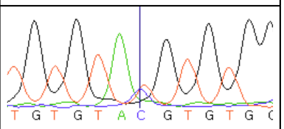
r^2 were based on the 1000 Genomes Project (Pilot 1, CEU). The data was plotted with LocusZoom v1.1 (http://csg.sph.umich.edu/locuszoom/). (**B**) Sites of open chromatin marked by FAIRE across the locus in the CHRF-288-11 (MK) and K562 (EB) cell lines (data uploaded as UCSC Genome Browser custom tracks). (**C**) Site of open chromatin found in MK but not EB cells harbouring the common variants rs342293 and rs342294. (**D**) *In silico* annotation of transcription factor binding sites at the open chromatin region described in (C). The predicted binding events are shown as transcription factor matrices (MatInspector v8.01). Note that some predicted binding sites are overlapping and not visible.

Next, I performed an *in silico* analysis of transcription factor binding sites at the 7q22.3 locus (**Appendix, Table 8-5**). Of the 34 SNPs in LD with rs342293, four (rs342240, rs342247, rs342292 and rs342293) disrupted an *in silico* predicted transcription factor binding site. However, only rs342293 was located within an experimentally verified site of open chromatin in MK cells. Among these four SNPs, rs342293 was the most strongly associated with mean platelet volume (Soranzo, Spector, et al., 2009).

The SNP rs342293 was located within overlapping predicted binding sites for the transcription factors BARX2, EVI1, GATA1, HHEX, HOXC8, HOXC9 and LBX2 (**Figure 5-1 D**). Based on the HaemAtlas data, of these seven transcription factors only EVI1, GATA1 and HHEX are expressed in megakaryocytes (**Figure 5-5 B**). The C>G substitution leads to disruption of the predicted binding sites of EVI1 and GATA1 (**Figure 5-2 A**). It is worth noting that *in silico* analysis predicted a RUNX1 transcription factor binding site only 5 bp apart from the EVI1-like binding site (discussed in **Section 5.3**).

To obtain the full spectrum of sequence variation at this region, we sequenced the NDR (chr7:106,159,393–106,159,887, build: hg18; 494 bp) in 643 healthy individuals of the Cambridge BioResource. No other common or low-frequency variants were detected, although a rare, possibly private, SNP was detected in one individual as a heterozygous position (chr7:106,159,601A>C) located in a polyA-region (**Table 5-1**).

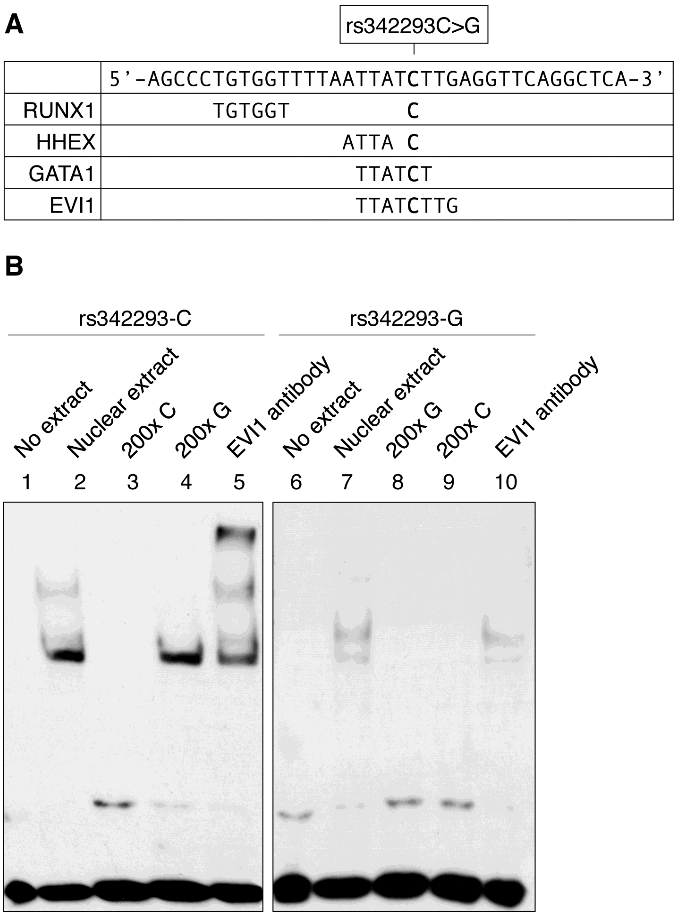
Table 5-1. Resequencing of the MK-specific open chromatin region at chromosome 7q22.3. The low call rate for rs342294 was due to the location of the SNP at the very end of the sequence-tagged site (STS). MAFs from Phase II+III HapMap (CEU) and the 1000 Genomes Project (Pilot 1, CEU) were retrieved from dbSNP v131. Genomic coordinates were based on the human reference genome, build hg18.

	rs342293	private	rs342294
Calls	600/643 (93.3%)	633/643 (98.4%)	330/643 (51.3%)
Position	chr7:106,159,455	chr7:106,159,601	chr7:106,159,858
Genotype	C/G	A/C	T/C
MAF (observed)	0.4167	0.0008	0.3909
MAF (HapMap)	0.482	n/a	0.482
MAF (1K Genomes)	0.458	n/a	0.458
Sequencing traces			

Based on the above evidence and the absence of any additional common or low-frequency sequence variant at the MK-specific NDR, rs342293 remained the only likely putative functional candidate underlying the MPV association signal at the 7q22.3 locus. However, it cannot be excluded that additional functional variants may exist outside NDRs and exert their function through, for example, disruption of a yet to be identified transcription factor binding site or modulation of DNA methylation.

5.3. The alleles of rs342293 differentially bind the transcription factor EVI1

I then performed EMSAs in nuclear protein extracts from the megakaryocytic cell line CHRF-288-11. I observed a band shift with oligonucleotide probes harbouring rs342293 for both the ancestral allele rs342293-C and the alternative allele rs342293-G (**Figure 5-2 B**). However, the bands were unequally shifted suggesting differential protein binding properties at this position depending on the allele of rs342293 tested. The specific unlabelled competitors supported specificity of the retarded bands. In addition, the results suggested superior protein binding to the probe containing the C-allele. With supershift experiments using an EVI1 antibody, I confirmed binding of EVI1 to probes containing the ancestral C-allele, but not to those harbouring the alternative G-allele (**Figure 5-2 B**).



	5'-AGCCCTGTGGTTTTAATTAT**C**TTGAGGTTCAGGCTCA-3'
RUNX1	TGTGGT C
HHEX	ATTA C
GATA1	TTATCT
EVI1	TTATCTTG

Figure 5-2. The effect of rs342293C>G on transcription factor binding. (A) Nucleotide sequence of the synthetic oligonucleotide probe harbouring rs342293 and *in silico* prediction of transcription factor binding sites. The major (C-) allele of rs342293 is shown in bold. The minor (G-) allele of rs342293 was predicted to disrupt the binding motifs of the transcription factors GATA1 (– strand) and EVI1 (–). RUNX1 (+) and HHEX (+) transcription factor binding sites were predicted to be located 9 bp and 1 bp apart from rs342293, respectively. (B) EMSAs in nuclear protein extracts from the megakaryocytic cell line CHRF-288-11 showed differential binding of the two alleles of rs342293. Competition reactions were performed with a 200-fold molar excess of unlabelled probes. Only the probe harbouring the reference allele was able to shift the protein-DNA complex when incubating with EVI1 antibodies.

Supershift experiments with a GATA1 antibody did not support *in vitro* binding of the GATA1 transcription factor to this site. I also confirmed RUNX1 transcription factor binding *in vitro* by demonstrating a supershift for both probes with a RUNX1 antibody (**Figure 5-3**).

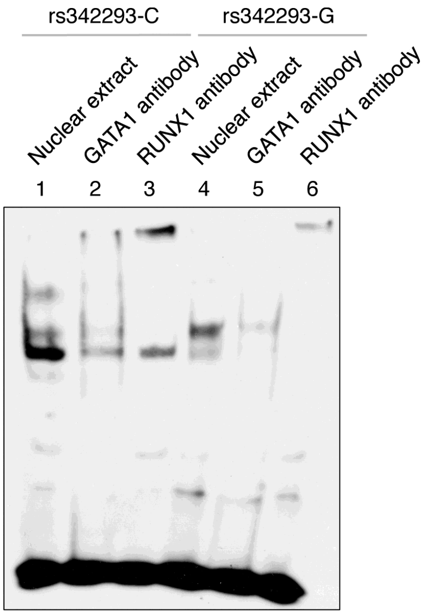


Figure 5-3. Gel shift assays in CHRF-288-11 nuclear protein extracts using GATA1 and RUNX1 antibodies. No supershift was observed for probes containing either rs342293-C or -G when incubating CHRF-288-11 nuclear protein extract with GATA1 antibodies. However, we showed evidence for RUNX1 transcription factor binding *in vitro*.

I further corroborated the *in silico* predictions and EMSA results by integrating data from genome-wide maps of transcription factor binding. Tijssen et al. performed chromatin immunoprecipitation combined with next-generation DNA sequencing (ChIP-seq) with GATA1 and RUNX1 antibodies in primary human megakaryocytes (Tijssen et al., 2011). No significant GATA1 but weak RUNX1 binding was observed at this locus (**Figure 5-4**).

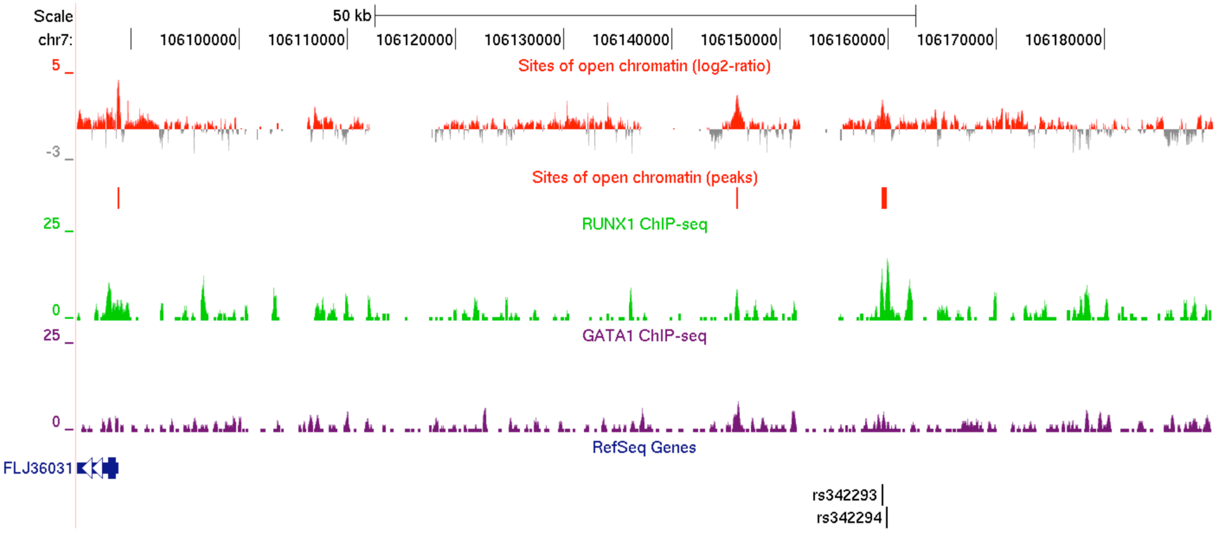


Figure 5-4. GATA1 and RUNX1 ChIP-seq profiles at the MK-specific open chromatin region at chromosome 7q22.3 in primary megakaryocytes. Data were transformed into density plots and displayed as UCSC Genome Browser custom tracks. Visual inspection of the 7q22.3 region showed no *in vivo* binding of GATA1, but weak RUNX1 binding. In total, the ChIP-seq data sets comprised 4,722 and 7,345 peaks for GATA1 and RUNX1, respectively (Tijssen et al., 2011).

5.4. The SNP rs342293 is associated with *PIK3CG* transcript levels in platelets and macrophages

A total of six protein-coding genes map within 1 Mb of the GWA index SNP rs342293, i.e. *COG5*, *FLJ36031* (also known as *CCDC71L*), *HBP1*, *NAMPT* (also known as *PBEF1*), *PIK3CG* and *PRKAR2B*. Whole-genome gene expression analysis on Illumina HumanWG-6 v1 Expression BeadChips revealed expression of all six genes in cord blood-derived megakaryocytes (**Figure 5-5 A**; Soranzo, Rendon, et al., 2009; Watkins et al., 2009). However, only *PIK3CG* and *PRKAR2B* are transcribed in platelets (Soranzo, Rendon, et al., 2009).

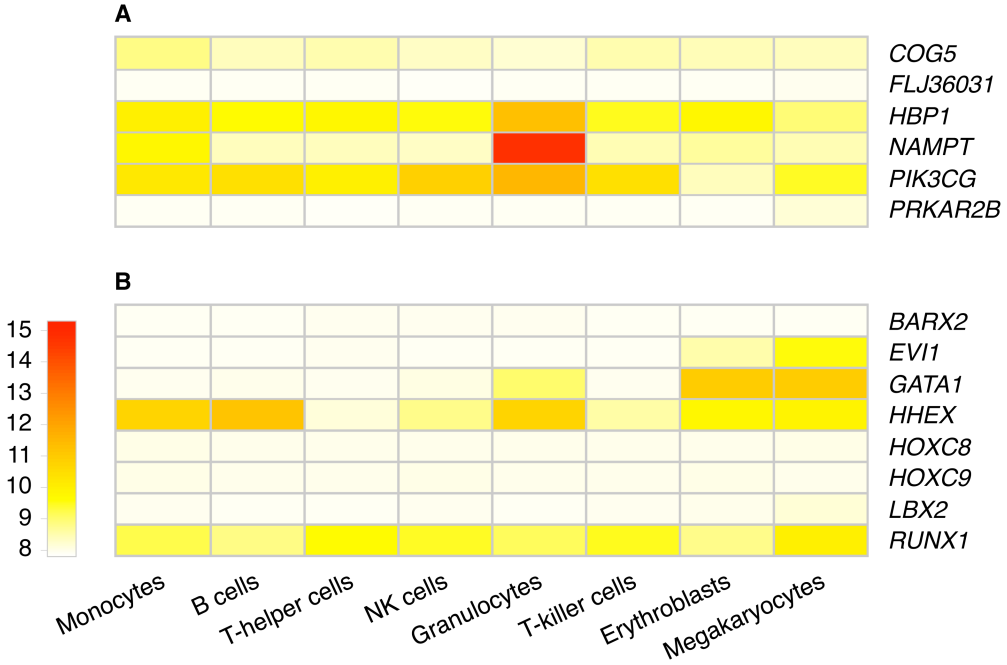


Figure 5-5. Gene expression profiles in differentiated human blood cells. Based on the HaemAtlas data (Watkins et al., 2009), the heat map shows normalised gene expression profiles of (A) genes within a 1 Mb interval of rs342293, and (B) genes encoding transcription factors predicted to bind DNA sequence motifs around rs342293.

Soranzo, Rendon, et al. investigated the association of rs342293 with transcript levels of both *PIK3CG* and *PRKAR2B* in platelets, and reported a weak eQTL association with *PIK3CG* transcript levels (permutation *P*=0.047) but not *PRKAR2B* (Soranzo, Rendon, et al., 2009). We replicated this eQTL association in an independent sample cohort of 24 healthy individuals (**Section 2.12**), obtaining the same genotypic effect for rs342293 (*P*=0.0542; **Figure 5-6 A**).

Next, we assessed the *PIK3CG*-eQTL in three different types of white blood cells, i.e. macrophages, monocytes and B cells (lymphoblastoid cell line, LCLs), as well as in different tissues, i.e. adipose (subcutaneous fat) and skin (**Appendix, Table 8-6**). We found an association between rs342293 and *PIK3CG* transcript abundance in macrophages (*P*=0.0018; **Figure 5-6 B**), but not in monocytes and LCLs. Further, we did not observe an association in adipose and skin tissues. Our data strengthened the previous evidence of rs342293 modulating *PIK3CG* transcript levels in platelets and showed that this eQTL was also present in macrophages. The association with *PIK3CG* transcript levels in macrophages but not monocytes suggested that the transcriptional effects may occur during late events of cellular differentiation (**Section 1.9**).

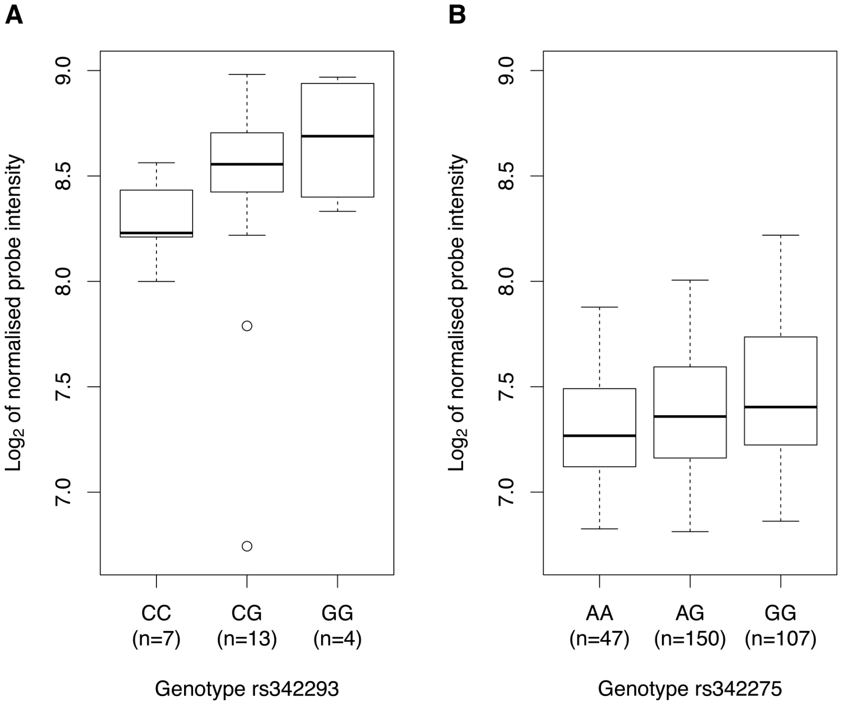


Figure 5-6. Association of rs342293 genotypes with *PIK3CG* transcript levels in (A) platelets and (B) macrophages. In platelets and macrophages, we observed a genotypic effect on *PIK3CG* transcript levels (*P*=0.0542 and *P*=0.0018, respectively). The associations are presented as box-and-whisker plots. In macrophages, the proxy SNP rs342275A>G (r^2=0.94, Phase II HapMap, CEU) was used for the eQTL analysis.

5.5. Gene expression profiles in whole blood of *Pik3cg*$^{-/-}$ mice

To further our understanding of the role of PIK3CG in platelets, we performed whole-genome gene expression profiling in whole blood of *Pik3cg* knockout mice (**Section 2.13**). Due to the need for relatively large quantities of RNA for whole-genome gene expression analysis on microarrays, we were only able to assay RNA extracted from whole blood as opposed to platelets. I identified 220 differentially expressed genes between knockout (n=3) and wild type mice (n=3) with a fold-change of at least ±1.5. Functional ontology classification of these genes (**Appendix, Table 8-7**) revealed enrichment for 'regulation of biological characteristic', e.g. cell size and volume (GO term: 0065008; *P*=2.97x10^{-14}), and 'blood coagulation' (GO term: 0007596; *P*=7.87x10^{-12}). Notably, the 220 differentially expressed genes included *Gp1bb* (fold-change of -2.203 in *Pik3cg*$^{-/-}$ compared to wild type mice), *Gp5* (-3.211), *Gp9* (-0.996), *Gp6* (-0.711) and *Vwf* (-1.381). These five genes are transcribed in the megakaryocytic lineage in humans, based on the HaemAtlas data. The platelet glycoproteins GP1BB, GP5 and GP9, together with GP1BA, constitute the platelet membrane receptor for the plasma protein Von Willebrand Factor (VWF), which is encoded by *VWF* (Roth, 1991).

5.6. Canonical pathway enrichment analysis and protein-protein interaction network

To explore the signalling pathways of PIK3CG in humans, I analysed 191 human orthologs of the 220 differentially expressed genes between *Pik3cg*$^{-/-}$ and wild type mice. Canonical pathway enrichment analysis based on the curated gene sets of the Molecular Signatures Database (MSigDB) v3.6 (Subramanian et al., 2005) revealed that the top six enriched gene sets are related to platelets: 'platelet degranulation' (*P*=3.44x10^{-8}), 'platelet activation' (*P*=5.05x10^{-8}), 'formation of platelet plug' (*P*=2.85x10^{-7}), 'haemostasis' (*P*=7.48x10^{-6}), 'formation of fibrin clot and clotting cascade' (*P*=1.39x10^{-5}) and 'platelet adhesion to exposed collagen' (*P*=1.86x10^{-5}).

We constructed a protein-protein interaction network centred on the proteins encoded by the 191 transcripts described above. First-order interactors of these 'core' proteins were obtained from Reactome, an open-source, manually curated database of human biological pathways (Matthews et al., 2009; Croft et al., 2011). We filtered interactors on their expression levels in megakaryocytes (**Section 2.14**). The resulting network incorporated 45 core proteins centred on PIK3CG consisting of 642 nodes and 1067 edges (**Figure 5-7**).

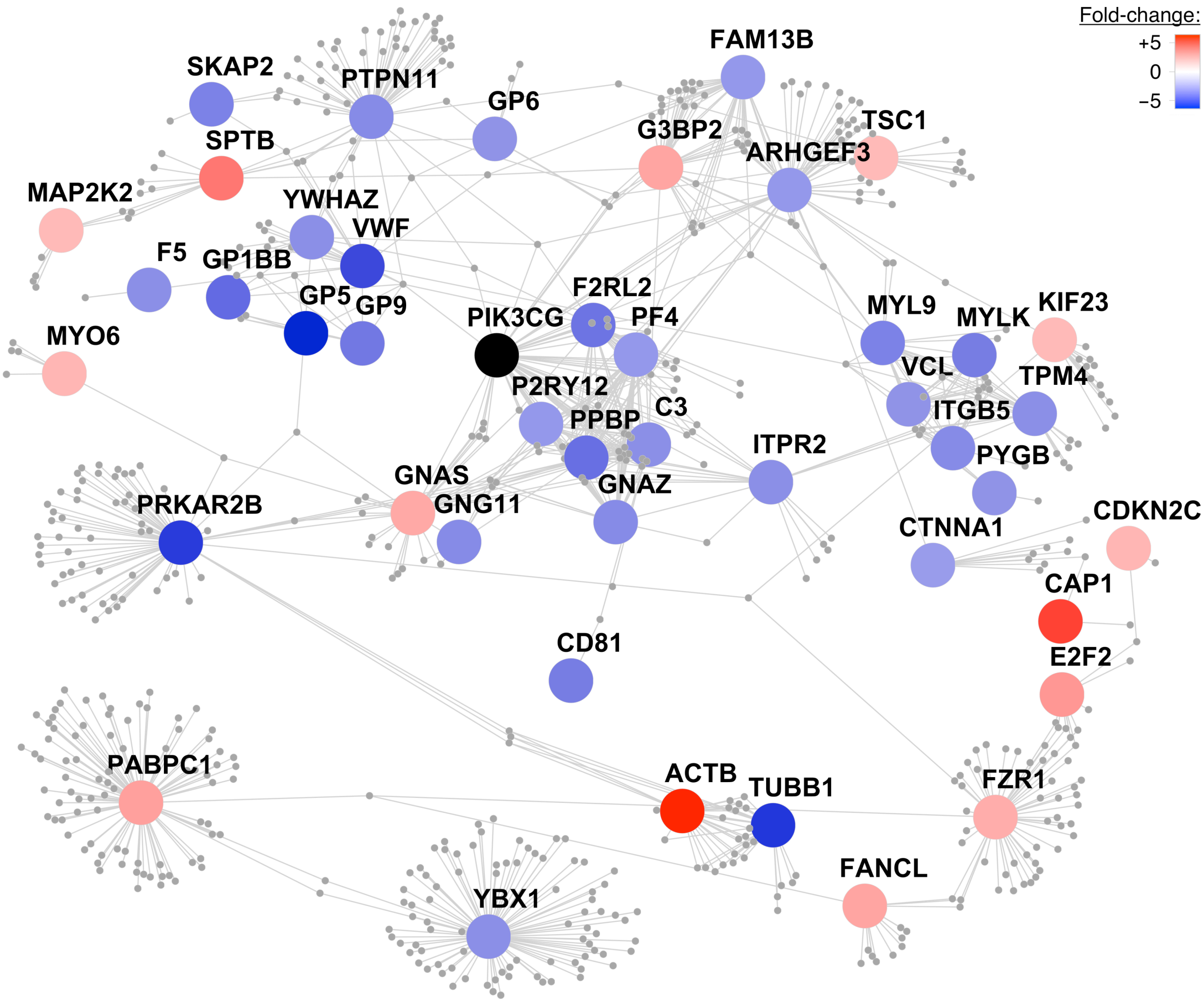


Figure 5-7. Protein-protein interaction network centred on PIK3CG. A protein-protein interaction network was constructed centring on the human orthologous proteins encoded by the 220 differentially expressed transcripts between *Pik3cg* knockout and wild type mice with a fold-change of at least ±1.5. The colour of these 'core' proteins represents the fold-change of gene transcript levels between *Pik3cg*$^{-/-}$ and wild type mice on a continuous scale from over- (red) to underexpression (blue) of transcripts in knockout mice. Core proteins, which did not exhibit first-order interactors in the Reactome database and were disconnected from the largest connected component, are not shown. First-order interactors of PIK3CG that form network nodes and edges are shown in grey and were obtained from Reactome. Interactors that are not expressed in megakaryocytes based on the HaemAtlas data were omitted. The resulting network consisted of 45 core proteins with 642 nodes and 1067 edges.

5.7. Discussion

We elucidated the molecular basis of the association between rs342293 and platelet volume and function at chromosome 7q22.3. We identified an NDR in MK but not EB cells containing the index

SNP rs342293 for this association. Resequencing of this NDR in 643 individuals provided strong evidence that rs342293 is the only putative causative variant in this region. However, resources such as the completed 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) will make the requirement for resequencing any identified NDR redundant with respect to identifying common variants (**Chapter 4**). I then demonstrated that the alleles of rs342293 differentially bind the transcription factor EVI1.

At the time, rs342293 was the first functional variant to be elucidated among the known platelet QTLs (Paul et al., 2011). Our data suggested a molecular mechanism by which a non-coding GWA index SNP modulates platelet phenotype. We note that our results implicate the original GWA tag SNP as the likely functional variant. Indeed, two studies identified the GWA index SNP for colorectal cancer at chromosome 8q24 to be a regulatory variant, and provided a mechanistic explanation for its association with disease risk (Pomerantz et al., 2009; Tuupanen et al., 2009).

Expression QTL data in platelets and macrophages provided statistical support that rs342293, or a variant in LD, affect *PIK3CG* gene expression levels. *PIK3CG* is transcribed in megakaryocytes but only weakly expressed in erythroblasts (Watkins et al., 2009; **Figure 5-5 A**), which is in agreement with the MK-specific properties of the identified NDR. However, additional work is required to scrutinise all possible targets of this regulatory module, in particular non-protein coding transcripts. *PIK3CG* is located 134 kb downstream of rs342293 and encodes the phosphoinositide-3-kinase γ-catalytic subunit. The lipid kinase PIK3CG (PI3Kγ) is a member of the class I PI3Ks and catalyses the conversion of phosphatidylinositol-4,5-bisphosphate (PtdIns(4,5)P2; PIP2) to phosphatidylinositol-3,4,5-trisphosphate (PtdIns(3,4,5)P3; PIP3) downstream of cell surface receptor activation (Wymann et al., 2003; Rückle et al., 2006; Hawkins & Stephens, 2007). In megakaryocytes and platelets, PIP3 is crucial in the collagen-induced regulation of phospholipase C and initiation of megakaryopoiesis and pro-platelet formation (Pasquet et al., 2000). Functional studies with *Pik3cg* knockout mice previously indicated a role in wound healing (Zhao et al., 2006), ADP-induced platelet aggregation and thrombosis (Hirsch et al., 2001; Schoenwaelder et al., 2007). It is also worth noting that PIK3CG has a prominent role in macrophage activation (Hawkins & Stephens, 2007).

The whole-genome gene expression profiling of *Pik3cg*$^{-/-}$ mice performed in this study showed differential expression of genes involved in important platelet-related pathways compared to wild type mice, most notably *Vwf* and its platelet membrane receptor components. A recent meta-analysis of GWA studies in 68,000 Northern Europeans showed that common sequence variants at the *VWF* and *GP1BA* loci exert an effect on platelet volume and count, respectively (Gieger et al., 2011). This recent finding, together with the established knowledge that Mendelian mutations in the genes encoding the platelet VWF (Von Willebrand disease, type 2; OMIM: 613554) and its receptor (Bernard-Soulier syndrome; OMIM: 231200) cause giant platelets, give biological credence to the effects we observed of *Pik3cg* knockout on the transcription of these platelet genes. The constructed protein-protein interaction network, which was centred on PIK3CG, highlighted additional proteins implicated in severe platelet disorders, including TUBB1 (macrothrombocytopenia, autosomal dominant, TUBB1-related; OMIM: 613112), F5 (factor V deficiency; OMIM: 612309) and P2RY12 (bleeding disorder due to P2RY12 defect; OMIM: 609821). Therefore, it is plausible to assume that differences in the *PIK3CG* transcript levels in humans, based on the different alleles of rs342293, may lead to changes in the abundance of platelet membrane proteins that are key regulators of platelet formation.

Soranzo, Rendon, et al. previously showed an association of rs342293-G with decreased platelet reactivity in humans (Soranzo, Rendon, et al., 2009), assessed as the proportion of binding to annexin V and fibrinogen, as well as P-selectin expression, after activation of platelets with collagen-related peptide (CRP-XL). We also observed that rs342293 is likely to modify events downstream of signalling via the collagen signalling receptor glycoprotein VI, which is encoded by *GP6* (Jones et al., 2009). A recent GWA study showed the same locus at chromosome 7q22.3 to also be associated with epinephrine-induced platelet aggregation (Johnson et al., 2010). The index SNP in that study, rs342286, and rs342293 are in high LD (r^2=0.87, Phase II HapMap, CEU) making rs342293 the putative causative variant underlying both functional associations. PIP3 is required for the platelet signalling cascades triggered via an immunoreceptor tyrosine-based activation motif on the Fc receptor γ-chain (ITAM-FcR-γ) for collagen and G-protein-coupled receptor for epinephrine. Therefore, our observations are compatible with the notion that these platelet functional events are modified by differences in the *PIK3CG* transcript level and that silencing of *Pik3cg* in mice reduces the expression of the *Gp6* gene.

EVI1 (ecotropic viral integration site-1) encodes a protein with two zinc-finger domains (ZF1 and ZF2), which feature distinct DNA-binding specificities (Bartholomew et al., 1997). EVI1 mainly promotes haematopoietic differentiation into the megakaryocytic lineage (Shimizu et al., 2002). EVI1 was frequently reported to be a repressor of transcription that has the potential to recruit diverse regulatory proteins. For example, EVI1 antagonises the growth-inhibitory effect of transforming growth factor-β (TGF-β), a potent regulator of megakaryopoiesis (Sakamaki et al., 1999), by interacting with SMAD3 via ZF1, and inhibiting SMAD3 from binding to DNA (Kurokawa, Mitani, Irie, et al., 1998; Kurokawa, Mitani, Imai, et al., 1998). EVI1 contains domains that interact with RUNX1 (Runt-related transcription factor 1), which is the α-subunit of the transcription factor CBF (core binding factor). The interaction of EVI1 with the DNA-binding domain Runt of *RUNX1* leads to the destabilisation of the DNA-RUNX1 complex and subsequent loss of RUNX1 function (Senyuk et al., 2007).

Based on the above data, I propose a model (**Figure 5-8**) in which the DNA sequence containing the C- but not the G-allele of rs342293 binds the transcription factor EVI1. Together with the transcription factor RUNX1, EVI1 may act as transcriptional repressor of *PIK3CG* in megakaryocytes. Individuals with homozygous rs342293-C have lower *PIK3CG* gene expression levels in platelets compared to subjects with homozygous rs342293-G. In *Pik3cg* knockout mouse models, transcripts that encode key proteins for platelet membrane biogenesis were found to be downregulated, which may ultimately affect the platelet phenotype.

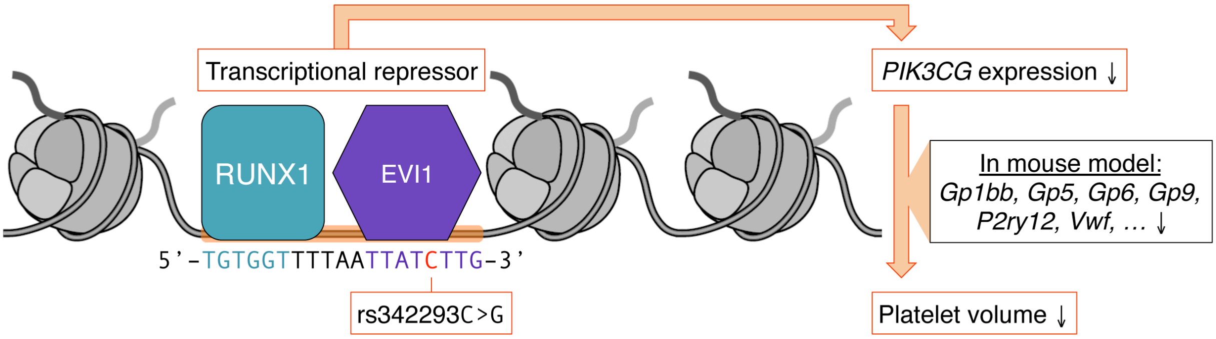

Transcriptional repressor
RUNX1
EVI1
5'-TGTGGTTTTAATTATCTTG-3'
rs342293C>G
PIK3CG expression ↓
In mouse model:
Gp1bb, Gp5, Gp6, Gp9, P2ry12, Vwf, … ↓
Platelet volume ↓

Figure 5-8. Model of the mechanism by which rs342293 may affect platelet volume. Based on our data, we propose that rs342293C>G affects EVI1 transcription factor binding at the megakaryocyte-specific site of open chromatin at chromosome 7q22.3. The underlying DNA sequence around rs342293 (indicated in red) is shown and transcription factor binding motifs for RUNX1 (turquoise) and EVI1 (purple) are highlighted. The transcriptional complex of EVI1 is a well-described repressor of gene transcription.

CHAPTER 6

Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in *RBM8A* causes TAR syndrome.

This chapter is based on the following publication:

Albers, C.A., Paul, D.S., Schulze, H., et al. (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit *RBM8A* causes TAR syndrome. Nat. Genet. *44*, 435-439.

Collaboration note:

Section 6.2: Cornelis A. Albers[1–3] performed next-generation sequence, Sanger sequence, genetic and statistical analyses. Graham Kiddle[1,2] supervised exome sequencing. Jonathan C. Stephens[1,2] performed Sanger sequencing and analysed the data. Harald Schulze[4,5], Kathleen Freson[6], Janine Fiedler[5,7], Kenneth Smith[8,9], Chantal Thys[6] and Ruth Newbury-Ecob[8,9] ascertained deletion status for TAR cases. Harald Schulze, Martijn H. Breuning[10], Najet Debili[11], Rémi Favier[11], Ingrid Krapels[12], Paquita Nurden[13], Claudia A.L. Ruivenkamp[10], Gabriele Strauss[14], Chris van Geet[6,15], Ruth Newbury-Ecob and Cedric Ghevaert[1,2] clinically characterised TAR cases. I did not contribute to the analyses and experiments described in this section.

Section 6.3: Kathleen Freson and Chantal Thys performed luciferase assays. Harald Schulze, Kathleen Freson, Chantal Thys, Cedric Ghevaert and Catherine M. Hobbs[1,2] performed protein blot experiments. Myrto Kostadima[16] and Paul Bertone[16] analysed the megakaryocyte RNA sequencing data. I performed FAIRE-seq experiments and analysis, EMSA studies and *in silico* transcription factor binding analysis.

[1]Department of Haematology, University of Cambridge, Cambridge, UK; [2]National Health Service (NHS) Blood and Transplant, Cambridge, UK; [3]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; [4]Institute for Transfusion Medicine, Charité Universitätsmedizin, Berlin, Germany; [5]Laboratory for Pediatric Molecular Biology, Charité Universitätsmedizin, Berlin, Germany; [6]Center for Molecular and Vascular Biology, University of Leuven, Leuven, Belgium; [7]Department of Biology, Chemistry, and Pharmacy, Freie University Berlin, Berlin, Germany; [8]Division of Child Health, University of Bristol, Bristol, UK; [9]Department of Clinical Genetics, St Michael's Hospital, Bristol, UK; [10]Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands; [11]Institut National de la Santé et de la Recherche Médicale (INSERM) U790, Villejuif, France; [12]Department of Clinical Genetics, Maastricht University Medical Center, Maastricht, The Netherlands; [13]Laboratoire d'Hématologie, Centre de Référence des Pathologies Plaquettaires, Hopital Xavier Arnozan, Pessac, France; [14]Department of Pediatric Oncology and Hematology, Charité Universitätsmedizin, Berlin, Germany; [15]Department of Pediatrics, Universitair Ziekenhuis Leuven, Leuven, Belgium; [16]European Molecular Biology Laboratory (EMBL)–European Bioinformatics Institute (EBI), Hinxton, Cambridge, UK.

6.1. Introduction

In **Chapters 3** and **4**, I applied FAIRE-generated maps of open chromatin to functionally assess sequence variants associated with complex traits. As a proof-of-concept, I subsequently investigated the molecular basis of the association between the non-coding GWA index SNP rs342293 and platelet volume and function at chromosome 7q22.3 (**Chapter 5**).

The final objective of this thesis was to explore the use of open chromatin maps to annotate low-frequency variants linked to a rare disease. For this purpose, I considered variants identified through exome sequencing of patients with thrombocytopenia with absent radii (TAR), a rare inherited blood and skeletal disorder. In this chapter, we functionally assess the candidate causal variants, found to be located in an NDR, and apply the experimental approach described in **Chapter 5** to establish the underlying biological mechanism.

6.1.1. Exome sequencing as a tool for gene discovery in rare diseases

In most cases, rare Mendelian diseases are caused by rare mutations, as selection acts strongly against these alleles. Strategies for identifying causal alleles depend on various factors including the structure of the pedigree or population, the mode of inheritance of a trait and the extent of locus heterogeneity.

Highly parallel sequencing has been successfully applied to identify causal mutations for monogenic disorders. This approach has been used to target genes within linkage intervals (Volpi et al., 2009; Nikopoulos et al., 2010), all protein-coding regions in the genome, referred to as 'exome' (Ng, Buckingham, et al., 2010; Ng, Bigham, et al., 2010), or whole genomes (Lupski et al., 2010; Roach et al., 2010). In recent years, many large-scale medical sequencing projects have focused on exome sequencing. One reason for this is cost, as whole-genome sequencing is still relatively expensive for large sample sizes. Another reason is biology, as most known examples of disease-causing variants alter the protein sequence, and functional assessment of non-coding genetic variation has been challenging.

There are four main strategies for identifying rare disease-causing variants through exome sequencing (Cirulli & Goldstein, 2010; Bamshad et al., 2011): (i) the sequencing of multiple affected but unrelated individuals; (ii) the sequencing of multiple affected individuals from the same pedigree; (iii) the sequencing of trios (parents and child) for identifying *de novo* mutations; and (iv) the sequencing of individuals at the extreme ends of a quantitative trait distribution.

For sequencing only a subset of the genome using next-generation DNA sequencing technology, 'genome-partitioning' methods are used. This requires the preparation of complex mixtures of sequencing templates that are highly enriched for the targeted genomic regions. In most exome sequencing protocols, the protein-coding fraction of the genome is selected by either solid-phase (Albert et al., 2007; Okou et al., 2007; Porreca et al., 2007) or liquid-phase hybridisation (Gnirke et al., 2009) to a complementary set of tiling oligonucleotide probes. After target enrichment, the regions are sequenced to great depth, i.e. a mean coverage of greater than 80-fold. In order to isolate pathogenic mutations from background polymorphisms, identified variants are filtered primarily based on function and frequency (reviewed in Stitziel et al., 2011 and Bamshad et al., 2011).

6.1.2. Genetics of thrombocytopenia with absent radii (TAR) syndrome

The thrombocytopenia with absent radii (TAR) syndrome is characterised by bilateral radial aplasia (the absence of the radius bones in the forearms) and severe thrombocytopenia (the reduction in the number of platelets). The incidence of TAR is estimated at 1:200,000–1:100,000 (http://www.ncbi.nlm.nih.gov/books/NBK23758/). However, many pregnancies are aborted if TAR is detected, therefore the real incidence may be higher. An excess of affected females has also been suggested (Greenhalgh et al., 2002). In contrast to other syndromes that combine absence of the radius with blood abnormalities, such as Fanconi anaemia, the thumb is preserved in TAR (Shaw & Oliver, 1959; Hall et al., 1969; Geddis, 2006). As illustrated in **Figure 6-1**, the severity of skeletal abnormalities varies from absence of radii to virtual absence of upper limbs with or without lower-limb defects, such as malformations of the hip and knee (Greenhalgh et al., 2002). Individuals with TAR have low numbers of MKs and frequently present with bleeding episodes in the first year of life, which diminish in frequency and severity with age. In TAR, platelet levels are generally below $50x10^9$ platelets per litre, with the normal range being $150–350x10^9$ platelets per litre.

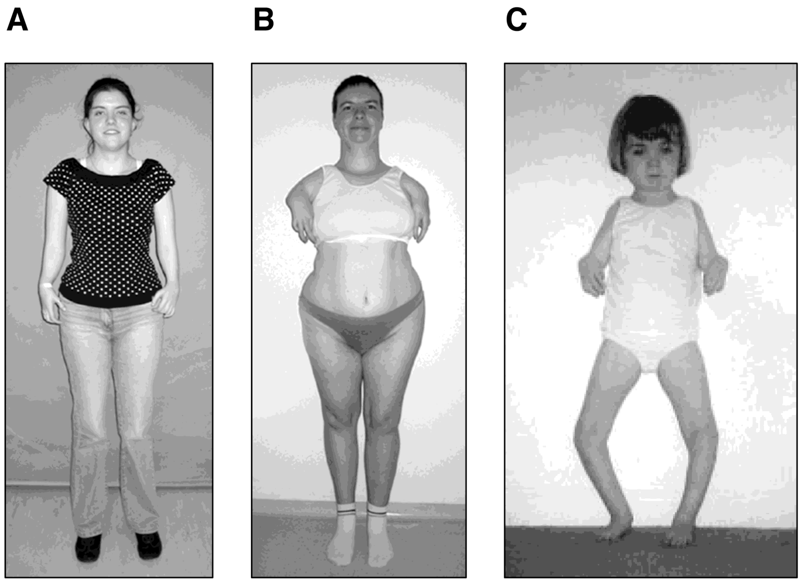


Figure 6-1. Skeletal abnormalities in TAR cases. (A) Patient shows mild upper-limb involvement with slightly reduced lengths of the arms. (B) Severe TAR phenotype with phocomelia (congenital absence of the proximal portion of a limb or limbs). (C) Lower limb involvement in a child with TAR, i.e. severe bowing of the legs. The pictures were adapted from Greenhalgh et al., 2002 and Klopocki et al., 2007.

An inherited or *de novo* deletion at chromosome 1q21.1 is found in the majority of affected individuals (Klopocki et al., 2007), but the apparent autosomal recessive nature of the syndrome requires the existence of an additional causative allele. This other allele has remained elusive, even with sequencing of the protein-coding exons of ten genes (including *RBM8A*) in the minimally deleted region (chr1:145,399,075–145,594,214, build: hg19; 195 kb), as reported by Klopocki et al., 2007.

6.2. Most TAR cases have a low-frequency regulatory variant and a rare null allele at the *RBM8A* locus

To identify the additional causative allele, we selected five individuals with TAR ('cases') of European ancestry, who had the 1q21.1 deletion (**Figure 6-2 A**), and sequenced their exomes using the SureSelect Human All Exon Kit [Agilent Technologies] (**Section 2.15**). All study subjects fulfilled the diagnostic criteria for TAR syndrome as described in **Section 6.1.2**. The clinical and genotype information of the TAR cases and their healthy parents are provided in **Appendix, Table 8-8**. Per individual, 13.1–13.5 Gb of sequence was generated, resulting in a mean coverage of 123–127-fold, with 89.9–90.5% of the targets covered by at least 10-fold.

We were unable to find recessive novel mutations in the protein-coding regions in the five TAR patients (**Section 2.15**). However, four of the cases carried the minor allele of a low-frequency SNP (chr1:145,507,646; rs139428292G>A) in the 5'-UTR of the *RBM8A* gene, while the remaining case carried a previously unknown SNP (chr1:145,507,765G>C) in the first intron of the same gene (**Figure 6-2 B**). Genotyping by Sanger sequencing of additional 48 cases of European ancestry with the 1q21.1 deletion identified rs139428292 ('5'-UTR SNP' hereafter) and chr1:145,507,765 ('intronic SNP' hereafter) in 35 and 11 samples, respectively (**Figure 6-2 C**; **Appendix, Table 8-8**).

In total, 34 trios of mother, father and child were investigated (**Appendix, Table 8-8**). In all 25 trios of European ancestry, where the deletion in the child was not inherited *de novo*, we confirmed that the deletion and the newly identified SNPs were inherited from different parents. Therefore, the observed mutations were compatible with a compound autosomal recessive mode of inheritance. Among the 34 trios, there was one previously reported example of vertical transmission of TAR (Klopocki et al., 2007). Both the affected mother and her (aborted) foetus, which showed skeletal features of TAR on ultrasound, carried the typical 1q21.1 deletion. It is important to note that, in contrast to all other cases studied, these patients were of non-European ancestry. Sequencing of the entire *RBM8A* gene, including exons, introns, 5'-UTR, promoter, as well as a putative regulatory element 4 kb upstream of the promoter, showed an absence of the minor alleles of both the 5'-UTR and intronic SNPs in all three samples. We did not identify an alternative sequence variant as a potential additional causative allele. Thus, we have failed to identify the second causative allele in this sporadic case of vertical transmission of TAR. We reasoned that another longer-distance *cis*-acting or possibly *trans*-acting modifier of the *RBM8A* locus may explain the disorder in this pedigree.

From the genotyping of 7,504 healthy individuals of the Cambridge BioResource, we estimated MAFs of 3.05% and 0.42% for the 5'-UTR SNP and the intronic SNP, respectively (**Table 6-1**). Analysis of copy number variants at the chromosome 1q21.1 locus in 5,919 healthy individuals from the Wellcome Trust Case Control Consortium did not reveal deletions of the *RBM8A* gene in these individuals, indicating a low frequency of the 1q21 deletions found in TAR cases and their healthy relatives. This is in agreement with the low incidence of TAR syndrome in the population. We observed five duplications, which suggested that overexpression of *RBM8A* is not deleterious (The Wellcome Trust Case Control Consortium, 2010; Huang et al., 2010). Thus, the concurrent presence of one of the two non-coding SNPs at one allele and the 1q21.1 deletion at the other is strongly associated with TAR syndrome (estimated P<5x10^{-228}; Albers et al., 2012).

Table 6-1. Genotyping of the 5'-UTR and intronic SNPs at the *RBM8A* locus in 7,504 healthy individuals of the Cambridge BioResource and association with platelet count. Platelet count data was available for 6,805 and 6,938 of the 7,504 individuals genotyped for the 5'-UTR SNP and the intronic SNP, respectively. The number of individuals for each genotype of the 5'-UTR and intronic SNPs with measured platelet count is indicated in parentheses. The log-transformed platelet count on genotype was regressed using an additive genetic model, adjusted for gender and age in years at date of venesection. Abbreviations: MAF: minor allele frequency; HWE: Hardy-Weinberg equilibrium.

	5'-UTR SNP (G/A)	Intronic SNP (G/C)
Genotypes passed QC (call rate)	7,317 (97.5%)	7,458 (99.4%)
Homozygous major	6,879 (6,402)	7,396 (6,879)
Heterozygous	431 (396)	62 (59)
Homozygous minor	7 (7)	0 (0)
Estimated MAF	3.05%	0.42%
Deviation from HWE (exact test)	*P*=0.84	*P*=1.00
Association with platelet count	*P*=0.87	*P*=0.99

Next, we sequenced all exons of *RBM8A* in two additional TAR cases, who did not carry the 1q21.1 deletion but were found to carry the 5'-UTR SNP. In the first case, we identified a 4 bp frameshift insertion at the start of the fourth exon, and established that the non-coding SNP and insertion were at different chromosomes. By genotyping the parents of this case, we identified the 4 bp insertion in the mother and the 5'-UTR SNP in the father, both as heterozygous positions (**Appendix, Table 8-8**). In the second case, we identified a nonsense mutation in the last exon of *RBM8A* (**Figure 6-2 B,C**). Both mutations were absent from 458 exome samples of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) and 416 samples from the Cohorte Lausannoise ('CoLaus'; Firmann et al., 2008). We concluded that, in the vast majority of cases, compound inheritance of a rare null allele (containing a deletion, frameshift mutation or encoded premature stop codon) and one of two low-frequency non-coding SNPs in *RBM8A* causes TAR syndrome. On the basis of the genetic results, we postulated a hypomorphic mechanism for TAR, in which one copy of the *RBM8A* gene is not functional (due to a null allele), and expression of the other copy is reduced (as a result of non-coding SNPs in the 5'-UTR or first intron).

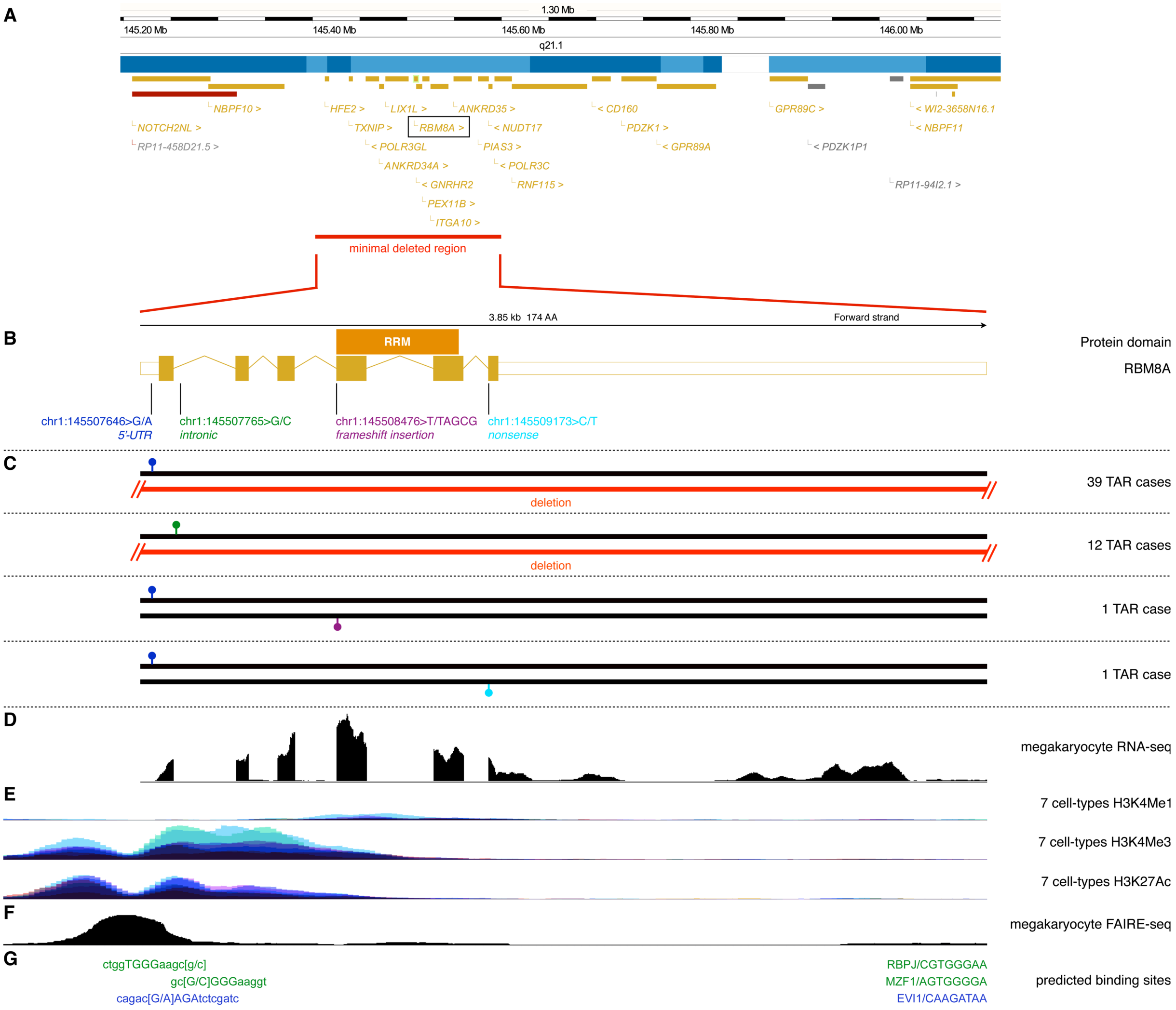


Figure 6-2. Low-frequency non-coding SNPs and a rare null mutation at the *RBM8A* locus. (A) Fifty-three of 55 TAR cases were heterozygous carriers of a rare 1q21.1 deletion of varying size. The red bar indicates the region that was absent in all 53 cases having a deletion. Legend: Yellow bars: genes; grey bars: pseudogenes; blue bars: contigs. (B) The *RBM8A* transcript is shown. The sequence encoding the RNA-binding domain (RRM) is indicated by the orange bar above the transcript. (C) We identified two low-frequency regulatory SNPs in 53 of a total of 55 TAR cases studied. The first, at chr1:145,507,646 (rs139428292G>A), is located at the 5'-UTR of *RBM8A* and has a population MAF of 3.05% (dark blue). The second, at chr1:145,507,765G>C, is located at the first intron of *RBM8A* and has a population MAF of 0.41% (green). Thirty-nine TAR cases carried the minor allele of the 5'-UTR SNP at one chromosome and the 1q21.1 deletion at the other. Twelve TAR cases carried the minor allele of the intronic SNP at one chromosome and the 1q21.1 deletion at the other. Compound inheritance of the 1q21.1 deletion and one of the two regulatory SNPs was strongly associated with TAR. Two additional TAR cases were found to have the minor allele of the 5'-UTR SNP in combination with either a frameshift insertion (purple) or nonsense mutation (light blue) instead of the 1q21.1 deletion, implicating *RBM8A* as the causative gene for TAR syndrome. (D) Sequencing of RNA from cord blood-derived MKs provided evidence that *RBM8A* is transcribed in MKs. Shown is the sequencing read depth across the *RBM8A* locus. (E) Histone modifications in seven cell lines (GM12878, H1-hESC, HSMM,

HUVEC, K562, NHEK and NHLF) reported by the ENCODE Project (The ENCODE Project Consortium, 2007) indicated the presence of active regulatory elements at the promoter region, including the 5'-UTR and first intron of *RBM8A*. Shown are the read depths resulting from ChIP-seq experiments of the three histone marks in the seven cell lines. The coverage profiles of the different cell types are represented by different shades of blue and are superimposed. (F) Coverage profile of FAIRE-seq experiments showed that the 5'-UTR and intronic SNPs are accessible to regulatory factors in MKs. (G) *In silico* transcription factor binding site analysis predicted that the minor allele of the 5'-UTR SNP creates a binding site for the EVI1 transcription factor. The minor allele of the intronic SNP was predicted to disrupt binding of MZF1 and RBPJ. Capital letters indicate the consensus transcription factor binding sites, and the alleles of the SNPs are shown in parentheses.

6.3. The effect of the regulatory SNPs on transcription factor binding, *RBM8A* promoter activity and protein expression in platelets

Analysis of histone modifications in seven human cell lines from the ENCODE Project (The ENCODE Project Consortium, 2007) indicated that the 5'-UTR and the intronic SNP are located in potential active regulatory elements (**Figure 6-2 D,E**). Annotation of open chromatin structure using the FAIRE-seq technique provided additional evidence in MKs (**Figure 6-2 F**). Computational predictions suggested that the 5'-UTR SNP introduces a binding site for the transcriptional repressor EVI1 and that the intronic SNP disrupts a binding site for the transcription factors MZF1 and RBPJ (**Figure 6-2 G**).

I confirmed the prediction of EVI1 binding by EMSAs in the megakaryocytic cell line CHRF-288-11, in which the EVI1 protein bound to the oligonucleotide probe carrying the minor allele of the 5'-UTR SNP but only weakly associated with the major allele (**Figure 6-3 A**). EMSA studies for the intronic SNP showed a decrease in the binding of nuclear proteins to the minor allele, although I could not confirm the presence of either MZF1 or RBPJ in supershift experiments (**Figure 6-3 B**).

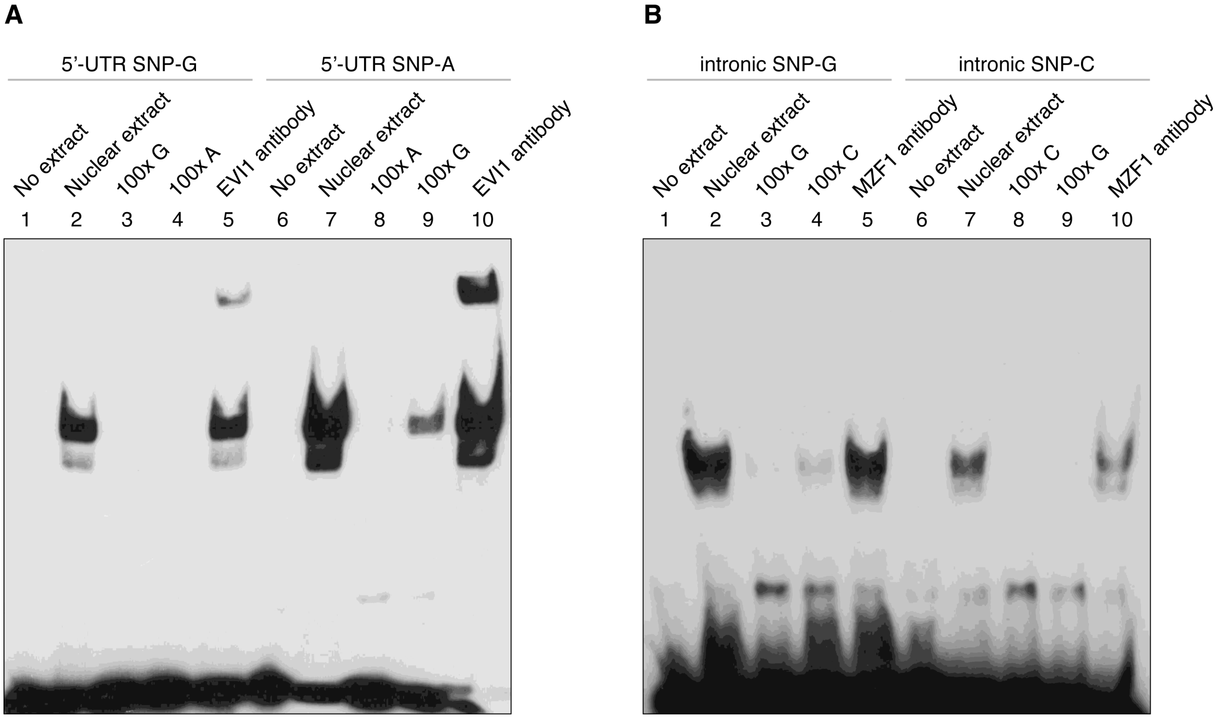


Figure 6-3. The effect of the regulatory 5'-UTR and intronic SNPs on transcription factor binding. (A) EMSAs in nuclear protein extracts from CHRF-288-11 cells showed higher protein affinity to probes with the A-allele (lane 7) than to probes with the G-allele (lane 2) of the 5'-UTR SNP. Binding of the A-allele probe was competed by a specific (lane 8) but not by an unspecific unlabelled probe (lane 9). I observed a supershift with an EVI1 antibody in DNA-protein complexes with the A-allele probe (lane 10), indicating that the minor allele of the 5'-UTR SNP increases binding affinity for the transcription factor EVI1 *in vitro*. (B) EMSAs for the intronic SNP showed higher protein affinity to probes containing the G-allele (lane 2) than the C-allele (lane 7). Protein binding of G-allele probes was competed by specific (lane 3) but not by unspecific unlabelled probes (lane 9). I performed supershift experiments with antibodies for the predicted transcription factors MZF1 and RBPJ. However, in my experiments none of the tested antibodies competed for binding and/or shifted the protein-DNA complex (lane 10; data not shown for RBPJ).

The results of luciferase reporter assays in cell lines representative of MKs and osteoblasts showed that the differential binding detected by EMSAs was functionally relevant and that both the 5'-UTR and intronic SNPs significantly reduced *RBM8A* promoter activity. The minor alleles, relative to the corresponding major alleles, were associated with significantly lower luciferase activity in human megakaryocytic CHRF-288-11 and DAMI cell lines and the mouse osteoblast cell line MC3T3 (**Figure 6-4**). No effect of the minor allele of the 5'-UTR SNP was observed in human endothelial EAHY926 and HEK293 cells. The minor allele of the intronic SNP did exert an effect in HEK293 cells but not in EAHY926 cells (**Figure 6-4**).

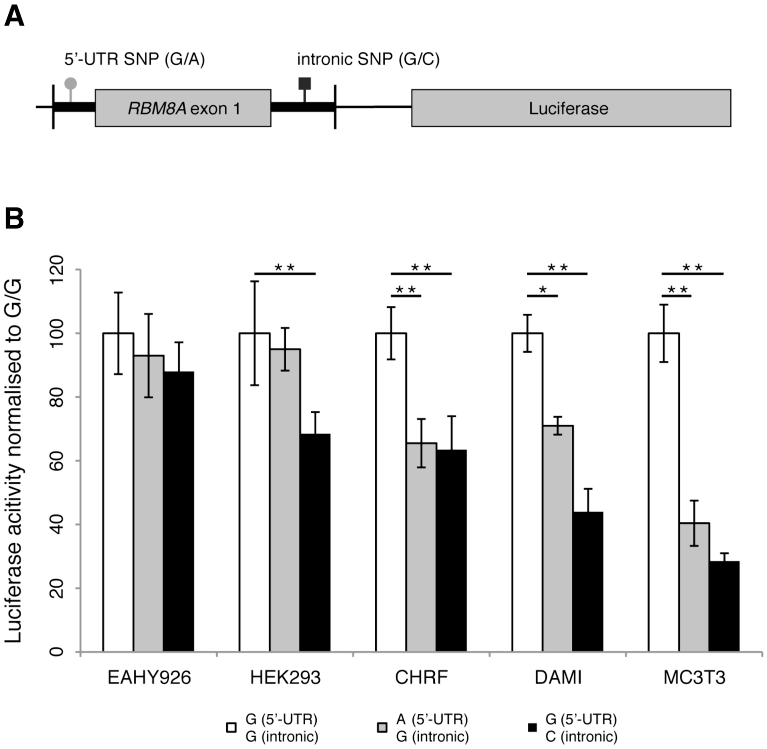


Figure 6-4. Luciferase reporter assays in cell lines representative of MKs (CHRF-288-11 and DAMI) and osteoblasts (MC3T3). (A) Schematic of the luciferase reporter construct with the 5'-UTR and intronic SNPs represented by circle and square symbols, respectively. (B) We observed significantly decreased *RBM8A* promoter activity for the minor alleles of both the 5'-UTR and intronic non-coding SNPs relative to the major alleles. No effect of the 5'-UTR SNP was observed in EAHY926 and HEK293 human endothelial cells. Error bars indicate standard deviations (s.d.). Statistical analysis was performed using the Tukey-Kramer multiple comparisons test, indicating *P<0.01 and **P<0.001. Luciferase activity was normalised with respect to the construct consisting of the major (G-) allele of both SNPs (indicated by G/G).

We next performed immunoblot staining of platelet lysates from seven TAR cases (all carrying the 1q21.1 deletion and either the 5'-UTR or intronic SNP), six unaffected parents (three with the 1q21.1 deletion, one heterozygous for the 5'-UTR SNP, one homozygous for the 5'-UTR SNP, and one compound heterozygous for the 5'-UTR and intronic SNPs), as well as six controls (**Figure 6-5 A**). Densitometry analysis of the protein blots showed a significant reduction in the levels of Y14, the protein encoded by *RBM8A*, in TAR cases compared to parental and healthy control samples (**Figure 6-5 B**).

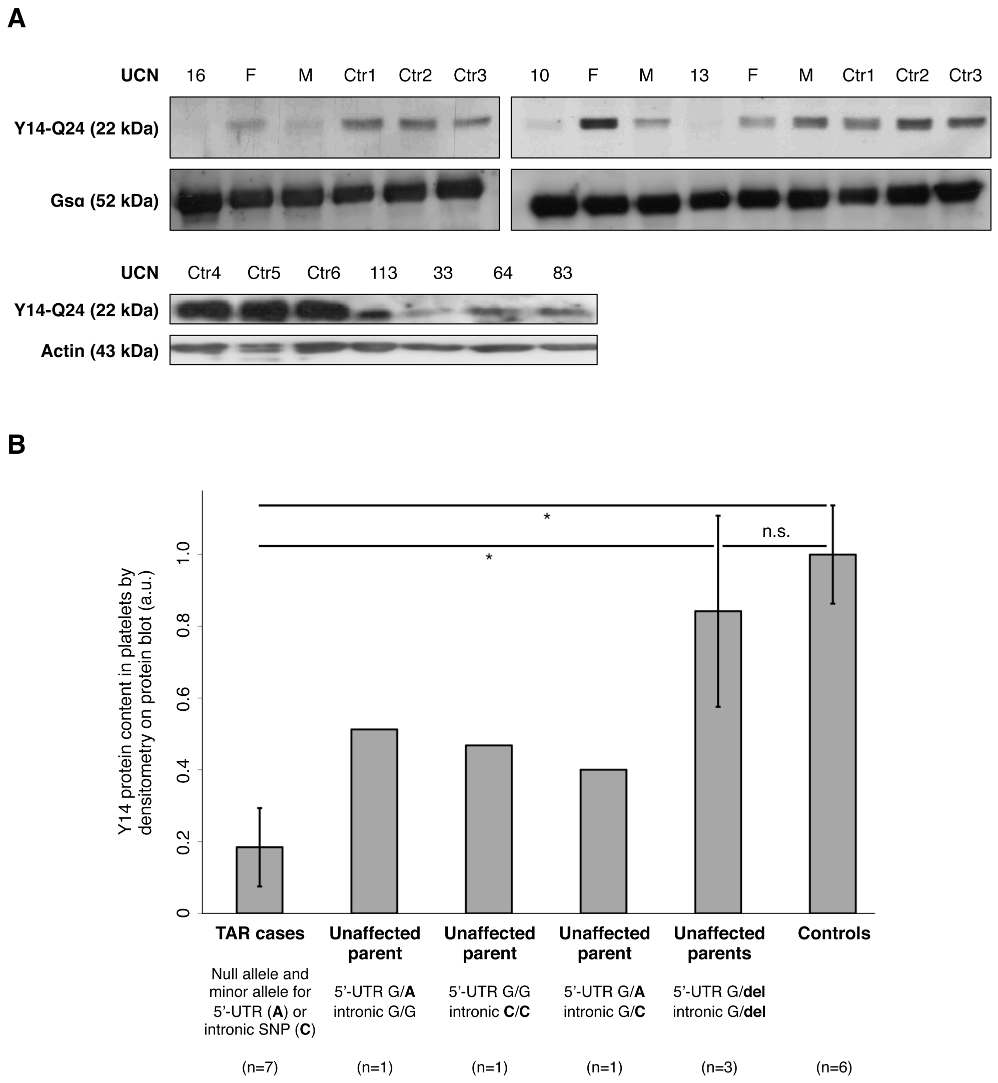


Figure 6-5. Immunoblot staining for Y14, the protein encoded by *RBM8A*, and densitometry analysis. (A) Western blot analyses of Y14 protein expression were performed in platelet lysates. We selected three TAR cases, all with the 1q21.1 deletion and the 5'-UTR SNP (UCNs 10, 13 and 16; **Appendix, Table 8-8**), and their six parents (labelled as 'F' and 'M' on the lane to the right of the TAR cases on the gel). In addition, we selected four TAR cases for which parental samples were not available: three with the 1q21.1 deletion and either the 5'-UTR SNP (UCNs 83 and 113) or the intronic SNP (UCN 64), and one with the 4 bp insertion in *RBM8A* in combination with the 5'-UTR SNP (UCN 33). Protein expression of Gsα or β-actin was used as a loading control. (B) Densitometry analysis showed significantly reduced Y14 protein levels in TAR cases compared to parental and control samples. Error bars indicate s.d. Statistical analysis was performed using the heteroscedastic t-test, marking **P*<0.01 and n.s. (not significant). Only genotype configurations indicated by lines were compared. The minor alleles of the 5'-UTR and intronic SNPs are shown in bold type. Abbreviations: UCN: unique case number; F: father; M: mother; Ctr: control; a.u.: arbitrary units.

Taken together, the genetic and biological data strongly supported our hypothesis that TAR results from insufficiency of the Y14 protein. The results from the luciferase assays suggested that the minor allele of the 5'-UTR SNP may cause decreased transcription relative to the major allele. Expression assays in platelet RNA samples from twelve healthy volunteers heterozygous for the 5'-UTR SNP, however, did not reveal a significant difference between transcript levels of the two alleles (*P*=0.91, paired t-test on allelic ratios; Albers et al., 2012). Therefore, what the exact mechanism is by which the non-coding SNPs lead to the decreased protein expression observed in TAR cases is still an open question.

We investigated whether there are any variants in strong LD with either the 5'-UTR SNP or the intronic SNP (Albers et al., 2012). We could identify no such candidates for the 5'-UTR SNP. In haplotype analysis using the four exome-sequenced TAR cases carrying the minor allele of the 5'-UTR SNP, this allele was present on at least two distinct haplotype backgrounds. This provided an additional line of evidence that the minor allele of the 5'-UTR SNP is causative in TAR. We identified a rare non-coding SNP (chr1:145,483,747C>T) 25 kb upstream of *RBM8A* in high LD with the intronic SNP. Sanger sequencing confirmed that this variant was present in all eleven genotyped TAR cases carrying the minor allele of the intronic SNP. The data from the ENCODE Project and our own FAIRE-seq open chromatin data in MKs indicated that this additional SNP was not located in a regulatory region, in contrast to the intronic SNP. Increased protein binding to the minor allele of the intronic SNP further corroborated the assumption that this particular SNP is causative. We cannot exclude the possibility that the 5'-UTR and intronic SNPs are not causative variants in TAR; however, in light of the genetic and biological evidence, we believe this is unlikely.

6.4. Discussion

Y14 is one of the four components of the exon-junction complex (EJC), which is involved in basic cellular functions, such as nuclear export and subcellular localisation of specific transcripts (Le Hir et al., 2001; Palacios et al., 2004), translational enhancement (Wiegand et al., 2003) and nonsense-mediated RNA decay (NMD) (Kim et al., 2001; Lykke-Andersen et al., 2001; Palacios et al., 2004). The *RBM8A* transcript is widely expressed (Salicioni et al., 2000) and is present in all haematopoietic lineages (Albers et al., 2012). Its encoded protein sequence is highly conserved between species (Albers et al., 2012). Given the important functions of the EJC, it is likely that a complete lack of Y14 in humans is not viable. Indeed, in *Drosophila melanogaster*, knockdown of its ortholog *tsu* leads to major defects in abdomen formation (Hachet & Ephrussi, 2001), and we found that knockdown of the orthologous *rbm8a* transcript in *Danio rerio* using antisense morpholinos resulted in extreme malformations and death at 2 d post-fertilisation (Albers et al., 2012). These findings are comparable with those from studies of a *Xenopus laevis* knockdown model of *Eif4a3*, which encodes an interacting EJC component, showing that EJC has a central role in vertebrate embryogenesis (Haremaki et al., 2010). Considered in this context, our results are compatible with both a dose-effect phenomenon and a lineage-dependent deficiency in Y14. The possibility of a dose-effect phenomenon is supported by the observation that simple haploinsufficiency is not sufficient to create an aberrant phenotype, as shown by the seemingly healthy carriers of the 1q21.1 deletion.

We did not observe an effect on platelet count for both the 5'-UTR and intronic SNPs in the respective 403 and 59 individuals of the Cambridge BioResource who carried the minor allele of each SNP (**Table 6-1**). This suggests that compound inheritance of a null allele together with the minor allele of one of the two regulatory SNPs brings Y14 levels below a critical threshold in certain tissues. Although the SNPs were directly genotyped, power to detect subtle effects of the SNPs on platelet count was limited due to the low MAFs of both SNPs. In addition, since the low platelet count in TAR cases often recovers in adolescence, and >94% of the genotyped individuals of the Cambridge BioResource were older than 20 years, the power to detect an effect of the SNPs on platelet count was expected to be limited.

The cell line-dependent effect shown in the luciferase assays was likely to be the result of differences in the regulation of *RBM8A* gene expression by combinatorial binding of transcription factors (including EVI1) in the context of the regulatory SNPs. An additional mechanism by which a deficiency in Y14 (and therefore in EJC function) may not be ubiquitous has been suggested by studies showing that NMD not only targets nonsense mRNAs but also regulates physiological mRNA abundance in a gene-specific manner (Nicholson et al., 2010). For example, haematopoietic-specific knockdown of *Upf2* in mouse, which encodes a core NMD component, resulted in complete disappearance of the haematopoietic stem cell compartment, whereas more differentiated cells were only mildly affected (Weischenfeldt et al., 2008). Finally, in addition to a tissue-dependent effect, it is possible that the regulatory SNPs have developmental stage-dependent consequences. In mouse, the *Mecom* gene encoding Evi1 is expressed in a transient manner in emerging limb buds (Perkins et al., 1991). This may provide an explanation for the skeletal abnormalities observed in TAR.

In conclusion, we applied next-generation sequencing to uncover the genetic basis of TAR syndrome, and identified a genetic mechanism of compound inheritance involving a null allele combined with a low-frequency regulatory variant. This compound inheritance mechanism reduces Y14 abundance, probably in a cell type- and developmental stage-dependent manner. Whether the same mechanism underlies other Mendelian disorders, in particular, other microdeletion syndromes showing variable penetrance and expression, remains to be established. However, these results highlight the importance of analysing regulatory regions even when searching for causative mutations in rare diseases. Although we have shown altered protein-binding affinity for the minor alleles of the regulatory SNPs, the mechanisms by which these SNPs lead to reduced levels of the Y14 protein in platelets are not clear and may be different for the 5'-UTR and intronic SNPs. Although genetic defects in the minor spliceosome (Edery et al., 2010; He et al., 2011), and NMD (Tarpey et al., 2007) have been linked to human disease, to the best of our knowledge, TAR syndrome is the first human disorder shown to be caused by a defect affecting one of the four EJC subunits.

CHAPTER 7

Conclusions and outlook.

124–129

7.1. Identifying candidate functional variants using maps of open chromatin

The biological interpretation of non-protein coding sequence variation associated with complex traits is challenging (Cooper & Shendure, 2011). It has been suggested that some of these non-coding variants influence phenotypic variation through regulation of gene expression.

Open chromatin assays provide an efficient and powerful tool for the identification of regulatory elements, as sites of open chromatin are general indicators of regulatory protein binding (**Section 1.7**). In recent studies, ~10% of the genome has been annotated as sites of open chromatin by DNase I- and/or FAIRE-seq across human cell types (Pennisi, 2011; Song et al., 2011). Despite the strong cross-validation of DNase I- and FAIRE-seq, NDRs specific to either assay are biologically relevant and functional, whereby DNase I- and FAIRE-specific sites tend to occur at TSSs and distal regions, respectively (Song et al., 2011). These differences may be the result of specific regulatory complexes that are bound in each NDR and influence the ability of formaldehyde to cross-link, or DNase I to cut.

However, open chromatin assays can neither directly reveal the function of the identified NDRs nor the transcription factors that are bound to them. Therefore, additional annotation data sets are required to corroborate the presence of a regulatory element and functionally classify it, e.g. enhancer vs. promoter, and type of transcription factor binding site. Here, *in silico* predictions may guide the identification of the specific transcription factor involved. In parallel, a multitude of high-resolution genome-wide annotation data sets across human cell types, developmental and disease states, as well as environmental conditions, will soon become publicly available through the efforts of large consortia including ENCODE and BLUEPRINT. These data sets can be used to further characterise regulatory elements that may be causally linked to phenotypic variation.

Open chromatin and ChIP protocols necessitate a large population of cells. In this thesis, at least 10 million cells were applied in each FAIRE assay (**Section 2.3**). Thus, the obtained regulatory maps have to be considered as an average of the chromatin status across a population of cells that may be heterogeneous. Often only a limited number of cells of a particular tissue or from a developmental stage are available. Protocols that require only a few thousand cells but still can be scaled to the whole genome are under development, with the aim of eventually probing single cells (Kalisky & Quake, 2011).

In **Chapter 3**, I intersected open chromatin maps in a megakaryocytic and an erythroblastoid cell line with CAD/MI risk alleles, but observed no overlap. One reason could be that the studied cell lines are not appropriate for studying the disease aetiology of that particular phenotype. Indeed, vascular smooth muscle cells, cardiomyocytes or other cell types may be more relevant. However, these cell types are difficult to obtain in adequate numbers from humans (especially from healthy individuals) for functional studies. Advances in induced pluripotent stem (iPS) cell technology may provide an attractive solution. This approach requires the heterologous overexpression of a few key transcription factors in mature cells for a period of few weeks. The mature cells, e.g. adult fibroblasts or keratinocytes, can be easily obtained through a skin biopsy and are then returned to an embryonic stem cell-like pluripotent state (de Souza, 2010). Even though differences compared to embryonic cells exist (Chin et al., 2009; Doi et al., 2009), iPS cell lines have the potential to differentiate into various different cell types. Therefore, this technology may hold great potential to generate essentially any cell type at various stages of cellular differentiation in large enough quantities to be used in experimental assays.

A key challenge in GWA follow-up studies is to link the putative regulatory element to a target gene or transcript. Candidate functional variants located within regulatory elements can be efficiently associated with transcript levels in eQTL studies. However, this method does not prove causality of the association. In contrast, allele-specific expression analyses overcome this limitation by directly linking risk alleles with transcript abundance, as shown at the *ZPBP2-GSDMB-ORMDL3* asthma and autoimmune disease risk locus (Verlaan et al., 2009). In GWA studies, the closest gene is typically reported as the most likely candidate. However, this assumption is rarely supported by experimental data in the absence of an eQTL, and many counter-examples in the literature exist (Spilianakis et al., 2005). Chromosome conformation capture techniques, i.e. 3C, 4C, 5C and Hi-C, offer an elegant way to link *cis*-regulatory elements with promoter regions of target genes (van Steensel & Dekker, 2010).

Low-throughput functional assays, e.g. EMSAs and luciferase reporter assays, are still the method-of-choice for establishing functionality of regulatory elements and variants, and studying conclusively their mechanisms (**Sections 5.3** and **6.3**). Following these *in vitro* experimental validation studies, regulatory elements may also be tested using *in vivo* assays. For example, the activity of tissue-specific enhancer sequences can be assessed in transgenic mouse assays (Visel et al., 2009).

Low-frequency and rare variants that impact phenotypic variation may reside on the same haplotype as the GWA tag SNP (**Section 1.1**). Therefore, the intersection of open chromatin or other regulatory maps with sequence variation depends on the availability of complete sequence information in order to make an informative assessment and reach a definite conclusion about causality. The sequence catalogue provided by the 1000 Genomes Project has made targeted resequencing for common variants (e.g. as described in **Section 5.2**) virtually obsolete for many populations. Initiatives such as the UK10K Project (http://www.uk10k.org/) expand this catalogue of sequence variation down to 0.1% allele frequency in individuals with European ancestry. Importantly, low-frequency and rare variants tend to be population-specific. That is, if associated with complex traits, these variants may have different effects in the different ethnic groups (Gravel et al., 2011; Bustamante et al., 2011). These population-specific effects are due to differences in allele frequency of the genetic markers in different populations.

7.2. Translating candidate functional variants using gene regulatory network approaches

Functional sequence variation may impact complex traits and diseases through the perturbations they cause to transcriptional and other molecular, cellular, tissue and organism network states (Sieberts & Schadt, 2007; Schadt, 2009). As DNA is transcribed into RNA and RNA is subsequently translated into protein, the molecular effects of DNA sequence variation on complex physiologic processes are mediated by transcriptional networks. These networks can be studied through integrative analyses of sequence variation, and cell type- and tissue-specific transcriptional and phenotypic data. This information benefits our understanding of the molecular mechanisms that drive complex trait variation and disease.

Exemplified by ENCODE, integration of multidimensional annotation data sets of the regulatory non-coding portion of the genome into a unified and quantitative framework can also help to improve predictions of genome function (**Figure 7-1**).

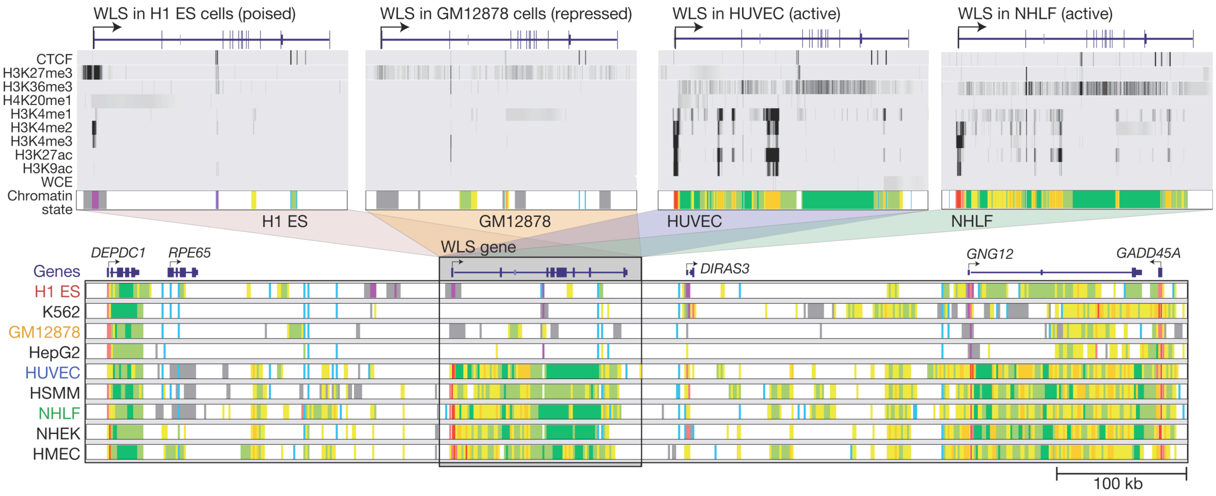


Figure 7-1. Example of the systematic genome annotation using chromatin maps across different cell types. Top: Annotation of the *WLS* gene locus using nine chromatin marks in four cell types. The complex chromatin status (grey scale) is summarised into functional modules and represented as a single, coloured annotation track. For example, red and orange indicate active promoters and strong enhancers, respectively. Bottom: Dynamic chromatin annotation of a 900 kb region centred on the *WLS* locus, showing activation and repression patterns for six genes with hundreds of gene regulatory elements. Figure taken from Ernst et al., 2011.

The characterisation of regulatory elements using chromatin maps contributes to the definition of the nuclear-based phenotype of the cell. In addition, nuclear-based phenotypes comprise other molecular interactions that occur on the chromatin level, such as transcript abundance and protein binding. The synergistic effect of nuclear-based phenotypes is expressed as cytoplasmic phenotypes. For example, transcript abundance results in protein abundance, but this depends on post-translational modifications. Other examples include metabolites or side-products that result from signalling or biochemical cascades (Dermitzakis, 2012).

The key aim of such integrative analysis is to identify relevant genes and effector cell types, and ultimately gain knowledge about pathways pertinent to the complex trait or disease of interest (**Figure 7-2**). This may benefit identification of possible drug targets or classification of the complex trait into sub-phenotypes, potentially resulting in strategies for disease diagnosis, prevention and therapy.

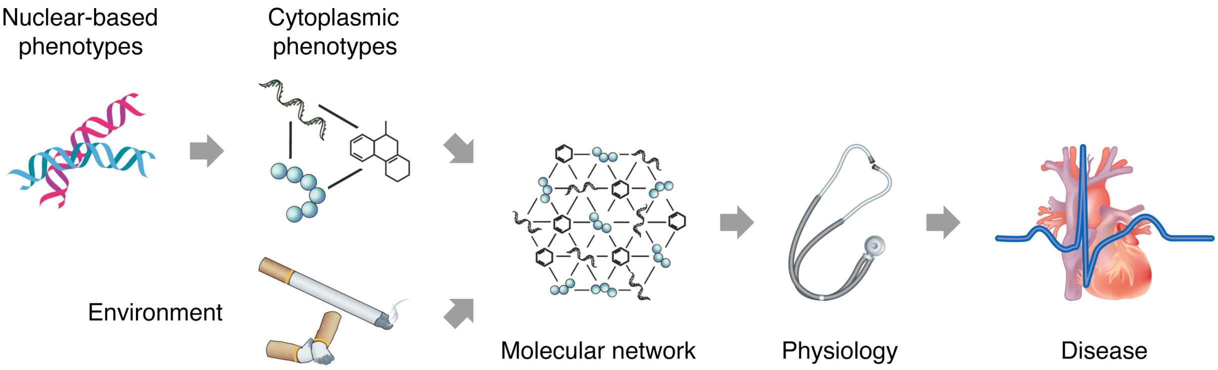


Figure 7-2. Integrative genomics and molecular networks. Molecular networks informed by nuclear-based and cytoplasmic phenotypes, such as DNA variation and protein abundance, define physiological states of human disease. Figure adapted from Schadt, 2009.

PIK3CG was identified as potential target gene underlying the association with platelet volume and function (**Chapter 5**). Indeed, platelets respond to a range of G-protein-coupled receptor agonists that activate PI3K signalling, including thrombin, thromboxane and ADP. Therefore, PI3Kγ may represent an attractive target in antiplatelet therapies (Rückle et al., 2006; Michelson, 2010). Our protein-protein interaction network, reported in **Section 5.6**, highlighted additional signalling pathways that may be investigated with respect to platelet characteristics and function.

A recent analysis investigated how many of the genes at GWA loci were amendable to pharmacological modulation using small molecules or biopharmaceuticals (i.e. therapeutic antibodies or protein therapeutics). The authors found that these genes were significantly more likely to be potentially 'druggable' and 'biopharmable' compared to the entire genome. The study also indicated opportunities for drug-repositioning (Sanseau et al., 2012).

Another route to translating disease-related genetic variants into patient benefits involves the identification of diagnostic biomarkers to inform disease processes. While variants in GWA studies generally have small effects and are therefore not directly suitable for diagnostics, variants identified in exome studies may be of interest. Indeed, the discovery of the genetic basis of TAR, described in **Chapter 6**, will make it simpler to more accurately diagnose future cases with a DNA test. In fact, such a test is currently being developed for the NHS as part of the international ThromboGenomics initiative (https://haemgen.haem.cam.ac.uk/thrombogenomics/). This diagnostic platform may lead to improvements in genetic counselling and patient care.

CHAPTER 8

Appendix.

130–165

Table 8-1. Genetic loci selected for the high-density DNA tiling array. The array contained (A) genetic loci associated with haematological and cardiovascular-related traits, and (B) lineage-specific reference genes. The criteria for selection are described in Section 2.4. Genomic coordinates were based on the human reference genome, build hg18 (NCBI build 36). Abbreviations: CAD: coronary artery disease; MI: myocardial infarction; MPV: mean platelet volume; PLT: platelet count; PLS: platelet signalling; WBC: white blood cell count; RBC: red blood cell count; MCV: mean corpuscular volume of erythrocytes; MCH: mean corpuscular haemoglobin; SBP: systolic blood pressure; DBP: diastolic blood pressure; HYP: hypertension. Key: [a]target gene ± 10 kb; [b]biological evidence for association.

(A) Association loci									
#	Gene(s)	GWA index SNP(s)		Trait(s)	Reference(s)	Genomic position			Interval (kb)
		ID	*Position*			*Chromosome*	*Start*	*End*	
1	*PCSK9*	rs11206510	55,268,627	Early-onset MI	Myocardial Infarction Genetics Consortium, 2009	1p32.3	55,240,000	55,300,000	60.00
2	*VAV3*[b]	rs17229705	107,940,484	PLS	Jones et al., 2009	1p13.3	107,890,000	108,055,000	165.00
3	*CELSR2, PSRC1, SORT1*	rs599839	109,623,689	CAD	Samani et al., 2007; Coronary Artery Disease Consortium, 2009	1p13.3	109,489,481	109,739,481	250.00
		rs646776	109,620,053	Early-onset MI	Myocardial Infarction Genetics Consortium, 2009				
4	*PEAR1*[b]	rs3737224	155,146,204	PLS	Jones et al., 2009	1q23.1	155,100,000	155,250,000	150.00
5	*FCER1G*[a,b]	rs3557	159,455,517	PLS	Jones et al., 2009	1q23.3	159,441,711	159,465,662	23.95
6	*DNM3*	rs10914144	170,216,373	MPV	Soranzo, Spector, et al., 2009	1q24.3	170,130,000	170,390,000	260.00
7	*TMCC2*	rs1668873	203,502,613	MPV	Soranzo, Spector, et al., 2009	1q32.1	203,445,000	203,540,000	95.00
8	*MIA3*	rs17465637	220,890,152	CAD/early-onset MI	Samani et al., 2007; Myocardial Infarction Genetics Consortium, 2009	1q41	220,678,228	221,078,228	400.00
		rs3008621	220,870,669	CAD	Coronary Artery Disease Consortium, 2009				
9	*EHD3*	rs647316	31,318,333	MPV	Soranzo, Spector, et al., 2009	2p23.1	31,305,000	31,346,000	41.00
10	*STK39*[b]	rs6749447	168,749,632	SBP/DBP	Wang et al., 2009	2q24.3	168,695,000	168,800,000	105.00
		rs3754777	168,724,160						
11	*WDR12*[a]	rs6725887	203,454,130	Early-onset MI	Myocardial Infarction Genetics Consortium, 2009	2q33.1	203,443,575	203,494,639	51.06
12	*gene desert*	rs2943634	226,776,324	CAD	Samani et al., 2007; Coronary Artery Disease Consortium, 2009	2q36.3	226,532,739	226,982,739	450.00
13	*ITPR1*[b]	rs17786144	4,804,575	PLS	Jones et al., 2009	3p26.2	4,760,000	4,860,000	100.00
14	*RAF1*[a,b]	rs3729931	12,601,516	PLS	Jones et al., 2009	3p25.1	12,590,108	12,690,678	100.57

15	*ULK4*	rs9815354	41,887,655	DBP	Newton-Cheh et al., 2009; Levy et al., 2009	3p22.1	41,710,000	41,970,000	260.00
16	*ARHGEF3*	rs12485738	56,840,816	MPV	Soranzo, Spector, et al., 2009; Meisinger et al., 2009	3p14.3	56,750,000	56,880,000	130.00
17	*MRAS*	rs9818870	139,604,812	CAD/MI	Erdmann et al., 2009	3q22.3	139,550,000	139,613,000	63.00
18	*P2RY12*[b]	rs1472122	152,517,292	PLS	Jones et al., 2009	3q25.1	152,410,000	152,580,000	170.00
19	*ITGA2*[b]	rs41305896	52,311,475	PLS	Jones et al., 2009	5q11.2	52,267,000	52,436,366	169.37
20	*PHACTR1*	rs12526453	13,035,530	Early-onset MI	Myocardial Infarction Genetics Consortium, 2009	6p24.1	12,985,000	13,060,000	75.00
21	*HFE*	rs1800562	26,201,120	MCV	Soranzo, Spector, et al., 2009	6p22.1	26,190,000	26,260,000	70.00
22	*BAK1*	rs210135	33,648,670	PLT	Soranzo, Spector, et al., 2009	6p21.31	33,644,000	33,665,000	21.00
23	*MAPK14*[b]	rs851007	36,172,014	PLS	Jones et al., 2009	6p21.31	36,050,000	36,210,000	160.00
24	*BSYL, CCND3*	rs11970772	42,033,268	MCV	Soranzo, Spector, et al., 2009	6p21.1	41,985,000	42,105,000	120.00
25	*HBS1L, MYB*	rs9402686	135,469,510	MCV	Soranzo, Spector, et al., 2009	6q23.3	135,400,000	135,582,003	182.00

26	*MTHFD1L*	rs6922269	151,294,678	CAD	The Wellcome Trust Case Control Consortium, 2007;
					Samani et al., 2007;
					Coronary Artery Disease Consortium, 2009
						6q25.1	151,199,579	151,399,579	200.00

27	*SLC22A3, LPAL2, LPA*	rs2048327-rs3127599-rs7767084-rs10755578 haplotype		CAD	Trégouët et al., 2009	6q25.3–q26	160,780,000	161,055,000	275.00
28	*CD36*[b]	rs1049654	80,113,391	PLS	Jones et al., 2009	7q21.11	80,045,000	80,137,000	92.00
29	*TFR2*	rs7385804	100,073,906	RBC	Soranzo, Spector, et al., 2009	7q22.1	100,050,000	100,100,000	50.00
30	*FLJ36031, PIK3CG*	rs342293	106,159,455	MPV	Soranzo, Spector, et al., 2009; Soranzo, Rendon, et al., 2009	7q22.3	106,085,000	106,190,000	105.00
31	*AK3, RCL1, JAK2* (b, for PLS)	rs385893	4,753,176	PLT	Soranzo, Spector, et al., 2009	9p24.1	4,730,000	5,050,000	320.00
		rs10429491	5,040,706	PLS	Jones et al., 2009				

32	*CDKN2A, CDKN2B*	rs1333049	22,115,503	CAD/MI	The Wellcome Trust Case Control Consortium, 2007;
					Samani et al., 2007; Schunkert et al., 2008;
					Coronary Artery Disease Consortium, 2009;
					Myocardial Infarction Genetics Consortium, 2009
						9p21.3	21,900,000	22,200,000	300.00

		rs4977574	22,088,574	Early-onset MI	Myocardial Infarction Genetics Consortium, 2009				
33	*CACNB2*	rs11014166	18,748,804	DBP	Newton-Cheh et al., 2009; Levy et al., 2009	10p12.33	18,570,000	18,840,000	270.00

34	*CXCL12/SDF1*	rs501120	44,073,873	CAD	Samani et al., 2007; Coronary Artery Disease Consortium, 2009	10q11.21	43,950,000	44,200,000	250.00
		rs1746048	44,095,830	Early-onset MI	Myocardial Infarction Genetics Consortium, 2009				
35	*JMJD1C*[a]	rs2393967	64,803,162	MPV	Soranzo, Spector, et al., 2009	10q21.2–q21.3	64,586,991	64,905,728	318.74
36	*CYP17A1, C10orf32*	rs1004467	104,584,497	SBP	Newton-Cheh et al., 2009; Levy et al., 2009	10q24.32	104,520,000	104,690,000	170.00
37	*BET1L, SIRT3, PSMD13*	rs11602954	192,856	MPV	Soranzo, Spector, et al., 2009	11p15.5	180,000	240,000	60.00
38	*PLEKHA7*	rs381815	16,858,844	SBP	Newton-Cheh et al., 2009; Levy et al., 2009	11p15.1	16,795,000	16,960,000	165.00
39	*ATP2B1*	rs2681492	88,537,220	SBP	Newton-Cheh et al., 2009; Levy et al., 2009	12q21.33	88,460,000	88,650,000	190.00
		rs2681472	88,533,090	HYP/DBP	Newton-Cheh et al., 2009; Levy et al., 2009				
40	*SH2B3/LNK, ATXN2*	rs3184504	110,368,991	MI; SBP/DBP	Gudbjartsson et al., 2009; Newton-Cheh et al., 2009; Levy et al., 2009	12q24.12	110,310,000	110,570,000	260.00
		rs11065987	110,556,807	PLT; CAD/MI	Soranzo, Spector, et al., 2009				
41	*PTPN11*[a]	rs11066301	111,355,755	PLT; CAD/MI	Soranzo, Spector, et al., 2009	12q24.13	111,330,919	111,442,100	111.18
42	*TBX3, TBX5*	rs2384550	113,837,114	DBP	Newton-Cheh et al., 2009; Levy et al., 2009	12q24.21	113,810,000	113,930,000	120.00
43	*HNF1A/TCF1, C12orf43*	rs2259816	119,919,970	CAD/MI	Erdmann et al., 2009	12q24.31	119,850,000	119,980,000	130.00
44	*WDR66*	rs7961894	120,849,966	MPV	Meisinger et al., 2009; Soranzo, Spector, et al., 2009	12q24.31	120,760,000	120,970,000	210.00
45	*TPM1*	rs11071720	61,129,049	MPV	Soranzo, Spector, et al., 2009	15q22.2	61,080,000	61,160,000	80.00
46	*SMAD3*[b]	rs17228212	65,245,693	CAD	Samani et al., 2007	15q22.33–q23	65,150,000	65,350,000	200.00
47	*CSK, ULK3*	rs6495122	72,912,698	DBP	Newton-Cheh et al., 2009; Levy et al., 2009	15q24.1	72,810,000	73,030,000	220.00
48	*CDH13*	rs11646213	81,200,152	HYP/SBP/DBP	Org et al., 2009	16q23.3	81,190,000	81,240,000	50.00
49	*MAP2K4*[a,b]	rs41307923	11,968,409	PLS	Jones et al., 2009	17p12	11,854,860	11,997,776	142.92
50	*TAOK1*	rs2138852	24,727,475	MPV	Meisinger et al., 2009; Soranzo, Spector, et al., 2009	17q11.2	24,680,000	24,930,000	250.00
51	*GSDMA, ORMDL3*	rs17609240	35,364,215	WBC	Soranzo, Spector, et al., 2009	17q12	35,314,374	35,387,545	73.17
52	*CD226*	rs893001	65,667,825	MPV	Soranzo, Spector, et al., 2009	18q22.2	65,620,000	65,730,000	110.00
53	*MAP2K2*[b]	rs350916	4,045,775	PLS	Jones et al., 2009	19p13.3	4,025,000	4,095,000	70.00
54	*LDLR*	rs1122608	11,024,601	Early-onset MI	Myocardial Infarction Genetics Consortium, 2009	19p13.2	11,010,000	11,120,000	110.00
55	*AKT2*[b]	rs41275750	45,429,935	PLS	Jones et al., 2009	19q13.2	45,380,000	45,520,000	140.00
56	*APOC1, APOC4, APOE*[b]	rs4420638	50,114,786	CAD	The Wellcome Trust Case Control Consortium, 2007	19q13.32	50,000,000	50,230,000	230.00
57	*GP6*[a,b]	rs1613662	60,228,407	PLS	Jones et al., 2009	19q13.42	60,206,885	60,251,444	44.56
58	*SIRPA*	rs6136489	1,871,734	MPV	Soranzo, Spector, et al., 2009	20p13	1,820,000	1,950,000	130.00

59	*SLC5A3, MRPS6, KCNE2*	rs9982601	34,520,998	Early-onset MI	Myocardial Infarction Genetics Consortium, 2009	21q22.11	34,340,000	34,530,000	190.00
60	*GNAZ*[b]	rs3788337	21,742,017	PLS	Jones et al., 2009	22q11.22	21,720,000	21,840,000	120.00
61	*FBXO7*	rs9609565	31,197,528	MCV	Soranzo, Spector, et al., 2009	22q12.3	31,186,000	31,250,000	64.00
62	*TMPRSS6*	rs5756506	35,797,338	MCH	Soranzo, Spector, et al., 2009	22q12.3	35,730,000	35,830,000	100.00

(B) Lineage-specific reference genes					
Lineage	Gene	Genomic position			Interval (kb)
		Chromosome	*Start*	*End*	
monocytic	*ASGR2*	17p13.1	6,943,365	6,960,852	17.49
	CD163	12p13.31	7,512,676	7,549,681	37.01
	FER1L3	10q23.33	95,054,176	95,234,029	179.85
	KLF4	9q31.2	109,284,956	109,293,576	8.62
	PID1	2q36.3	229,594,933	229,846,301	251.37
	RIN2	20p11.23	19,816,210	19,933,100	116.89
	SLC46A2	9q32	114,679,021	114,694,866	15.85
	TMEM176A	7q36.1	150,126,787	150,135,141	8.35
erythroblastoid	*CA1*	8q21.2	86,425,709	86,479,594	53.89
	CALB2	16q22.3	69,948,127	69,983,842	35.72
	EPB42	15q15.2	41,274,720	41,302,773	28.05
	ERAF	16p11.2	31,444,704	31,449,625	4.92
	FAM83A	8q24.13	124,261,933	124,293,499	31.57
	GDF15	19p13.11	18,355,968	18,362,986	7.02
	HBZ	16p13.3	140,854	146,504	5.65
	LOC51252	2q11.2	96,903,349	96,929,558	26.21
megakaryocytic	*ADCY6*	12q13.12	47,444,248	47,471,087	26.84
	CMTM5	14q11.2	22,913,857	22,920,821	6.96
	DDEF2	2p25.1	9,262,345	9,465,257	202.91
	LY6G6D	6p21.33	31,789,112	31,795,560	6.45
	MEIS1	2p14	66,514,036	66,655,395	141.36
	MYLK	3q21.1	124,811,833	125,087,839	276.01
	NFIB	9p23–p22.3	14,069,847	14,305,945	236.10
	SELP	1q24.2	167,822,711	167,868,001	45.29

Table 8-2. Ontology analysis of genes flanking FAIRE peaks using GREAT. I report the ontology of genes flanking ('stringent') FAIRE peaks of both lower (Bin 2) and higher peak score (Bin 4). For these gene sets, the GO biological process and the mouse phenotype enrichment are reported, i.e. the top 15 terms with binominal raw *P*<10^{-4}.

Term name	Binom. raw *P*-val	Binom. FDR Q-val	Binom. fold enrichment	Binom. observed region hits
(A) EB Bin 2				
GO biological process				
erythrocyte differentiation	3.44x10^{-08}	1.30x10^{-05}	2.727	40
erythrocyte homoeostasis	1.82x10^{-07}	5.02x10^{-05}	2.488	42
iron ion homoeostasis	6.99x10^{-05}	6.96x10^{-03}	2.299	28
Mouse phenotype				
anaemia	9.55x10^{-17}	1.55x10^{-13}	2.063	160
abnormal mean corpuscular volume	3.69x10^{-10}	8.24x10^{-08}	3.117	42
abnormal erythroid progenitor cell morphology	8.55x10^{-10}	1.68x10^{-07}	2.818	47
decreased mean corpuscular volume	1.84x10^{-09}	3.50x10^{-07}	4.118	27
decreased erythrocyte cell number	2.00x10^{-09}	3.71x10^{-07}	2.143	76
microcytosis	5.08x10^{-09}	8.24x10^{-07}	6.748	16
increased liver iron level	5.80x10^{-09}	8.95x10^{-07}	5.169	20
abnormal reticulocyte morphology	3.45x10^{-08}	4.48x10^{-06}	2.615	43
abnormal liver iron level	7.14x10^{-08}	8.74x10^{-06}	4.227	21
spherocytosis	1.35x10^{-07}	1.51x10^{-05}	5.685	15
abnormal pro-erythroblast morphology	1.43x10^{-07}	1.55x10^{-05}	2.216	54
abnormal mean corpuscular haemoglobin	2.29x10^{-07}	2.18x10^{-05}	3.323	26
microcytic anaemia	2.64x10^{-07}	2.45x10^{-05}	5.029	16
abnormal mast cell morphology	8.17x10^{-06}	5.40x10^{-04}	2.554	29
abnormal erythrocyte physiology	1.22x10^{-05}	7.60x10^{-04}	3.135	20
(B) EB Bin 4				
GO biological process				
intracellular transport	4.05x10^{-12}	4.84x10^{-09}	2.425	76
protein catabolic process	1.97x10^{-11}	1.76x10^{-08}	2.963	51
translation	4.04x10^{-11}	2.89x10^{-08}	3.448	40
proteolysis involved in cellular protein catabolic process	4.60x10^{-11}	3.00x10^{-08}	2.968	49
cellular protein catabolic process	6.04x10^{-11}	3.61x10^{-08}	2.944	49
ubiquitin-dependent protein catabolic process	2.38x10^{-10}	1.14x10^{-07}	2.986	45
modification-dependent protein catabolic process	2.58x10^{-10}	1.16x10^{-07}	2.978	45
establishment of localisation in cell	3.46x10^{-10}	1.46x10^{-07}	2.017	92
DNA metabolic process	8.33x10^{-09}	3.14x10^{-06}	2.290	59
cell cycle process	6.47x10^{-08}	2.32x10^{-05}	2.080	64
cellular macromolecule catabolic process	7.64x10^{-08}	2.61x10^{-05}	2.277	52

DNA replication	9.53x10^{-08}	3.11x10^{-05}	3.190	29
macromolecule catabolic process	1.77x10^{-07}	5.30x10^{-05}	2.158	55
translational elongation	2.67x10^{-07}	7.66x10^{-05}	5.002	16
cell cycle phase	1.03x10^{-06}	2.45x10^{-04}	2.141	49
(C) MK Bin 2				
GO biological process				
lamellipodium assembly	4.01x10^{-05}	7.99x10^{-03}	3.351	16
Mouse phenotype				
abnormal platelet physiology	5.66x10^{-10}	2.04x10^{-07}	2.716	51
increased IgG level	2.61x10^{-08}	5.28x10^{-06}	2.055	72
increased T cell proliferation	4.03x10^{-06}	4.93x10^{-04}	2.128	45
thymus hyperplasia	4.27x10^{-06}	5.13x10^{-04}	2.893	25
(D) MK Bin 4				
GO biological process				
cell cycle	3.14x10^{-12}	5.63x10^{-09}	2.304	84
cell cycle process	2.74x10^{-11}	3.28x10^{-08}	2.501	66
intracellular transport	3.24x10^{-08}	2.32x10^{-05}	2.195	59
protein folding	5.10x10^{-08}	3.32x10^{-05}	4.107	22
regulation of cell cycle	7.24x10^{-07}	3.24x10^{-04}	2.125	51
mitotic cell cycle	1.19x10^{-06}	4.06x10^{-04}	2.341	40
DNA metabolic process	1.90x10^{-06}	5.46x10^{-04}	2.127	47
positive regulation of cell activation	6.26x10^{-05}	6.32x10^{-03}	2.686	21
(E) MO Bin 2				
GO biological process				
immune response	1.05x10^{-51}	3.75x10^{-48}	2.203	451
cell activation	3.13x10^{-29}	3.20x10^{-26}	2.076	288
leukocyte activation	5.29x10^{-29}	4.75x10^{-26}	2.144	265
regulation of cytokine production	9.71x10^{-25}	4.35x10^{-22}	2.286	195
inflammatory response	1.41x10^{-23}	4.40x10^{-21}	2.038	240
lymphocyte activation	1.05x10^{-19}	1.94x10^{-17}	2.026	201
activation of pro-apoptotic gene products	6.50x10^{-17}	8.63x10^{-15}	5.363	40
positive regulation of cytokine production	1.20x10^{-16}	1.48x10^{-14}	2.492	107
regulation of transcription factor import into nucleus	1.21x10^{-14}	1.17x10^{-12}	3.687	51
regulation of vascular endothelial growth factor production	2.67x10^{-13}	2.40x10^{-11}	5.916	28
immune effector process	2.76x10^{-13}	2.44x10^{-11}	2.165	111
leukocyte-mediated immunity	8.98x10^{-13}	7.66x10^{-11}	2.553	76
cytokine-mediated signalling pathway	1.25x10^{-12}	1.03x10^{-10}	2.723	67
myeloid leukocyte activation	7.05x10^{-12}	5.11x10^{-10}	2.684	64
B cell activation	1.20x10^{-11}	8.20x10^{-10}	2.131	99

Mouse phenotype				
abnormal immune cell physiology	5.86x10^{-137}	3.80x10^{-133}	2.345	1,025
abnormal adaptive immunity	6.04x10^{-137}	1.96x10^{-133}	2.342	1,027
abnormal cell-mediated immunity	1.38x10^{-136}	2.99x10^{-133}	2.343	1,024
abnormal leukocyte physiology	2.28x10^{-134}	3.70x10^{-131}	2.334	1,015
abnormal antigen presenting cell physiology	4.01x10^{-111}	4.33x10^{-108}	2.524	736
abnormal lymphocyte physiology	6.53x10^{-107}	4.23x10^{-104}	2.415	767
abnormal immune serum protein physiology	8.74x10^{-103}	5.15x10^{-100}	2.400	747
abnormal lymphocyte morphology	6.48x10^{-98}	3.23x10^{-95}	2.185	857
abnormal mononuclear leukocyte morphology	2.84x10^{-94}	1.08x10^{-91}	2.047	952
abnormal bone marrow cell morphology/development	1.93x10^{-92}	6.96x10^{-90}	2.112	872
abnormal cytokine secretion	5.21x10^{-84}	1.78x10^{-81}	2.698	500
abnormal B cell physiology	2.70x10^{-80}	8.33x10^{-78}	2.627	500
abnormal leukopoiesis	6.81x10^{-79}	2.01x10^{-76}	2.155	717
abnormal myeloblast morphology/development	1.85x10^{-78}	5.20x10^{-76}	2.153	715
abnormal mononuclear leukocyte differentiation	2.01x10^{-78}	5.43x10^{-76}	2.196	685
(F) MO Bin 4				
GO biological process				
response to wounding	4.08x10^{-11}	5.85x10^{-08}	2.138	90
translation	8.91x10^{-09}	5.81x10^{-06}	2.766	42
positive regulation of cytokine production	1.04x10^{-07}	4.67x10^{-05}	3.449	26
wound healing	2.42x10^{-07}	8.66x10^{-05}	2.386	44
positive regulation of cellular protein metabolic process	6.56x10^{-07}	2.24x10^{-04}	2.225	47
response to other organism	1.24x10^{-06}	3.56x10^{-04}	2.095	51
cell activation	2.77x10^{-06}	6.86x10^{-04}	2.052	50
regulation of myeloid cell differentiation	7.24x10^{-06}	1.53x10^{-03}	2.865	24
leukocyte-mediated immunity	9.22x10^{-06}	1.79x10^{-03}	3.443	18
positive regulation vascular endothelial growth factor production	1.12x10^{-05}	2.01x10^{-03}	9.622	7
response to bacterium	1.85x10^{-05}	2.71x10^{-03}	2.158	37
regulation of vascular endothelial growth factor production	2.61x10^{-05}	2.97x10^{-03}	8.422	7
neutrophil chemotaxis	5.71x10^{-05}	4.65x10^{-03}	6.248	8
adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	6.51x10^{-05}	5.07x10^{-03}	3.360	15
adaptive immune response	7.89x10^{-05}	5.90x10^{-03}	3.301	15
Mouse phenotype				
abnormal cell-mediated immunity	3.72x10^{-28}	2.41x10^{-24}	2.423	186
abnormal immune cell physiology	3.77x10^{-28}	1.22x10^{-24}	2.423	186
abnormal adaptive immunity	5.38x10^{-28}	1.16x10^{-24}	2.415	186
abnormal leukocyte physiology	9.80x10^{-27}	1.59x10^{-23}	2.384	182
abnormal blood cell morphology/development	6.39x10^{-25}	8.29x10^{-22}	2.071	224
abnormal haematopoiesis	1.43x10^{-23}	1.16x10^{-20}	2.041	219

abnormal antigen presenting cell physiology	1.03x10^{-22}	7.40x10^{-20}	2.597	133
abnormal lymphocyte morphology	5.19x10^{-22}	2.80x10^{-19}	2.309	159
abnormal immune system cell morphology	1.24x10^{-21}	6.18x10^{-19}	2.101	189
abnormal leukocyte morphology	1.74x10^{-20}	8.06x10^{-18}	2.069	185
abnormal mononuclear leukocyte morphology	3.74x10^{-20}	1.62x10^{-17}	2.118	173
abnormal innate immunity	5.37x10^{-20}	2.18x10^{-17}	3.020	92
abnormal leukocyte cell number	6.94x10^{-18}	2.50x10^{-15}	2.107	155
abnormal bone marrow cell morphology/development	1.00x10^{-17}	3.43x10^{-15}	2.110	153
abnormal lymphocyte cell number	2.70x10^{-17}	8.76x10^{-15}	2.246	131
(G) CHRF Bin 2				
GO biological process				
amino acid import	1.39x10^{-06}	5.37x10^{-05}	3.093	25
lymphocyte activation involved in immune response	5.13x10^{-06}	1.61x10^{-04}	2.573	30
natural killer cell activation	5.35x10^{-05}	1.13x10^{-03}	2.338	28
endothelial cell proliferation	8.85x10^{-05}	1.71x10^{-03}	2.165	31
Mouse phenotype				
decreased platelet cell number	1.07x10^{-18}	8.80x10^{-17}	2.073	181
abnormal platelet physiology	1.22x10^{-16}	8.24x10^{-15}	2.107	153
decreased haemoglobin content	1.49x10^{-16}	9.87x10^{-15}	2.192	139
abnormal erythroid progenitor cell morphology	4.56x10^{-14}	2.43x10^{-12}	2.173	118
abnormal splenic cell ratio	7.66x10^{-13}	3.38x10^{-11}	2.168	107
abnormal embryonic haematopoiesis	6.51x10^{-12}	2.48x10^{-10}	2.065	110
abnormal mean corpuscular volume	6.51x10^{-12}	2.47x10^{-10}	2.189	96
decreased mean corpuscular volume	1.53x10^{-11}	5.63x10^{-10}	2.765	59
increased IgG2a level	5.27x10^{-11}	1.79x10^{-09}	2.068	100
increased monocyte cell number	1.46x10^{-10}	4.82x10^{-09}	2.118	90
abnormal strial intermediate cells	4.73x10^{-10}	1.48x10^{-08}	2.545	58
increased eosinophil cell number	5.07x10^{-10}	1.58x10^{-08}	2.234	75
abnormal intraocular pressure	5.74x10^{-10}	1.77x10^{-08}	2.385	65
decreased erythroid progenitor cell number	6.16x10^{-10}	1.89x10^{-08}	2.148	81
abnormal prostaglandin level	9.26x10^{-09}	2.40x10^{-07}	2.299	60
(H) CHRF Bin 4				
GO biological process				
translation	1.97x10^{-11}	4.15x10^{-09}	2.184	91
regulation of protein kinase B signalling cascade	9.29x10^{-09}	1.11x10^{-06}	3.691	28
translational elongation	5.38x10^{-08}	5.01x10^{-06}	2.959	34
protein folding	6.09x10^{-08}	5.53x10^{-06}	2.317	52
nuclear export	3.05x10^{-07}	2.32x10^{-05}	3.270	26
positive regulation of protein kinase B signalling cascade	4.33x10^{-07}	3.01x10^{-05}	3.512	23
regulation of endopeptidase activity	4.99x10^{-07}	3.34x10^{-05}	2.261	47
regulation of caspase activity	5.04x10^{-07}	3.35x10^{-05}	2.283	46

release of cytochrome c from mitochondria	8.06x10^{-07}	4.98x10^{-05}	4.611	16
regulation of peptidase activity	8.16x10^{-07}	4.96x10^{-05}	2.220	47
protein mono-ubiquitination	8.32x10^{-07}	5.01x10^{-05}	6.290	12
histone modification	8.94x10^{-07}	5.26x10^{-05}	2.066	55
cofactor biosynthetic process	1.29x10^{-06}	7.02x10^{-05}	2.451	37
covalent chromatin modification	1.74x10^{-06}	9.03x10^{-05}	2.019	55
double-strand break repair	2.21x10^{-06}	1.12x10^{-04}	2.683	30
(I) K562 Bin 2				
GO biological process				
erythrocyte differentiation	1.53x10^{-18}	2.03x10^{-16}	2.561	115
erythrocyte homoeostasis	7.48x10^{-16}	7.05x10^{-14}	2.303	119
peptidyl-tyrosine phosphorylation	8.16x10^{-11}	4.03x10^{-09}	2.034	102
peptidyl-tyrosine modification	1.18x10^{-10}	5.62x10^{-09}	2.005	104
organ regeneration	1.35x10^{-10}	6.40x10^{-09}	2.083	94
regulation of actin filament depolymerisation	2.55x10^{-08}	7.98x10^{-07}	2.763	40
actin filament capping	2.75x10^{-08}	8.50x10^{-07}	2.941	36
negative regulation of actin filament depolymerisation	8.49x10^{-08}	2.44x10^{-06}	2.716	38
Mouse phenotype				
abnormal platelet physiology	9.73x10^{-56}	7.01x10^{-53}	3.427	234
abnormal haematopoietic system physiology	4.98x10^{-51}	1.62x10^{-48}	2.165	471
decreased immature B cell number	1.13x10^{-43}	2.37x10^{-41}	2.496	300
abnormal erythrocyte morphology	1.25x10^{-36}	1.80x10^{-34}	2.053	379
abnormal megakaryocyte progenitor cell morphology	4.15x10^{-36}	5.85x10^{-34}	2.188	322
decreased pre-B cell number	9.74x10^{-35}	1.26x10^{-32}	2.596	220
abnormal pre-B cell morphology	2.47x10^{-32}	2.92x10^{-30}	2.357	245
abnormal erythrocyte cell number	1.38x10^{-31}	1.54x10^{-29}	2.099	308
decreased erythrocyte cell number	9.01x10^{-31}	9.74x10^{-29}	2.285	248
abnormal megakaryocyte morphology	1.04x10^{-29}	1.02x10^{-27}	2.083	294
abnormal erythroid progenitor cell morphology	2.09x10^{-29}	1.97x10^{-27}	2.938	150
extramedullary haematopoiesis	2.68x10^{-29}	2.49x10^{-27}	2.292	234
cortical renal glomerulopathies	5.85x10^{-28}	5.26x10^{-26}	2.326	216
abnormal pro-B cell morphology	2.07x10^{-22}	1.37x10^{-20}	2.123	208
decreased erythroid progenitor cell number	2.70x10^{-22}	1.77x10^{-20}	3.018	107
(J) K562 Bin 4				
cell cycle	1.85x10^{-21}	2.22x10^{-18}	2.009	217
intracellular transport	4.36x10^{-19}	2.24x10^{-16}	2.123	169
cell cycle process	9.49x10^{-17}	3.24x10^{-14}	2.047	160
mitotic cell cycle	2.03x10^{-16}	6.61x10^{-14}	2.331	118
mRNA transport	2.99x10^{-16}	9.33x10^{-14}	4.532	45
nucleobase, nucleoside, nucleotide and nucleic acid transport	2.60x10^{-14}	6.66x10^{-12}	3.596	51

RNA localisation	3.82x10^{-14}	9.45x10^{-12}	3.736	48
RNA transport	3.83x10^{-14}	9.14x10^{-12}	3.801	47
RNA processing	1.91x10^{-13}	3.91x10^{-11}	2.008	132
cell cycle phase	8.83x10^{-13}	1.47x10^{-10}	2.047	119
translation	7.52x10^{-12}	1.10x10^{-09}	2.477	73
ncRNA metabolic process	1.71x10^{-11}	2.27x10^{-09}	2.829	56
protein targeting to mitochondrion	7.29x10^{-11}	8.43x10^{-09}	6.306	21
mitochondrion organisation	1.83x10^{-09}	1.73x10^{-07}	2.815	45
translational elongation	2.71x10^{-09}	2.53x10^{-07}	3.693	30

Table 8-3. SNPs associated with platelet and erythrocyte phenotypes located in open chromatin in primary human MKs, EBs and MOs. Regions of open chromatin overlapping SNPs at (A) platelet (Gieger et al., 2011) and (B) erythrocyte (van der Harst et al., 2012) QTLs were determined using the software F-Seq (Boyle, Guinney, et al., 2008), applying the 'stringent' cut-off for peak calling. I retrieved additional candidate functional variants at (C) platelet and (D) erythrocyte QTLs by applying the 'moderate' cut-off for peak calling, including the SNP rs342293 (highlighted in blue). The functional mechanism of this SNP is described in detail in **Chapter 5**. Allele frequencies and r^2 values were retrieved from the European samples of the 1000 Genomes Project (interim phase I release of June 2011) (The 1000 Genomes Project Consortium, 2010). Genomic coordinates were mapped to the human reference genome, build hg19 (NCBI build 37).

(A) Platelet phenotypes, stringent peak calling														
SNP in open chromatin				GWA sentinel SNP				Proxy↔sentinel		FAIRE peak				
Chr	*Position*	*rs number*	*Allele freq*	*Chr*	*Position*	*rs number*	*Gene locus*	*r^2*	*Distance (bp)*	*Cell type*	*Start position*	*End position*	*F-Seq score*	*Bin*
1	171,939,290	rs2038479	0.827	1	171,949,750	rs10914144	*DNM3*	0.973	10,460	MK	171,938,981	171,939,669	0.1322	3
1	171,939,641	rs2038480	0.827	1	171,949,750	rs10914144	*DNM3*	0.973	10,109	MK	171,938,981	171,939,669	0.1322	3
1	205,185,104	rs12144980	0.381	1	205,244,953	rs1172130	*TMCC2*	0.833	59,849	EB	205,184,914	205,185,139	0.0541	1
1	205,225,457	rs1172147	0.382	1	205,244,953	rs1172130	*TMCC2*	0.839	19,496	EB	205,224,992	205,225,502	0.1785	3
1	205,225,457	rs1172147	0.382	1	205,244,953	rs1172130	*TMCC2*	0.839	19,496	MO	205,225,437	205,225,541	0.0183	1
1	205,254,238	rs1151787	0.362	1	205,244,953	rs1172130	*TMCC2*	0.928	9,285	MO	205,253,848	205,254,286	0.0303	3
2	43,711,523	rs998768	0.094	2	43,687,879	rs17030845	*THADA*	0.908	23,644	MO	43,711,507	43,711,739	0.0220	2
2	43,713,808	rs17030925	0.087	2	43,687,879	rs17030845	*THADA*	0.983	25,929	MO	43,713,592	43,713,949	0.0258	2
2	43,761,299	rs17031005	0.086	2	43,687,879	rs17030845	*THADA*	1.000	73,420	EB	43,761,285	43,761,352	0.0398	1
2	43,764,442	rs17031016	0.090	2	43,687,879	rs17030845	*THADA*	0.952	76,563	MO	43,764,269	43,764,701	0.0336	3
3	12,267,648	rs7616006	0.442	3	12,267,648	rs7616006	*SYN2*	1.000	0	MO	12,267,593	12,267,780	0.0200	1
3	12,267,780	rs7650082	0.443	3	12,267,648	rs7616006	*SYN2*	0.995	132	MO	12,267,593	12,267,780	0.0200	1
3	18,250,509	rs7618405	0.213	3	18,311,412	rs7641175	*SATB1*	0.909	60,903	MK	18,250,349	18,250,815	0.1633	4
3	18,250,509	rs7618405	0.213	3	18,311,412	rs7641175	*SATB1*	0.909	60,903	MO	18,250,463	18,250,606	0.0187	1
3	122,833,003	rs3804749	0.567	3	122,839,876	rs3792366	*PDIA5*	0.973	6,873	MK	122,832,668	122,833,082	0.1414	4
3	124,339,527	rs6771416	0.481	3	124,340,222	rs10512627	*KALRN*	1.000	695	EB	124,339,467	124,339,671	0.0492	1
3	124,339,527	rs6771416	0.481	3	124,340,222	rs10512627	*KALRN*	1.000	695	MK	124,339,398	124,339,684	0.0598	2
4	6,891,435	rs11734099	0.160	4	6,891,519	rs11734132	*KIAA0232*	0.990	84	MK	6,891,308	6,891,504	0.0457	1
4	6,891,455	rs11731274	0.163	4	6,891,519	rs11734132	*KIAA0232*	0.990	64	MK	6,891,308	6,891,504	0.0457	1

5	158,635,102	rs6556405	0.243	5	158,604,963	rs10076782	*RNF145*	0.951	30,139	MO	158,634,891	158,635,107	0.0208	1
6	25,537,194	rs214060	0.447	6	25,548,288	rs441460	*LRRC16*	0.984	11,094	MK	25,536,943	25,537,478	0.1216	3
6	135,419,631	rs9389268	0.271	6	135,419,018	rs9399137	*HBS1L-MYB*	1.000	613	EB	135,419,430	135,419,712	0.0593	1
6	135,419,636	rs9376091	0.271	6	135,419,018	rs9399137	*HBS1L-MYB*	1.000	618	EB	135,419,430	135,419,712	0.0593	1
6	135,419,688	rs9402685	0.271	6	135,419,018	rs9399137	*HBS1L-MYB*	1.000	670	EB	135,419,430	135,419,712	0.0593	1
7	123,410,525	rs112385417	0.097	7	123,411,223	rs4731120	*WASL*	1.000	698	MO	123,410,517	123,410,916	0.0386	4
9	4,811,553	rs13284787	0.231	9	4,814,948	rs13300663	*RCL1*	0.993	3,395	EB	4,811,442	4,811,611	0.0458	1
10	65,016,174	rs7088799	0.410	10	65,050,659	rs7075195	*JMJD1C*	1.000	34,485	MK	65,016,137	65,016,378	0.0437	1
10	65,027,143	rs7098181	0.407	10	65,050,659	rs7075195	*JMJD1C*	0.989	23,516	MK	65,027,095	65,027,550	0.0489	1
11	268,927	rs72882960	0.251	11	270,715	rs17655730	*PSMD13-NLRP6*	0.979	1,788	MO	268,859	269,026	0.0195	1
11	268,940	rs17655663	0.251	11	270,715	rs17655730	*PSMD13-NLRP6*	0.979	1,775	MO	268,859	269,026	0.0195	1
12	29,435,480	rs2015599	0.437	12	29,435,480	rs2015599	*MLSTD1*	1.000	0	MK	29,435,458	29,435,823	0.0842	2
12	29,435,675	rs1006409	0.437	12	29,435,480	rs2015599	*MLSTD1*	1.000	195	MK	29,435,458	29,435,823	0.0842	2
12	57,030,686	rs1107479	0.368	12	57,055,291	rs2950390	*PTGES3-BAZ2A*	0.933	24,605	EB	57,030,512	57,031,051	0.2790	4
12	57,030,686	rs1107479	0.368	12	57,055,291	rs2950390	*PTGES3-BAZ2A*	0.933	24,605	MK	57,030,568	57,031,065	0.1384	4
12	57,030,686	rs1107479	0.368	12	57,055,291	rs2950390	*PTGES3-BAZ2A*	0.933	24,605	MO	57,030,608	57,030,965	0.0204	1
12	57,119,236	rs3214051	0.611	12	57,055,291	rs2950390	*PTGES3-BAZ2A*	0.842	63,945	EB	57,118,946	57,119,448	0.2389	4
12	57,119,236	rs3214051	0.611	12	57,055,291	rs2950390	*PTGES3-BAZ2A*	0.842	63,945	MK	57,119,042	57,119,471	0.1508	4
12	57,119,236	rs3214051	0.611	12	57,055,291	rs2950390	*PTGES3-BAZ2A*	0.842	63,945	MO	57,119,035	57,119,458	0.0293	3
13	95,895,666	rs4148450	0.907	13	95,898,207	rs4148441	*ABCC4*	1.000	2,541	MK	95,895,547	95,895,837	0.0711	2
14	68,394,054	rs12431697	0.201	14	68,520,906	rs8022206	*RAD51L1*	0.841	126,852	EB	68,393,931	68,394,322	0.0743	2
14	68,394,298	rs7142860	0.192	14	68,520,906	rs8022206	*RAD51L1*	0.891	126,608	EB	68,393,931	68,394,322	0.0743	2
14	68,454,505	rs17192586	0.177	14	68,520,906	rs8022206	*RAD51L1*	0.982	66,401	EB	68,454,306	68,454,658	0.0626	1
14	68,454,505	rs17192586	0.177	14	68,520,906	rs8022206	*RAD51L1*	0.982	66,401	MK	68,454,266	68,454,734	0.1252	3
14	103,031,831	rs11160693	0.742	14	103,040,087	rs11628318	*RCOR1*	1.000	8,256	MO	103,031,713	103,032,081	0.0287	3
14	103,031,845	rs7155171	0.774	14	103,040,087	rs11628318	*RCOR1*	0.841	8,242	MO	103,031,713	103,032,081	0.0287	3
14	105,721,603	rs2735816	0.298	14	105,729,792	rs3000073	*BRF1*	0.889	8,189	MK	105,721,485	105,721,688	0.0507	1
15	65,165,459	rs4776639	0.157	15	65,183,801	rs1719271	*ANKDD1A*	0.970	18,342	MO	65,165,319	65,165,589	0.0257	2
15	65,187,874	rs1522744	0.179	15	65,183,801	rs1719271	*ANKDD1A*	0.833	4,073	MO	65,187,724	65,187,887	0.0185	1
17	27,661,844	rs117772072	0.426	17	27,714,587	rs8076739	*TAOK1*	0.859	52,743	EB	27,661,572	27,662,487	0.1082	2

17	27,661,844	rs117772072	0.426	17	27,714,587	rs8076739	*TAOK1*	0.859	52,743	MK	27,661,749	27,661,943	0.0395	1
17	27,662,271	n/a	0.439	17	27,714,587	rs8076739	*TAOK1*	0.907	52,316	EB	27,661,572	27,662,487	0.1082	2
17	27,662,271	n/a	0.439	17	27,714,587	rs8076739	*TAOK1*	0.907	52,316	MK	27,662,178	27,662,322	0.0380	1
20	1,924,707	rs13042885	0.275	20	1,924,707	rs13042885	*SIRPA*	1.000	0	MO	1,924,320	1,924,729	0.0368	4
20	57,589,995	rs55905547	0.194	20	57,587,771	rs4812048	*CTSZ-TUBB1*	0.925	2,224	MK	57,589,647	57,590,002	0.0878	3

(B) Erythrocyte phenotypes, stringent peak calling														
SNP in open chromatin				GWA sentinel SNP				Proxy↔sentinel		FAIRE peak				
Chr	*Position*	*rs number*	*Allele freq*	*Chr*	*Position*	*rs number*	*Gene locus*	r^2	*Distance (bp)*	*Cell type*	*Start position*	*End position*	*F-Seq score*	*Bin*
1	3,691,528	rs1175550	0.206	1	3,691,528	rs1175550	*CCDC27-LRRC48*	1.000	0	EB	3,691,366	3,691,660	0.0608	1
1	47,679,258	rs4926524	0.561	1	47,676,233	rs741959	*TAL1*	0.850	3,025	EB	47,679,193	47,679,296	0.0414	1
1	158,596,438	rs2482963	0.278	1	158,575,729	rs857684	*OR10Z1-SPTA1*	0.941	20,709	MO	158,596,370	158,596,475	0.0177	1
1	199,010,721	rs1434282	0.724	1	199,007,208	rs7529925	*MIR181A1*	0.929	3,513	EB	199,010,573	199,011,066	0.0964	2
1	199,010,721	rs1434282	0.724	1	199,007,208	rs7529925	*MIR181A1*	0.929	3,513	MK	199,010,612	199,011,128	0.1143	3
2	111,843,166	rs2880112	0.435	2	111,849,659	rs10207392	*ACOXL*	0.842	6,493	MO	111,843,101	111,843,381	0.0245	2
3	141,217,954	rs6808837	0.384	3	141,266,493	rs6776003	*RASA2*	0.857	48,539	EB	141,217,874	141,218,028	0.0453	1
3	141,217,954	rs6808837	0.384	3	141,266,493	rs6776003	*RASA2*	0.857	48,539	MO	141,217,713	141,218,095	0.0337	3
3	142,233,990	rs6791816	0.589	3	142,120,786	rs13061823	*XRN1*	0.824	113,204	MK	142,233,859	142,234,023	0.0392	1
4	55,408,875	rs218264	0.258	4	55,395,024	rs218238	*KIT*	0.834	13,851	EB	55,408,759	55,409,035	0.0724	2
4	122,745,038	rs769236	0.368	4	122,751,061	rs13152701	*BBS7-CCNA2*	1.000	6,023	EB	122,744,961	122,745,400	0.1267	3
4	122,750,079	rs13145213	0.369	4	122,751,061	rs13152701	*BBS7-CCNA2*	0.994	982	MK	122,750,001	122,750,254	0.0535	2
4	122,791,601	rs2271176	0.368	4	122,751,061	rs13152701	*BBS7-CCNA2*	1.000	40,540	EB	122,791,573	122,791,906	0.1025	2
6	41,925,159	rs9349205	0.234	6	41,914,378	rs9349204	*CCND3*	0.850	10,781	EB	41,924,850	41,925,202	0.0909	2
6	109,625,879	rs1546723	0.422	6	109,626,965	rs1008084	*CCDC162*	0.989	1,086	EB	109,625,663	109,626,087	0.1237	3
6	109,625,879	rs1546723	0.422	6	109,626,965	rs1008084	*CCDC162*	0.989	1,086	MK	109,625,692	109,625,981	0.0581	2
6	135,419,631	rs9389268	0.271	6	135,427,159	rs9389269	*HBS1L*	0.911	7,528	EB	135,419,430	135,419,712	0.0593	1
6	135,419,636	rs9376091	0.271	6	135,427,159	rs9389269	*HBS1L*	0.911	7,523	EB	135,419,430	135,419,712	0.0593	1
6	135,419,688	rs9402685	0.271	6	135,427,159	rs9389269	*HBS1L*	0.911	7,471	EB	135,419,430	135,419,712	0.0593	1
6	135,431,318	rs6920211	0.258	6	135,427,159	rs9389269	*HBS1L*	0.850	4,159	EB	135,431,304	135,431,674	0.0942	2
6	135,431,640	rs9494142	0.255	6	135,427,159	rs9389269	*HBS1L*	0.863	4,481	EB	135,431,304	135,431,674	0.0942	2

6	139,839,960	rs589235	0.499	6	139,844,429	rs590856	*CITED2*	0.904	4,469	EB	139,839,765	139,840,281	0.1883	3
6	164,463,355	rs4709819	0.446	6	164,482,836	rs736661	*QKI*	1.000	19,481	EB	164,463,287	164,463,672	0.1176	2
6	164,463,355	rs4709819	0.446	6	164,482,836	rs736661	*QKI*	1.000	19,481	MK	164,463,305	164,463,683	0.0787	2
6	164,463,572	rs4709820	0.446	6	164,482,836	rs736661	*QKI*	1.000	19,264	EB	164,463,287	164,463,672	0.1176	2
6	164,463,572	rs4709820	0.446	6	164,482,836	rs736661	*QKI*	1.000	19,264	MK	164,463,305	164,463,683	0.0787	2
8	41,630,405	rs4737009	0.261	8	41,630,405	rs4737009	*ANK1*	1.000	0	EB	41,630,153	41,630,603	0.1498	3
8	41,630,405	rs4737009	0.261	8	41,630,405	rs4737009	*ANK1*	1.000	0	MK	41,630,356	41,630,449	0.0357	1
8	41,630,447	rs4737010	0.250	8	41,630,405	rs4737009	*ANK1*	0.946	42	EB	41,630,153	41,630,603	0.1498	3
8	41,630,447	rs4737010	0.250	8	41,630,405	rs4737009	*ANK1*	0.946	42	MK	41,630,356	41,630,449	0.0357	1
9	4,852,599	rs10758656	0.193	9	4,844,265	rs2236496	*RCL1*	0.950	8,334	EB	4,852,346	4,852,777	0.1514	3
10	45,966,422	rs901683	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	0	EB	45,966,011	45,966,515	0.1390	3
10	46,039,930	rs75595592	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	73,508	EB	46,039,726	46,040,064	0.0550	1
10	46,053,061	rs9422657	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	86,639	MK	46,052,986	46,053,213	0.0528	1
11	9,023,421	rs7479407	0.413	11	8,938,049	rs11042125	*AKIP1-C11orf16-C11orf17-NRIP3-ST5*	0.873	85,372	MO	9,023,307	9,023,742	0.0429	4
11	73,115,314	rs7114009	0.114	11	73,009,084	rs7125949	*ARHGEF17-P2RY6*	0.827	106,230	MO	73,115,131	73,115,357	0.0213	1
14	65,499,909	rs12435835	0.481	14	65,502,239	rs7155454	*FNTB-MAX*	0.989	2,330	MK	65,499,871	65,500,238	0.1006	3
14	65,509,878	rs11628273	0.481	14	65,502,239	rs7155454	*FNTB-MAX*	0.989	7,639	EB	65,509,783	65,510,185	0.1592	3
14	65,509,878	rs11628273	0.481	14	65,502,239	rs7155454	*FNTB-MAX*	0.989	7,639	MK	65,509,766	65,510,128	0.0857	3
15	66,070,693	rs2572207	0.209	15	66,070,693	rs2572207	*DENND4A-PTPLAD1*	1.000	0	EB	66,070,410	66,070,913	0.1031	2
15	75,315,778	rs2304903	0.225	15	75,321,262	rs8028632	*PPCDC-SCAMP5*	1.000	5,484	EB	75,315,650	75,316,026	0.1255	3
15	75,315,778	rs2304903	0.225	15	75,321,262	rs8028632	*PPCDC-SCAMP5*	1.000	5,484	MK	75,315,722	75,315,936	0.0476	1
15	75,322,179	rs35911108	0.225	15	75,321,262	rs8028632	*PPCDC-SCAMP5*	1.000	917	MK	75,322,112	75,322,277	0.0384	1
15	75,354,621	rs35577967	0.212	15	75,321,262	rs8028632	*PPCDC-SCAMP5*	0.881	33,359	MO	75,354,500	75,354,655	0.0201	1
16	163,598	rs11248850	0.487	16	163,598	rs11248850	*C16orf35-HBA1*	1.000	0	MK	163,466	163,821	0.0790	2
16	163,667	rs11865131	0.487	16	163,598	rs11248850	*C16orf35-HBA1*	1.000	69	MK	163,466	163,821	0.0790	2
16	170,044	rs11866877	0.460	16	163,598	rs11248850	*C16orf35-HBA1*	0.850	6,446	EB	169,902	170,268	0.0891	2
16	67,927,124	rs7196789	0.172	16	67,902,326	rs2271294	*CTRL-EDC4-NUTF2*	0.991	24,798	EB	67,926,933	67,927,181	0.0645	1
16	67,927,124	rs7196789	0.172	16	67,902,326	rs2271294	*CTRL-EDC4-NUTF2*	0.991	24,798	MK	67,926,943	67,927,172	0.0514	1
16	88,840,462	rs10445033	0.634	16	88,840,462	rs10445033	*FAM38A*	1.000	0	EB	88,840,369	88,840,702	0.0689	2
16	88,840,462	rs10445033	0.634	16	88,840,462	rs10445033	*FAM38A*	1.000	0	MK	88,840,437	88,840,569	0.0389	1

17	42,323,376	rs7209801	0.272	17	42,294,337	rs2269906	*SLC4A1-UBTF*	0.804	29,039	EB	42,323,033	42,323,732	0.0931	2
17	44,217,112	rs2532314	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	390,475	MK	44,217,038	44,217,183	0.0362	1
17	44,253,364	rs2532259	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	426,727	EB	44,253,293	44,253,384	0.0403	1
17	44,271,430	rs2532236	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	444,793	EB	44,271,327	44,271,746	0.1180	2
17	44,271,430	rs2532236	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	444,793	MK	44,271,426	44,271,848	0.0510	1
17	44,272,000	rs2532235	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	445,363	MO	44,271,954	44,272,323	0.0267	2
17	44,272,266	rs2532234	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	445,629	MO	44,271,954	44,272,323	0.0267	2
17	44,272,552	rs17663792	0.231	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.993	445,915	MK	44,272,445	44,272,713	0.0509	1
17	44,272,552	rs17663792	0.231	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.993	445,915	MO	44,272,408	44,272,673	0.0216	1
17	44,272,679	rs2732660	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	446,042	MK	44,272,445	44,272,713	0.0509	1
17	44,276,330	rs1918785	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	449,693	MK	44,276,225	44,276,411	0.0454	1
17	44,280,188	rs2732675	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	453,551	MK	44,280,017	44,280,226	0.0443	1
17	44,289,101	rs2732629	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.985	462,464	MO	44,289,093	44,289,164	0.0171	1
17	44,289,150	rs2732630	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.985	462,513	MO	44,289,093	44,289,164	0.0171	1
18	43,802,778	rs12607898	0.733	18	43,833,278	rs4890633	*C18orf25*	0.980	30,500	EB	43,802,643	43,803,150	0.1602	3
19	4,458,063	rs11670503	0.254	19	4,366,219	rs732716	*MPND-SH3GL1-UBXD1*	0.839	91,844	EB	4,457,808	4,458,132	0.1203	2
19	13,001,547	rs11085824	0.310	19	13,024,250	rs741702	*CALR-FARSA-SYCE2*	0.840	22,703	MO	13,001,382	13,002,002	0.0332	3

19	13,030,280	rs8113575	0.701	19	13,024,250	rs741702	*CALR-FARSA-SYCE2*	0.921	6,030	EB	13,030,047	13,030,376	0.1119	2
19	13,044,544	rs2974750	0.700	19	13,024,250	rs741702	*CALR-FARSA-SYCE2*	0.903	20,294	MO	13,044,538	13,044,646	0.0180	1
22	32,887,498	rs6518786	0.386	22	32,880,585	rs5749446	*FBXO7*	1.000	6,913	EB	32,887,371	32,887,591	0.0472	1
22	32,887,566	rs5754113	0.386	22	32,880,585	rs5749446	*FBXO7*	1.000	6,981	EB	32,887,371	32,887,591	0.0472	1

(C) Platelet phenotypes, moderate peak calling														
SNP in open chromatin				GWA sentinel SNP				Proxy↔sentinel		FAIRE peak				
Chr	*Position*	*rs number*	*Allele freq*	*Chr*	*Position*	*rs number*	*Gene locus*	*r^2*	*Distance (bp)*	*Cell type*	*Start position*	*End position*	*F-Seq score*	*Bin*
1	205,236,432	rs1768587	0.361	1	205,244,953	rs1172130	*TMCC2*	0.934	8,521	MK	205,236,432	205,236,777	0.0895	4
1	205,253,839	rs3851296	0.422	1	205,244,953	rs1172130	*TMCC2*	0.829	8,886	MO	205,253,816	205,254,326	0.0303	3
2	43,665,943	rs76684055	0.079	2	43,687,879	rs17030845	*THADA*	0.916	21,936	EB	43,665,917	43,666,222	0.0553	2
2	43,668,169	rs7567685	0.086	2	43,687,879	rs17030845	*THADA*	1.000	19,710	EB	43,668,076	43,668,192	0.0325	1
2	43,668,176	rs7581586	0.086	2	43,687,879	rs17030845	*THADA*	1.000	19,703	EB	43,668,076	43,668,192	0.0325	1
2	43,676,433	rs12105786	0.086	2	43,687,879	rs17030845	*THADA*	1.000	11,446	EB	43,676,386	43,676,580	0.0339	1
2	43,678,617	rs17406646	0.079	2	43,687,879	rs17030845	*THADA*	0.916	9,262	MO	43,678,533	43,678,675	0.0153	1
2	43,683,885	rs6728106	0.093	2	43,687,879	rs17030845	*THADA*	0.922	3,994	MO	43,683,870	43,683,943	0.0147	1
2	43,710,444	rs7600657	0.097	2	43,687,879	rs17030845	*THADA*	0.880	22,565	MO	43,710,150	43,710,524	0.0167	1
2	43,737,465	rs10180005	0.087	2	43,687,879	rs17030845	*THADA*	0.983	49,586	MK	43,737,276	43,737,481	0.0323	1
4	6,891,519	rs11734132	0.161	4	6,891,519	rs11734132	*KIAA0232*	1.000	0	MK	6,891,275	6,891,535	0.0457	2
6	25,536,937	rs214059	0.447	6	25,548,288	rs441460	*LRRC16*	0.984	11,351	MK	25,536,902	25,537,516	0.1216	4
7	106,355,943	rs342272	0.425	7	106,372,219	rs342293	*FLJ36031-PIK3CG*	0.913	16,276	MO	106,355,923	106,356,436	0.0170	1
7	106,356,005	rs342273	0.425	7	106,372,219	rs342293	*FLJ36031-PIK3CG*	0.913	16,214	MO	106,355,923	106,356,436	0.0170	1
7	106,372,219	rs342293	0.447	7	106,372,219	rs342293	*FLJ36031-PIK3CG*	1.000	0	MK	106,372,203	106,372,378	0.0339	1
7	123,407,146	rs12671376	0.097	7	123,411,223	rs4731120	*WASL*	1.000	4,077	MK	123,407,083	123,407,197	0.0280	1
7	123,410,918	rs79415660	0.097	7	123,411,223	rs4731120	*WASL*	1.000	305	MO	123,410,502	123,410,962	0.0386	4
7	123,422,357	rs4727976	0.098	7	123,411,223	rs4731120	*WASL*	0.985	11,134	MO	123,422,338	123,422,610	0.0173	1
7	123,424,583	rs725859	0.097	7	123,411,223	rs4731120	*WASL*	1.000	13,360	MO	123,424,389	123,424,594	0.0155	1
8	106,581,528	rs6993770	0.295	8	106,581,528	rs6993770	*ZFPM2*	1.000	0	EB	106,581,358	106,581,529	0.0371	1
8	106,581,528	rs6993770	0.295	8	106,581,528	rs6993770	*ZFPM2*	1.000	0	MK	106,581,443	106,581,722	0.0359	1
9	332,152	rs1536609	0.627	9	331,490	rs10813766	*DOCK8*	0.983	662	MK	332,133	332,338	0.0362	1

10	65,027,610	rs10761731	0.407	10	65,050,659	rs7075195	*JMJD1C*	0.989	23,049	EB	65,027,412	65,027,689	0.0406	1
10	65,062,820	rs7073753	0.409	10	65,050,659	rs7075195	*JMJD1C*	0.995	12,161	EB	65,062,634	65,062,851	0.0440	1
10	65,071,215	rs10822155	0.407	10	65,050,659	rs7075195	*JMJD1C*	0.989	20,556	EB	65,071,215	65,071,462	0.0482	1
12	122,365,583	rs7961894	0.086	12	122,365,583	rs7961894	*WDR66*	1.000	0	MO	122,365,425	122,365,646	0.0172	1
14	68,413,769	rs7151917	0.177	14	68,520,906	rs8022206	*RAD51L1*	0.982	107,137	EB	68,413,594	68,413,802	0.0369	1
14	68,431,295	rs4902540	0.192	14	68,520,906	rs8022206	*RAD51L1*	0.891	89,611	MO	68,431,223	68,431,394	0.0153	1
14	68,599,145	rs2588834	0.823	14	68,520,906	rs8022206	*RAD51L1*	0.809	78,239	MO	68,599,107	68,599,284	0.0161	1
15	65,185,884	rs1684032	0.173	15	65,183,801	rs1719271	*ANKDD1A*	0.882	2,083	MO	65,185,876	65,186,279	0.0265	3
15	65,185,904	rs1719262	0.173	15	65,183,801	rs1719271	*ANKDD1A*	0.882	2,103	MO	65,185,876	65,186,279	0.0265	3
19	16,197,320	rs2228367	0.015	19	16,185,559	rs8109288	*TPM4*	1.000	11,761	MO	16,197,302	16,197,581	0.0187	2

(D) Erythrocyte phenotypes, moderate peak calling														
SNP in open chromatin				GWA sentinel SNP				Proxy↔sentinel		FAIRE peak				
Chr	*Position*	*rs number*	*Allele freq*	*Chr*	*Position*	*rs number*	*Gene locus*	r^2	*Distance (bp)*	*Cell type*	*Start position*	*End position*	*F-Seq score*	*Bin*
1	203,651,824	rs11240734	0.890	1	203,655,121	rs7551442	*ATP2B4*	1.000	3,297	EB	203,651,789	203,651,903	0.0322	1
1	205,253,839	rs3851296	0.422	1	205,249,450	rs9660992	*TMCC2*	1.000	4,389	MO	205,253,816	205,254,326	0.0303	3
3	142,004,728	rs4434160	0.552	3	142,120,786	rs13061823	*XRN1*	0.984	116,058	EB	142,004,626	142,004,732	0.0327	1
3	142,247,052	rs6440087	0.589	3	142,120,786	rs13061823	*XRN1*	0.824	126,266	EB	142,247,013	142,247,180	0.0339	1
3	142,272,314	rs9869842	0.585	3	142,120,786	rs13061823	*XRN1*	0.818	151,528	MO	142,272,275	142,272,335	0.0147	1
3	142,313,987	rs6440092	0.585	3	142,120,786	rs13061823	*XRN1*	0.808	193,201	EB	142,313,850	142,314,018	0.0326	1
3	142,315,074	rs13069307	0.561	3	142,120,786	rs13061823	*XRN1*	0.886	194,288	EB	142,315,064	142,315,310	0.0468	1
4	122,748,996	rs1048433	0.368	4	122,751,061	rs13152701	*BBS7-CCNA2*	1.000	2,065	MK	122,748,903	122,749,066	0.0335	1
6	43,810,974	rs9369425	0.708	6	43,811,430	rs9369427	*VEGFA*	0.994	456	EB	43,810,962	43,811,323	0.0606	2
7	50,427,982	rs6592965	0.452	7	50,428,445	rs12718598	*IKZF1*	0.848	463	EB	50,427,715	50,428,000	0.0573	2
7	100,221,849	rs4729597	0.626	7	100,240,296	rs2075672	*ACTL6B-TFR2*	0.925	18,447	EB	100,221,848	100,221,924	0.0318	1
10	45,972,325	rs71494788	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	5,903	MK	45,972,249	45,972,350	0.0278	1
10	45,991,978	rs12764652	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	25,556	EB	45,991,905	45,992,079	0.0347	1
10	46,013,438	rs12781186	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	47,016	MO	46,013,384	46,013,477	0.0148	1
10	46,023,327	rs12764693	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	56,905	MO	46,023,200	46,023,549	0.0164	1
10	46,024,335	rs35902429	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	57,913	EB	46,024,290	46,024,461	0.0368	1

10	46,051,401	rs9422654	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	84,979	MO	46,051,373	46,051,506	0.0153	1
10	46,051,408	rs9422655	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	84,986	MO	46,051,373	46,051,506	0.0153	1
10	46,074,329	rs71496620	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	107,907	EB	46,073,754	46,074,355	0.0461	1
10	46,081,096	rs12779637	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	114,674	MK	46,081,095	46,081,396	0.0339	1
10	46,081,101	rs12772102	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	114,679	MK	46,081,095	46,081,396	0.0339	1
10	46,081,358	rs35183751	0.079	10	45,966,422	rs901683	*CTGLF1-MARCH8*	1.000	114,936	MK	46,081,095	46,081,396	0.0339	1
11	73,018,413	rs77127734	0.111	11	73,009,084	rs7125949	*ARHGEF17-P2RY6*	0.949	9,329	MO	73,018,380	73,018,518	0.0164	1
11	73,136,261	rs12288753	0.112	11	73,009,084	rs7125949	*ARHGEF17-P2RY6*	0.838	127,177	MO	73,136,189	73,136,489	0.0178	1
12	121,163,518	rs2239760	0.433	12	121,126,438	rs3829290	*MLEC*	0.838	37,080	EB	121,163,444	121,163,621	0.0377	1
14	103,822,762	rs17616316	0.065	14	103,822,762	rs17616316	*EIF5*	1.000	0	EB	103,822,621	103,822,799	0.0333	1
15	75,322,310	rs5020842	0.226	15	75,321,262	rs8028632	*PPCDC-SCAMP5*	0.992	1,048	MK	75,321,852	75,322,327	0.0384	1
15	75,370,232	rs8036197	0.226	15	75,321,262	rs8028632	*PPCDC-SCAMP5*	0.806	48,970	MO	75,370,203	75,370,361	0.0167	1
17	27,077,331	rs3181215	0.194	17	27,075,423	rs2070265	*C17orf63-ERAL1-NEK8-TRAF4*	0.992	1,908	MO	27,077,239	27,077,355	0.0156	1
17	27,088,436	rs7215000	0.194	17	27,075,423	rs2070265	*C17orf63-ERAL1-NEK8-TRAF4*	0.992	13,013	MO	27,088,426	27,088,562	0.0156	1
17	44,047,216	rs62641967	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	220,579	MO	44,047,179	44,047,573	0.0176	1
17	44,080,039	rs62064663	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	253,402	MO	44,080,033	44,080,154	0.0159	1
17	44,147,721	rs56323408	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	321,084	MK	44,147,710	44,148,096	0.0294	1
17	44,170,238	rs112596352	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	343,601	MO	44,170,238	44,170,405	0.0166	1
17	44,177,337	rs62061808	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	350,700	MO	44,177,282	44,177,566	0.0156	1
17	44,178,839	rs62061809	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	352,202	EB	44,178,689	44,178,883	0.0396	1
17	44,184,375	rs79346219	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	357,738	MO	44,184,339	44,184,430	0.0147	1
17	44,184,404	rs73984689	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	357,767	MO	44,184,339	44,184,430	0.0147	1

17	44,184,428	rs111295615	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	357,791	MO	44,184,339	44,184,430	0.0147	1
17	44,188,477	rs17577159	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	361,840	MO	44,188,308	44,188,489	0.0161	1
17	44,195,424	rs62061852	0.231	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.993	368,787	MO	44,195,244	44,195,458	0.0156	1
17	44,214,814	rs740711	0.218	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.934	388,177	EB	44,214,746	44,214,863	0.0340	1
17	44,214,815	rs740710	0.218	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.934	388,178	EB	44,214,746	44,214,863	0.0340	1
17	44,217,226	rs2696599	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	390,589	MK	44,216,973	44,217,266	0.0362	1
17	44,218,044	rs1918801	0.202	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.849	391,407	MK	44,217,957	44,218,117	0.0287	1
17	44,218,138	rs1918800	0.200	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.835	391,501	MO	44,218,100	44,218,510	0.0158	1
17	44,218,242	rs1918799	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	391,605	MO	44,218,100	44,218,510	0.0158	1
17	44,221,836	rs2696589	0.202	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.849	395,199	MK	44,221,735	44,221,885	0.0303	1
17	44,228,169	rs1918793	0.237	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.964	401,532	MK	44,228,125	44,228,177	0.0270	1
17	44,241,664	rs2532286	0.231	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.993	415,027	EB	44,241,495	44,241,680	0.0326	1
17	44,244,397	rs2696684	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	417,760	MK	44,244,122	44,244,665	0.0347	1
17	44,244,581	rs17585644	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	417,944	MK	44,244,122	44,244,665	0.0347	1
17	44,244,896	rs2532282	0.202	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.849	418,259	MK	44,244,809	44,245,065	0.0329	1
17	44,244,926	rs2696657	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-*	1.000	418,289	MK	44,244,809	44,245,065	0.0329	1

							IMP5-KIAA1267-MAPT-STH							
17	44,246,527	rs2532277	0.202	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.849	419,890	MO	44,246,415	44,246,533	0.0150	1
17	44,248,042	rs2532271	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	421,405	MO	44,247,806	44,248,048	0.0169	1
17	44,253,275	rs2532260	0.197	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.821	426,638	EB	44,253,215	44,253,442	0.0403	1
17	44,253,275	rs2532260	0.197	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.821	426,638	MK	44,253,213	44,253,362	0.0295	1
17	44,254,291	rs2732645	0.197	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.821	427,654	MO	44,254,143	44,254,400	0.0167	1
17	44,254,379	rs2732646	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	427,742	MO	44,254,143	44,254,400	0.0167	1
17	44,256,655	rs740708	0.194	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.807	430,018	MO	44,256,325	44,257,117	0.0320	4
17	44,258,354	rs740706	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	431,717	EB	44,258,268	44,258,455	0.0354	1
17	44,258,422	rs758523	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	431,785	EB	44,258,268	44,258,455	0.0354	1
17	44,265,477	rs2696709	0.202	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.849	438,840	EB	44,265,394	44,265,607	0.0374	1
17	44,267,617	rs1918788	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	440,980	EB	44,267,586	44,267,789	0.0370	1
17	44,268,488	rs2141298	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	441,851	EB	44,268,377	44,268,742	0.0392	1
17	44,274,560	rs2696440	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	1.000	447,923	MO	44,274,546	44,274,672	0.0157	1
17	44,289,220	rs2532417	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.985	462,583	MO	44,288,999	44,289,261	0.0171	1
17	44,289,232	rs2732631	0.230	17	43,826,637	rs12150672	*ARHGAP27-ARL17-C17orf69-CRHR1-IMP5-KIAA1267-MAPT-STH*	0.985	462,595	MO	44,288,999	44,289,261	0.0171	1

22	32,870,769	rs2072814	0.366	22	32,880,585	rs5749446	*FBXO7*	0.908	9,816	EB	32,870,586	32,870,772	0.0389	1
22	32,884,381	rs7354804	0.386	22	32,880,585	rs5749446	*FBXO7*	1.000	3,796	EB	32,884,051	32,884,413	0.0464	1
22	32,887,661	rs17772071	0.386	22	32,880,585	rs5749446	*FBXO7*	1.000	7,076	EB	32,887,327	32,887,681	0.0472	1

Table 8-4. ***In silico*** **transcription factor binding site predictions.** I report transcription factors of whose DNA binding motif may be altered due to the presence of a SNP. The table is ranked from the highest to the lowest absolute difference log(*P*) between the probes harbouring the reference and alternative allele of the platelet SNPs. Only results with absolute differences of log(*P*)>1 or log(*P*)<-1 are shown. The results with the probes containing rs1107479C>T did not meet these criteria.

#	Rank	log(*P*)	Ref allele	Alt allele	Matrix ID	Matrix name	Binding factor
1			rs1006409-A	rs1006409-G			
	1	5.03	1.84x10^{-06}	1.98x10^{-01}	M00252	V$TATA_01	TATA
	2	4.80	1.84x10^{-06}	1.15x10^{-01}	M00216	V$TATA_C	TATA
	3	3.88	1.84x10^{-06}	1.41x10^{-02}	M00215	V$SRF_C	SRF
	4	3.36	1.20x10^{-04}	2.71x10^{-01}	M01292	V$HOXA13_01	HOXA13
	5	2.45	2.91x10^{-05}	8.11x10^{-03}	M01304	V$SRF_03	SRF
	6	2.32	2.98x10^{-04}	6.23x10^{-02}	M00395	V$HOXA3_01	HOXA3
	7	1.75	1.19x10^{-05}	6.61x10^{-04}	M00922	V$SRF_Q5_01	SRF
	8	1.59	2.66x10^{-03}	1.04x10^{-01}	M01375	V$HOXD10_01	HOXD10
	9	1.41	1.30x10^{-02}	3.37x10^{-01}	M00671	V$TCF4_Q5	TCF4
	10	-1.28	2.36x10^{-01}	1.23x10^{-02}	M00751	V$AML1_Q6	AML1
	11	1.20	2.25x10^{-03}	3.55x10^{-02}	M00152	V$SRF_01	SRF
	12	1.15	1.06x10^{-03}	1.50x10^{-02}	M00810	V$SRF_Q4	SRF
	13	-1.07	1.36x10^{-01}	1.17x10^{-02}	M00722	V$COREBINDINGFACTOR_Q6	Core-binding factor
	14	-1.06	1.16x10^{-01}	1.01x10^{-02}	M00731	V$OSF2_Q6	OSF2
	15	-1.06	7.72x10^{-02}	6.70x10^{-03}	M00984	V$PEBP_Q6	PEBP
	16	-1.00	5.64x10^{-01}	5.64x10^{-02}	M01658	V$AML1_Q4	AML1
2			rs1107479-C	rs1107479-T			
	–	–	–	–	–	–	–
3			rs11731274-T	rs11731274-G			
	1	2.49	1.61x10^{-03}	4.94x10^{-01}	M00500	V$STAT6_02	STAT6
	2	1.50	1.43x10^{-02}	4.51x10^{-01}	M00496	V$STAT1_03	STAT1
	3	-1.29	3.12x10^{-02}	1.61x10^{-03}	M01177	V$SREBP2_Q6	SREBP2
	4	1.08	3.36x10^{-02}	4.02x10^{-01}	M01281	V$NFAT1_Q6	NFAT1
4			rs11734099-G	rs11734099-A			
	1	1.27	3.14x10^{-02}	5.92x10^{-01}	M00272	V$P53_02	P53
5			rs17192586-G	rs17192586-A			
	1	-2.44	2.08x10^{-01}	7.63x10^{-04}	M00175	V$AP4_Q5	AP4
	2	-2.22	1.76x10^{-01}	1.06x10^{-03}	M01287	V$NEUROD_01	NEUROD
	3	1.85	2.74x10^{-03}	1.92x10^{-01}	M00227	V$VMYB_02	VMYB
	4	1.69	4.86x10^{-03}	2.38x10^{-01}	M00003	V$VMYB_01	VMYB
	5	-1.58	3.97x10^{-01}	1.04x10^{-02}	M00927	V$AP4_Q6_01	AP4
	6	-1.56	2.52x10^{-01}	6.95x10^{-03}	M01347	V$RHOX11_01	RHOX11
	7	-1.43	3.15x10^{-01}	1.17x10^{-02}	M01384	V$RHOX11_02	RHOX11
	8	-1.37	1.99x10^{-02}	8.44x10^{-04}	M01419	V$MEIS1_02	MEIS1
	9	-1.32	1.60x10^{-02}	7.63x10^{-04}	M01488	V$MEIS2_01	MEIS2

	10	-1.22	5.43x10^{-01}	3.27x10^{-02}	M00973	V$E2A_Q6	E2A
	11	-1.13	2.02x10^{-02}	1.49x10^{-03}	M01459	V$PREP1_01	PREP1
	12	-1.12	4.12x10^{-01}	3.10x10^{-02}	M00712	V$MYOGENIN_Q6	MYOGENIN
	13	-1.08	2.23x10^{-01}	1.88x10^{-02}	M00176	V$AP4_Q6	AP4
	14	-1.06	2.47x10^{-01}	2.15x10^{-02}	M00122	V$USF_02	USF
	15	-1.03	2.77x10^{-02}	2.61x10^{-03}	M01407	V$TGIF2_01	TGIF2
6			rs2015599-G	rs2015599-A			
	1	-1.40	8.62x10^{-02}	3.45x10^{-03}	M01410	V$IRX4_01	IRX4
	2	-1.39	6.08x10^{-02}	2.51x10^{-03}	M01485	V$IRX3_02	IRX3
	3	-1.17	9.20x10^{-02}	6.29x10^{-03}	M01318	V$IRX3_01	IRX3
7			rs2038479-C	rs2038479-A			
	1	-2.63	2.51x10^{-01}	5.86x10^{-04}	M00630	V$FOXM1_01	FOXM1
	2	-2.55	8.43x10^{-02}	2.35x10^{-04}	M00216	V$TATA_C	TATA
	3	-2.23	1.61x10^{-01}	9.50x10^{-04}	M00252	V$TATA_01	TATA
	4	-1.78	1.43x10^{-02}	2.35x10^{-04}	M00980	V$TBP_Q6	TBP
	5	1.53	1.06x10^{-02}	3.59x10^{-01}	M00462	V$GATA6_01	GATA6
	6	1.50	9.91x10^{-03}	3.16x10^{-01}	M00104	V$CDPCR1_01	CDPCR1
	7	1.34	9.23x10^{-03}	2.00x10^{-01}	M00347	V$GATA1_06	GATA1
	8	-1.19	1.43x10^{-01}	9.29x10^{-03}	M01404	V$HOXD13_01	HOXD13
	9	1.06	2.53x10^{-02}	2.90x10^{-01}	M00348	V$GATA2_02	GATA2
	10	1.03	1.38x10^{-02}	1.49x10^{-01}	M01429	V$HOMEZ_01	HOMEZ
8			rs2038480-A	rs2038480-T			
	1	-3.24	9.25x10^{-02}	5.38x10^{-05}	M00744	V$POU1F1_Q6	POU1F1
	2	-2.98	5.08x10^{-02}	5.38x10^{-05}	M00138	V$OCT1_04	OCT1
	3	2.38	5.38x10^{-05}	1.29x10^{-02}	M00451	V$NKX3A_01	NKX3A
	4	2.23	4.68x10^{-03}	7.94x10^{-01}	M01181	V$NKX32_01	NKX32
	5	-1.22	8.44x10^{-02}	5.10x10^{-03}	M00672	V$TEF_Q6	TEF
	6	-1.03	1.51x10^{-01}	1.42x10^{-02}	M00136	V$OCT1_02	OCT1
9			rs214060-C	rs214060-T			
	1	-2.38	5.24x10^{-01}	2.19x10^{-03}	M01275	V$IPF1_Q6	IPF1
	2	-1.53	7.49x10^{-02}	2.19x10^{-03}	M00436	V$IPF1_Q4	IPF1
	3	-1.30	3.47x10^{-01}	1.74x10^{-02}	M00930	V$OCT1_Q5_01	OCT1
	4	-1.25	6.51x10^{-01}	3.66x10^{-02}	M01307	V$POU5F1_01	POU5F1
	5	-1.20	3.90x10^{-01}	2.46x10^{-02}	M00795	V$OCT_Q6	OCT
	6	-1.01	3.94x10^{-01}	3.84x10^{-02}	M01338	V$HOXD3_01	HOXD3
10			rs2735816-G	rs2735816-C			
	1	-2.59	3.01x10^{-01}	7.78x10^{-04}	M00921	V$GR_Q6_01	GR
	2	-1.41	1.91x10^{-02}	7.48x10^{-04}	M00647	V$LXR_Q3	LXR
11			rs3214051-G	rs3214051-A			
	1	-1.39	2.28x10^{-01}	9.35x10^{-03}	M00097	V$PAX6_01	PAX6
	2	1.08	4.24x10^{-03}	5.15x10^{-02}	M00134	V$HNF4_01	HNF4
	3	1.08	6.96x10^{-02}	8.36x10^{-01}	M01107	V$RUSH1A_02	RUSH1A
	4	-1.01	5.61x10^{-01}	5.48x10^{-02}	M00684	V$XPF1_Q6	XPF1

12			rs3804749-C	rs3804749-T			
	1	-1.74	6.66x10^{-01}	1.22x10^{-02}	M00073	V$DELTAEF1_01	DELTAEF1
	2	-1.58	3.98x10^{-01}	1.04x10^{-02}	M00973	V$E2A_Q6	E2A
	3	-1.29	6.72x10^{-01}	3.42x10^{-02}	M01034	V$EBOX_Q6_01	EBOX
	4	-1.13	3.20x10^{-01}	2.37x10^{-02}	M00821	V$NRF2_Q4	NRF2
	5	-1.11	4.17x10^{-01}	3.25x10^{-02}	M00712	V$MYOGENIN_Q6	MYOGENIN
	6	-1.09	5.18x10^{-01}	4.20x10^{-02}	M01116	V$CLOCKBMAL_Q6	CLOCKBMAL
	7	-1.01	5.98x10^{-01}	5.81x10^{-02}	M00799	V$MYC_Q2	MYC
	8	-1.01	3.74x10^{-01}	3.68x10^{-02}	M01287	V$NEUROD_01	NEUROD
13			rs4148450-C	rs4148450-T			
	1	-1.62	4.14x10^{-01}	9.92x10^{-03}	M00762	V$DR1_Q3	DR1
	2	1.54	1.29x10^{-02}	4.47x10^{-01}	M00655	V$PEA3_Q6	PEA3
	3	1.45	3.36x10^{-02}	9.55x10^{-01}	M01079	V$CBF_01	CBF
	4	-1.42	3.54x10^{-01}	1.34x10^{-02}	M00763	V$PPAR_DR1_Q2	PPAR
	5	-1.30	4.04x10^{-01}	2.01x10^{-02}	M00973	V$E2A_Q6	E2A
	6	-1.22	2.28x10^{-01}	1.39x10^{-02}	M00693	V$E12_Q6	E12
	7	-1.21	2.29x10^{-01}	1.41x10^{-02}	M00764	V$HNF4_DR1_Q3	HNF4
	8	-1.07	4.54x10^{-02}	3.88x10^{-03}	M01263	V$TBX15_01	TBX15
	9	-1.04	3.79x10^{-02}	3.41x10^{-03}	M01195	V$TBX22_01	TBX22
	10	-1.04	1.90x10^{-01}	1.75x10^{-02}	M00765	V$COUP_DR1_Q6	COUP
14			rs55905547-A	rs55905547-G			
	1	1.16	6.61x10^{-02}	9.59x10^{-01}	M01033	V$HNF4_Q6_03	HNF4
15			rs6771416-G	rs6771416-A			
	1	-1.85	2.73x10^{-01}	3.87x10^{-03}	M01334	V$NKX11_01	NKX11
	2	-1.62	5.76x10^{-01}	1.38x10^{-02}	M00313	V$GEN_INI2_B	GEN
	3	-1.57	4.93x10^{-01}	1.34x10^{-02}	M00315	V$GEN_INI_B	GEN
	4	1.55	1.74x10^{-02}	6.12x10^{-01}	M01592	V$LBP9_01	LBP9
	5	-1.48	5.33x10^{-01}	1.78x10^{-02}	M00314	V$GEN_INI3_B	GEN
	6	1.30	4.01x10^{-02}	7.92x10^{-01}	M00077	V$GATA3_01	GATA3
	7	-1.19	2.49x10^{-01}	1.61x10^{-02}	M01386	V$EVX2_01	EVX2
	8	-1.07	2.50x10^{-01}	2.11x10^{-02}	M01331	V$ISX_01	ISX
	9	-1.03	6.84x10^{-01}	6.34x10^{-02}	M00624	V$DBP_Q6	DBP
	10	-1.01	3.64x10^{-01}	3.58x10^{-02}	M01382	V$GBX2_01	GBX2
	11	-1.00	3.36x10^{-01}	3.33x10^{-02}	M01427	V$NKX12_01	NKX12
	12	1.00	1.01x10^{-01}	9.99x10^{-01}	M00075	V$GATA1_01	GATA1
	13	-1.00	3.85x10^{-01}	3.89x10^{-02}	M01461	V$EMX2_01	EMX2
16			rs7618405-C	rs7618405-A			
	1	-2.96	4.70x10^{-01}	5.13x10^{-04}	M01131	V$SOX10_Q6	SOX10
	2	-1.83	8.17x10^{-01}	1.22x10^{-02}	M00137	V$OCT1_03	OCT1
	3	-1.30	3.81x10^{-01}	1.93x10^{-02}	M01016	V$SOX17_01	SOX17

Table 8-5. Investigation of the functional role of platelet volume-associated variants at chromosome 7q22.3. Proxy SNPs to rs342293 (r^2≥0.8) were retrieved from the 1000 Genomes Project (Pilot 1, CEU). Genomic coordinates were based on the human reference genome, build hg18 (NCBI build 36). *P*-values for association with mean platelet volume (MPV) were obtained from Soranzo, Spector, et al., 2009. *In silico* transcription binding site predictions were performed as described in **Section 2.10**. Based on the HaemAtlas data, we defined genes as expressed when they exhibit a normalised expression value of at least 8.5.

Proxy to rs342293							SNP overlaps with ...		
ID	*Chromosome*	*Position*	*r^2*	*Distance*	*Annotation*	*MPV P-value*	*NDR in MK cells*	*Transcription factor (TF) binding site (MatInspector)*	*Binding site of TF expressed in MK cells*
rs342207	7	106,108,641	0.81	50,814	Intergenic	1.46x10^{-10}	–	–	–
rs342209	7	106,109,208	0.84	50,247	Intergenic	n/a	–	–	–
rs342210	7	106,109,256	0.87	50,199	Intergenic	3.57x10^{-11}	–	–	–
rs342212	7	106,111,150	0.84	48,305	Intergenic	5.47x10^{-11}	–	–	–
rs342213	7	106,111,848	0.84	47,607	Intergenic	6.21x10^{-11}	–	–	–
rs342214	7	106,111,979	0.81	47,476	Intergenic	n/a	–	–	–
rs342236	7	106,122,874	0.84	36,581	Intergenic	1.24x10^{-10}	–	–	–
rs342239	7	106,124,138	0.90	35,317	Intergenic	6.97x10^{-11}	–	–	–
rs342240	7	106,124,486	0.90	34,969	Intergenic	6.97x10^{-11}	–	HMX2	–
rs342241	7	106,124,587	0.90	34,868	Intergenic	n/a	–	–	–
rs342242	7	106,126,225	0.90	33,230	Intergenic	6.43x10^{-11}	–	–	–
rs342244	7	106,128,061	0.81	31,394	Intergenic	1.43x10^{-10}	–	–	–
rs342247	7	106,130,427	0.84	29,028	Intergenic	4.23x10^{-11}	–	HHEX, HOXC4, LBX2, MSX	HHEX, MSX
rs342248	7	106,130,541	0.84	28,914	Intergenic	n/a	–	–	–
rs342251	7	106,132,045	0.90	27,410	Intergenic	n/a	–	–	–
rs342252	7	106,133,666	0.90	25,789	Intergenic	n/a	–	–	–
rs342254	7	106,135,492	0.90	23,963	Intergenic	3.84x10^{-11}	–	–	–
rs342257	7	106,137,089	0.84	22,366	Intergenic	6.09x10^{-11}	–	–	–
rs342271	7	106,143,111	0.90	16,344	Intergenic	n/a	–	–	–
rs342275	7	106,146,452	0.90	13,003	Intergenic	1.33x10^{-11}	–	–	–

rs342281	7	106,149,079	0.84	10,376	Intergenic	5.36x10^{-12}	–	–	–
rs342284	7	106,149,446	0.87	10,009	Intergenic	4.62x10^{-12}	–	–	–
rs342286	7	106,151,835	0.94	7,620	Intergenic	4.90x10^{-12}	–	–	–
rs342290	7	106,154,840	1.00	4,615	Intergenic	n/a	–	–	–
rs342292	7	106,157,880	1.00	1,575	Intergenic	8.64x10^{-13}	–	MEIS1, MEIS1A/HOXA9	MEIS1, MEIS1A/HOXA9
rs342293	7	106,159,455	1.00	0	Intergenic	6.75x10^{-13}	+	GATA1, EVI1	GATA1, EVI1
rs342294	7	106,159,858	1.00	403	Intergenic	n/a	+	–	–
rs342295	7	106,159,996	1.00	541	Intergenic	4.14x10^{-13}	–	–	–
rs342296	7	106,160,139	1.00	684	Intergenic	7.68x10^{-13}	–	–	–
rs342298	7	106,160,882	0.97	1,427	Intergenic	7.55x10^{-13}	–	–	–
rs342299	7	106,160,954	0.97	1,499	Intergenic	7.55x10^{-13}	–	–	–
rs386805	7	106,125,700	0.97	33,755	Intergenic	n/a	–	–	–
rs67036916	7	106,154,872	0.97	4,583	Intergenic	n/a	–	–	–
rs77655772	7	106,125,699	0.97	33,756	Intergenic	n/a	–	–	–

Table 8-6. Expression QTL associations at the *PIK3CG* gene locus in platelets, macrophages, monocytes, B cells (LCLs), adipose and skin. The strength of the relationship between alleles and gene expression intensities was estimated with the Spearman's rank correlation coefficient using the software Genevar (**Section 2.12**). In macrophages and monocytes (as part of the data generated by the Cardiogenics Consortium), the rs342293 proxy SNP rs342275 was used for the eQTL analysis. For LCLs, adipose and skin (as part of the data generated by the MuTHER Consortium), the proxy SNPs rs342296 and rs342275 were used for analysis. Values for r^2 were obtained from Phase II HapMap, CEU.

Cell type/tissue	SNP tested			eQTL nominal *P*-value	SNP with the strongest association with *PIK3CG* expression in 1-Mb window		
	ID	*r^2*	*Distance*		*ID*	*r^2 with rs342293*	*eQTL nominal P-value*
Platelets	rs342293	1.000	0 bp	0.0542	rs342293	1.000	–
Macrophages	rs342275	0.935	13,003 bp	0.0018	rs342275	0.935	–
Monocytes	rs342275	0.935	13,003 bp	0.4348	rs10953522	0.002	0.0005
LCLs	rs342296	1.000	684 bp	0.5983	rs7788626	0.032	0.0018
	rs342275	0.935	13,003 bp	0.9999	–	–	–
Adipose	rs342296	1.000	684 bp	0.5308	rs849375	0.059	0.0022
	rs342275	0.935	13,003 bp	0.6917	–	–	–
Skin	rs342296	1.000	684 bp	0.2091	rs13246564	0.045	0.0006
	rs342275	0.935	13,003 bp	0.2537	–	–	–

Table 8-7. Functional ontology classification of differentially expressed genes between *Pik3cg*$^{-/-}$ and wild type mice.

#	GO term	Biological process	*P*-value	Sample frequency	Background frequency	Genes
1	0009987	Cellular process	$3.86\text{x}10^{-16}$	123/187 (65.8%)	11382/33954 (33.5%)	*Tsc1, Rgs10, Tmsb4x, Cst3, Ybx1, Psmd4, Ywhah, Snx15, Fech, Myo6, Map2k2, Acp1, Hist2h2ac, Smox, Atpif1, Ccng2, Gng11, Hist1h1c, Cdkn2c, Glrx5, Bicd2, Clic4, Cela1, C3, Lyz2, Cul4a, Gpx4, Rnf10, Cd81, Sh3bgrl3, Ifit2, Csda, Atp2a3, Car2, Cap1, Bcl2l1, Stx7, Fzr1, Skap2, Dstn, Sorl1, Cmas, Ctnna1, Prdx3, Chka, Acsl1, Ilk, Sytl4, Msi2, Alox12, Nptn, Arhgef3, Vwf, Atox1, Birc2, Actb, F2rl2, Serpine2, Rnf11, Dusp23, Slc44a1, Sdpr, Itpr2, Itgb5, Rffl, Ptp4a3, Pabpc1, Ywhaz, Ifi30, Zyx, Mmd, Plp1, St3gal5, Prdx5, Tpi1, Ehd4, Fancl, Lyz1, Sort1, Gnaz, Fis1, Litaf, Prkar2b, Cux1, Treml1, Lgals3bp, Gp9, Nusap1, Fhl1, F5, Pros1, Ranbp10, Emb, P2ry12, Urod, Gnas, St6galnac2, Itga6, Cdc42ep5, Cd9, Trim10, Arf5, Gp1bb, Snx3, Pygb, Mylk, Dap, Bin1, Epb4.1, Stx11, Spnb1, Ptpn11, Agtrap, Slc2a3, G3bp2, Mast2, Epb4.9, Plek, Ndrg1, Gp5, Pnpo, E2f2, Vcl*
2	0065008	Regulation of biological quality	$2.64\text{x}10^{-14}$	41/187 (21.9%)	1503/33954 (4.4%)	*Tsc1, Tmsb4x, Fech, Myo6, Glrx5, C3, Rnf10, Cd81, Sh3bgrl3, Car2, Dstn, Prdx3, Ilk, Sytl4, Alox12, Nptn, Gp6, Vwf, Atox1, F2rl2, Serpine2, Ywhaz, Ifi30, Plp1, Prdx5, Treml1, Gp9, F5, Pros1, P2ry12, Cdc42ep5, Cd9, Trim10, Gp1bb, Epb4.1, Spnb1, Ptpn11, Agtrap, Epb4.9, Plek, Gp5*
3	0007596	Blood coagulation	$7.84\text{x}10^{-12}$	13/187 (7.0%)	91/33954 (0.3%)	*C3, Gp6, Vwf, F2rl2, Serpine2, Treml1, Gp9, F5, Pros1, P2ry12, Gp1bb, Plek, Gp5*
4	0007599	Haemostasis	$9.09\text{x}10^{-12}$	13/187 (7.0%)	92/33954 (0.3%)	*C3, Gp6, Vwf, F2rl2, Serpine2, Treml1, Gp9, F5, Pros1, P2ry12, Gp1bb, Plek, Gp5*
5	0050817	Coagulation	$1.21\text{x}10^{-11}$	13/187 (7.0%)	94/33954 (0.3%)	*C3, Gp6, Vwf, F2rl2, Serpine2, Treml1, Gp9, F5, Pros1, P2ry12, Gp1bb, Plek, Gp5*

6	0065007	Biological regulation	$3.71\text{x}10^{-10}$	84/187 (44.9%)	7128/33954 (21.0%)	*Tsc1, Rgs10, Tmsb4x, Cst3, Ybx1, Psmd4, Ywhah, Fech, Myo6, Map2k2, Atpif1, Ccng2, Gchfr, Gng11, Cdkn2c, Glrx5, Cela1, C3, Cul4a, Gpx4, Rnf10, Cd81, Sh3bgrl3, Csda, Car2, Bcl2l1, Fzr1, Skap2, Dstn, Ctnna1, Prdx3, Ilk, Sytl4, Alox12, Nptn, Gp6, Arhgef3, Vwf, Atox1, Birc2, F2rl2, Serpine2, Sdpr, Itpr2, Ywhaz, Ifi30, Plp1, Prdx5, Ehd4, Fancl, Gnaz, Litaf, Prkar2b, Cux1, Treml1, Lgals3bp, Gp9, Nusap1, F5, Pros1, Ranbp10, P2ry12, Gnas, B2m, Itga6, Cdc42ep5, Cd9, Trim10, Arf5, Gp1bb, Ctla2b, Bin1, Epb4.1, Spnb1, Ptpn11, Agtrap, Slc2a3, G3bp2, Mast2, Epb4.9, Plek, Gp5, E2f2, Vcl*
7	0042060	Wound healing	$3.06\text{x}10^{-09}$	13/187 (7.0%)	143/33954 (0.4%)	*C3, Gp6, Vwf, F2rl2, Serpine2, Treml1, Gp9, F5, Pros1, P2ry12, Gp1bb, Plek, Gp5*
8	0050878	Regulation of body fluid levels	$1.01\text{x}10^{-08}$	13/187 (7.0%)	157/33954 (0.5%)	*C3, Gp6, Vwf, F2rl2, Serpine2, Treml1, Gp9, F5, Pros1, P2ry12, Gp1bb, Plek, Gp5*
9	0071840	Cellular component organisation or biogenesis	$3.22\text{x}10^{-08}$	42/187 (22.5%)	2396/33954 (7.1%)	*Tsc1, Tmsb4x, Ywhah, Fech, Myo6, Hist2h2ac, Ccng2, Gchfr, Hist1h1c, C3, Lyz2, Gpx4, Cd81, Cap1, Bcl2l1, Fzr1, Dstn, Sorl1, Ctnna1, Ilk, Alox12, Nptn, Birc2, Actb, Serpine2, Pf4, Ywhaz, Ehd4, Lyz1, Sort1, Nusap1, Pros1, Ranbp10, Itga6, Cd9, Bin1, Epb4.1, Spnb1, Ptpn11, Epb4.9, Plek, Vcl*
10	0006950	Response to stress	$5.03\text{x}10^{-08}$	31/187 (16.6%)	1371/33954 (4.0%)	*Cst3, Psmd4, Chi3l3, Fech, Cela1, C3, Lyz2, Cul4a, Gpx4, Fzr1, Prdx3, Gp6, Vwf, Atox1, Birc2, F2rl2, Serpine2, Ywhaz, Fancl, Lyz1, Treml1, Gp9, F5, Pros1, P2ry12, B2m, Cdc42ep5, Gp1bb, Ptpn11, Plek, Gp5*
11	0050896	Response to stimulus	$5.20\text{x}10^{-08}$	45/187 (24.1%)	2752/33954 (8.1%)	*Ifi27l2a, Cst3, Ybx1, Psmd4, Chi3l3, Fech, Myo6, S100a8, Cela1, C3, Lyz2, Cul4a, Gpx4, Ifit2, Bcl2l1, Fzr1, Prdx3, Acsl1, Alox12, Gp6, Vwf, Atox1, Birc2, F2rl2, Serpine2, Pf4, Ywhaz, Oasl2, Fancl, Lyz1, Sort1, Prkar2b, Treml1, Gp9, F5, Pros1, P2ry12, Gnas, B2m, Itga6, Cdc42ep5, Gp1bb, Ptpn11, Plek, Gp5*

Table 8-8. Genotype and phenotype information for TAR cases and unaffected parents. Healthy individuals are on dark grey background. Abbreviations: Unkn.: Unknown; het: heterozygous; del: deletion; M: male; F: female.

Unique Case Number (UCN)	TAR diagnosed	Heterozygous 1q21 deletion	1q21 deletion origin	Genotype 5'-UTR SNP	Genotype intronic SNP	Sex	Age (years)	Gestation/delivery (weeks)	BW (g)	Neonatal problems	Lowest platelet count (x10^9/L)	Highest platelet count (x10^9/L)	Upper limb abnormality	Lower limb abnormality	Cardio-vascular abnormality	Cow's milk intolerance
1	Yes	Yes	*De novo*	A/Del	G/Del	M	31	Unkn.	Unkn.	Diarrhoea	110	111	Radii absent or hypoplastic	Yes	No	Unkn.
105 (mother of 1)	No	No		G/G	G/G											
106 (mother of 1)	No	No		G/A	G/G											
2	Yes	Yes	*De novo*	A/Del	G/Del	F	6 months	41	3,549	Bruising	7	20	Radii absent or hypoplastic	Yes	No	Unkn.
109 (mother of 2)	No	No		G/G	G/G											
110 (father of 2)	No	No		G/A	G/G											
3	Yes	Yes	Maternal	G/Del	C/Del	F	14	Unkn.	3,232	Bruising	90	140	Radii absent or hypoplastic	Yes	Unkn.	Yes
103 (mother of 3)	No	Yes		G/Del	G/Del											
104 (father of 3)	No	No		G/G	G/C											
4	Yes	Yes	Paternal	A/Del	G/Del	F	14	Unkn.	Unkn.	Unkn.	112	Unkn.	Radii absent or hypoplastic	Unkn.	Unkn.	Unkn.
107 (mother of 4)	No	No		G/A	G/G											
108 (father of 4)	No	Yes		G/Del	G/Del											
5	Yes	Yes	Paternal	A/Del	G/Del	F	29	40	3,175	Tube fed	11	78	Radii absent or hypoplastic	No	No	No
111 (mother of 5)	No	No		G/A	G/G											
112 (father of 5)	No	Yes		G/Del	G/Del											
6	Yes	Yes	Paternal	A/Del	G/Del	F	28.5	40	2,900	Unkn.	101	142	Radii absent or hypoplastic	Yes	No	Yes
81 (mother of 6)	No	No		A/A	G/G											
82 (father of 6)	No	Yes		G/Del	G/Del											
7	Yes	Yes	Maternal	A/Del	G/Del	F	22	Unkn.	Unkn.	Unkn.	12	59	Radii absent or hypoplastic	Yes	Yes	No
90 (mother of 7)	No	No		G/A	G/G											
91 (father of 7)	No	Yes		G/Del	G/Del											
8	Yes	Yes	*De novo*	A/Del	G/Del	F	2 days	Unkn.	Unkn.	Unkn.	Unkn.	20	Radii absent or hypoplastic	Yes	No	No

99 (mother of 8)	No	No		A/A	G/G											
100 (father of 8)	No	No		G/A	G/G											
10	Yes	Yes	Paternal	A/Del	G/Del	M	28	40	2,900	Bleeding	8	120	Absence of radius with hypoplasia of humerus and ulna	No	No	Yes
11 (father of 10)	No	Yes		G/G	G/G											
12 (mother of 10)	No	No		G/A	G/G											
13	Yes	Yes	Maternal	A/Del	G/Del	F	5	40	2,670	Phototherapy	37	43	Radii absent or hypoplastic	Yes	No	No
14 (father of 13)	No	No		A/A	G/G											
15 (mother of 13)	No	Yes		G/G	G/G											
16	Yes	Yes	Paternal	A/Del	G/Del	F	15	40	3,150	Bleeding	29	64	Absence of radius, ulna and humerus; hypoplasia of scapula	Yes	No	Yes
17 (father of 16)	No	Yes		G/G	G/G											
18 (mother of 16)	No	No		G/A	G/C											
19	Yes	Yes	Unkn.	A/Del	G/Del	Unkn.	23	40	Unkn.	None	26	133	Radii absent or hypoplastic	Yes	No	No
20	Yes	Yes	Unkn.	A/Del	G/Del	Unkn.	23	40	2,750	Petechiae	10	110	Radii absent or hypoplastic	No	No	No
21	Yes	Yes	Unkn.	A/Del	G/Del	Unkn.	26	40	Unkn.	None	20	53	Radii absent or hypoplastic	Yes	No	No
22	Yes	Yes	Unkn.	A/Del	G/Del	Unkn.	27	40	Unkn.	None	10	40	Radii absent or hypoplastic	No	No	Yes
23	Yes	Yes	*De novo*	A/Del	G/Del	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.
24 (parent of 23)	No	No		G/G	G/G											
25 (parent of 23)	No	No		G/A	G/G											
33	Yes	Het frameshift insertion chr1:145,508,476 (T/TAGCG)	n/a	G/A	G/G	M	29	41	3,110	Petechiae	13	Unkn.	Unkn.	Unkn.	Unkn.	Yes
31 (parent of 33)	No	Het frameshift insertion chr1:145,508,476 (T/TAGCG)		G/G	G/G											
32 (parent of 33)	No	No		G/A	G/G											
40	Yes	Yes	Parent	A/Del	G/Del	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.
41	Yes	Yes	Parent	A/Del	G/Del	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.
38 (parent 1 of 40 and 41)	No	Yes		G/Del	G/Del											
39 (parent 2 of 40 and 41)	No	No		G/A	G/G											
42	Yes	Yes	Maternal	G/Del	C/Del	F	17	40	2,750	Unkn.	9	58	Radii absent or hypoplastic	No	No	Yes
43 (mother of 42)	No	Yes		G/G	G/G											
44 (father of 42)	No	No		G/G	G/C											

47	Yes	Yes	Maternal	A/Del	G/Del	M	4 months	40	3,450	Unkn.	30	233	Radii absent or hypoplastic	Yes	No	No
48 (mother of 47)	No	Yes		G/Del	G/Del											
49 (father of 47)	No	No		G/A	G/Del											
50	Yes	Yes	Maternal	A/Del	G/Del	M	26	Unkn.	Unkn.	Unkn.	Unkn.	163	Radii absent or hypoplastic	Yes	No	No
51 (mother of 50)	No	Yes		G/Del	G/Del											
52 (father of 50)	No	No		G/A	G/G											
53	Yes	Yes	Paternal	A/Del	G/Del	F	1	Unkn.	2,610	None	10	28	Radii absent or hypoplastic	Yes	No	No
54 (mother of 53)	No	No		G/A	G/G											
55 (father of 53)	No	Yes		G/Del	G/Del											
56	Yes	Yes	*De novo*	A/Del	G/Del	F	34	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Radii absent or hypoplastic	Yes	No	No
57 (mother of 56)	No	No		G/G	G/G											
58 (father of 56)	No	No		G/A	G/G											
59	Yes	Yes	*De novo*	A/Del	G/Del	M	26	Unkn.	2,600	Unkn.	10	200	Radii absent or hypoplastic	Yes	Unkn.	Unkn.
60 (mother of 59)	No	No		G/G	G/G											
61 (father of 59)	No	No		G/A	G/G											
64	Yes	Yes	Unkn.	G/Del	C/Del	M	23	Unkn.	Unkn.	Unkn.	94	155	Radii absent or hypoplastic	Yes	Unkn.	Unkn.
65	Yes	Yes	Maternal	A/Del	G/Del	F	34	Unkn.	Unkn.	Unkn.	79	142	Radii absent or hypoplastic	Unkn.	Unkn.	Unkn.
66 (mother of 65)	No	Yes		G/Del	G/Del											
67 (father of 65)	No	No		G/A	G/G											
68	Yes	Yes	Maternal	G/Del	C/Del	M	1.5	Unkn.	Unkn.	Unkn.	79	169	Radii absent or hypoplastic	Unkn.	Unkn.	Unkn.
69 (mother of 68)	No	Yes		G/Del	G/Del											
70	Yes	Yes	Maternal	G/Del	C/Del	F	18	38	2,600	None	34	154	Radii absent or hypoplastic	Yes	No	Unkn.
71	Yes	Yes	Maternal	G/Del	C/Del	F	6 months	39	3,510	Unkn.	30	200	Radii absent or hypoplastic	Yes	Unkn.	Unkn.
72 (mother of 70 and 71)	No	Yes		G/Del	G/Del											
73 (father of 70 and 71)	No	No		G/G	G/C											
74	Yes	Yes	Unkn.	G/Del	C/Del	M	39	Unkn.	Unkn.	Unkn.	79	169	Radii absent or hypoplastic	Yes	No	No
75 (mother of 74)	No	Yes		G/Del	G/Del											
76	Yes	Yes	Maternal	G/Del	C/Del	M	2	40	3,220	Bleeding	8	130	Radii absent or hypoplastic	Yes	No	Unkn.
77 (mother of 76)	No	Yes		G/Del	G/Del											
78 (father of 76)	No	No		G/G	G/C											
83	Yes	Yes	Non-maternal	A/Del	G/Del	F	37	Unkn.	Unkn.	None	74	136	Radii absent or hypoplastic	Yes	Unkn.	Unkn.
84 (mother 83)	No	No		G/G	G/G											

85	Yes	Yes	*De novo*	A/Del	G/Del	F	17 days	37	2,800	Unkn.	Unkn.	167	Radii absent or hypoplastic	No	Yes	No
87 (father of 85)	No	No		G/A	G/G											
90 (mother of 85)	No	No		G/A	G/G											
88	Yes	Yes	Unkn.	A/Del	G/Del	F	4	Unkn.	2,720	Unkn.	Unkn.	34	Radii absent or hypoplastic	No	Unkn.	Unkn.
89	Yes	Yes	Unkn.	G/Del	C/Del	M	8	Unkn.	Unkn.	Unkn.	Unkn.	88	Radii absent or hypoplastic	Yes	Unkn.	Unkn.
92	Yes	Yes	*De novo*	A/Del	G/Del	M	1.5	Unkn.	Unkn.	Unkn.	7	65	Radii absent or hypoplastic	Unkn.	Unkn.	Unkn.
93 (mother of 92)	No	No		G/G	G/G											
94 (father of 92)	No	No		G/A	G/G											
95	Yes	Yes	Maternal	A/Del	G/Del	M	6	39	2,900	Unkn.	18	180	Radii absent or hypoplastic	Yes	Yes	Unkn.
96 (mother of 95)	No	Yes		G/Del	G/Del											
97 (father of 95)	No	No		G/A	G/G											
98	Yes	Yes	Unkn.	A/Del	G/Del	F	4	40	2,910	None	7	295	Radii absent or hypoplastic	No	No	Yes
101	Yes	Yes	Unkn.	A/Del	G/Del	F	17	Unkn.	Unkn.	Unkn.	31	91	Radii absent or hypoplastic	Yes	Unkn.	Unkn.
102 (mother of 101)	No	Yes		G/Del	G/Del											
113	Yes	Yes	Unkn.	A/Del	G/Del	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.
114	Yes	Yes	*De novo*	G/Del	C/Del	Unkn.	8	42	2,960	Bleeding	Unkn.	43	Radii absent or hypoplastic	Yes	No	No
115 (parent of 114)	No	No		G/G	G/G											
131 (parent of 114)	No	No		G/G	G/C											
116	Yes	Yes	Maternal	A/Del	G/Del	F	8	40	3,062	Bleeding	12	91	Radii absent or hypoplastic	Yes	No	No
117 (father of 116)	No	No		G/A	G/G											
118 (mother of 116)	No	Yes		G/Del	G/Del											
121	Yes	Yes	Maternal	G/Del	G/Del	F	36	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Radii absent or hypoplastic	No	No	Yes
145 (mother of 121)	Yes	Yes	Unkn.	G/Del	G/Del	M	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.
120 (father of 121)	No	No		G/G	G/G											
122	Yes	Yes	Unkn.	A/Del	G/Del	M	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.
123	Yes	Yes	Unkn.	G/Del	C/Del	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.
124	Yes	Yes	Unkn.	A/Del	G/Del	F	22	Unkn.	Unkn.	Bruising	Unkn.	Unkn.	Radii absent or hypoplastic	Yes	No	Yes
125	Yes	Yes	Paternal	A/Del	G/Del	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.
126 (father of 125)	No	Yes		G/Del	G/Del											
127 (mother of 125)	No	No		A/G	G/G											
128	Yes	Yes	Unkn.	G/Del	C/Del	F	8	41	3,544	Bleeding	11	178	Radii absent or hypoplastic	No	No	Yes
129 (parent of 128)	No	No		G/G	G/C											

134	Yes	Yes	Maternal	A/Del	G/Del	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.
132 (mother of 134)	No	Yes		G/Del	G/Del											
133 (father of 132)	No	No		G/A	G/G											
136	Yes	Yes	Non-paternal	A/Del	G/Del	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.	Unkn.
135 (father of 136)	No	No		G/G	G/G											
139	Yes	Yes	Unkn.	A/Del	G/Del	F	17	42	3,459	Bleeding	20	55	Radii absent or hypoplastic	No	No	Unkn.
138 (parent of 138)	No	No		G/A	G/G											
140	Yes	Het gain of stop codon chr1:145,509,173 (C/T)	n/a	G/A	G/C	F	9	Unkn.	Unkn.	Bleeding	Unkn.	Unkn.	Unkn.	Unkn.	Yes	Yes
142	Yes	Yes	Paternal	A/Del	G/Del	M	16	40	Unkn.	Bleeding	Unkn.	78	Radii absent or hypoplastic	Yes	No	No
143 (mother of 142)	No	No		G/A	G/G											
144 (father of 142)	No	Yes		G/Del	G/Del											

REFERENCES.

The International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789-796.

The ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636-640.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

The International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661-678.

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

Myocardial Infarction Genetics Consortium (2009). Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat. Genet. *41*, 334-340.

Coronary Artery Disease Consortium (2009). Large scale association analysis of novel genetic loci for coronary artery disease. Arterioscl. Throm. Vas. Biol. *29*, 774-780.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

The Wellcome Trust Case Control Consortium (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature *464*, 713-720.

The ENCODE Project Consortium (2011). A user's guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol. *9*, e1001046.

Adams, D., Altucci, L., Antonarakis, S.E., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., et al. (2012). BLUEPRINT to decode the epigenetic signature written in blood. Nat. Biotechnol. *30*, 224-226.

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L.A. & Rubin, E.M. (2007). Deletion of ultraconserved elements yields viable mice. PLoS Biol. *5*, e234.

Akalin, A., Fredman, D., Arner, E., Dong, X., Bryne, J.C., Suzuki, H., Daub, C.O., Hayashizaki, Y. & Lenhard, B. (2009). Transcriptional features of genomic regulatory blocks. Genome Biol. *10*, R38.

Al Olama, A.A., Kote-Jarai, Z., Giles, G.G., Guy, M., Morrison, J., Severi, G., Leongamornlert, D.A., Tymrakiewicz, M., Jhavar, S., Saunders, E., et al. (2009). Multiple loci on 8q24 associated with prostate cancer susceptibility. Nat. Genet. *41*, 1058-1060.

Albers, C.A., Cvejic, A., Favier, R., Bouwmans, E.E., Alessi, M.-C., Bertone, P., Jordan, G., Kettleborough, R.N.W., Kiddle, G., Kostadima, M., et al. (2011). Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. Nat. Genet. *43*, 735-737.

Albers, C.A., Paul, D.S., Schulze, H., Freson, K., Stephens, J.C., Smethurst, P.A., Jolley, J.D., Cvejic, A., Kostadima, M., Bertone, P., et al. (2012). Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat. Genet. *44*, 435-439.

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., et al. (2007). Direct selection of human genomic loci by microarray hybridization. Nat. Methods *4*, 903-905.

Altshuler, D., Daly, M.J. & Lander, E.S. (2008). Genetic mapping in human disease. Science *322*, 881-888.

Andersson, L.C., Nilsson, K. & Gahmberg, C.G. (1979). K562–a human erythroleukemic cell line. Int. J. Cancer *23*, 143-147.

Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M. & Teichmann, S.A. (2004). Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. *14*, 283-291.

Bai, L. & Morozov, A.V. (2010). Gene regulation by nucleosome positioning. Trends Genet. *26*, 476-483.

Baker, M. (2012). Functional genomics: the changes that count. Nature *482*, 257-262.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745-755.

Bannister, A.J. & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res. *21*, 381-395.

Barrett, J.C. & Cardon, L.R. (2006). Evaluating coverage of genome-wide association studies. Nat. Genet. *38*, 659-662.

Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C., et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. *41*, 703-707.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. & Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Bartholomew, C., Kilbey, A., Clark, A.-M. & Walker, M. (1997). The Evi-1 proto-oncogene encodes a transcriptional repressor activity associated with transformation. Oncogene *14*, 569-577.

Bell, A.C., West, A.G. & Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell *98*, 387-396.

Bell, O., Schwaiger, M., Oakeley, E.J., Lienert, F., Beisel, C., Stadler, M.B. & Schübeler, D. (2010). Accessibility of the Drosophila genome discriminates PcG repression, H4K16 acetylation and replication timing. Nat. Struct. Mol. Biol. *17*, 894-900.

Bell, O., Tiwari, V.K., Thomä, N.H. & Schübeler, D. (2011). Determinants and dynamics of genome accessibility. Nat. Rev. Genet. *12*, 554-564.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. *28*, 1045-1048.

Bird, A.P. (1987). CpG islands as gene markers in the vertebrate nucleus. Trends Genet. *3*, 342-347.

Blow, M.J., McCulley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2010). ChIP-Seq identification of weakly conserved heart enhancers. Nat. Genet. *42*, 806-810.

Bodmer, W. & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet. *40*, 695-701.

Boos, C.J. & Lip, G.Y.H. (2007). Assessment of mean platelet volume in coronary artery disease–what does it mean? Thromb. Res. *120*, 11-13.

Botstein, D., White, R.L., Skolnick, M. & Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. *32*, 314-331.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. & Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell *132*, 311-322.

Boyle, A.P., Guinney, J., Crawford, G.E. & Furey, T.S. (2008). F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics *24*, 2537-2538.

Boyle, A.P., Song, L., Lee, B.-K., London, D., Keefe, D., Birney, E., Iyer, V.R., Crawford, G.E. & Furey, T.S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. *21*, 456-464.

Bustamante, C.D., De La Vega, F.M. & Burchard, E.G. (2011). Genomics for the world. Nature *475*, 163-165.

Cantor, A.B. (2005). GATA transcription factors in hematologic disease. Int. J. Hematol. *81*, 378-384.

Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B. & Lewis, S. (2009). AmiGO: online access to ontology and annotation data. Bioinformatics *25*, 288-289.

Carey, M., Lin, Y.-S., Green, M.R. & Ptashne, M. (1990). A mechanism for synergistic activation of a mammalian gene by GAL4 derivatives. Nature *345*, 361-364.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. *38*, 626-635.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M. & Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics *21*, 2933-2942.

Cheung, V.G. & Spielman, R.S. (2009). Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat. Rev. Genet. *10*, 595-604.

Chin, M.H., Mason, M.J., Xie, W., Volinia, S., Singer, M., Peterson, C., Ambartsumyan, G., Aimiuwu, O., Richter, L., Zhang, J., et al. (2009). Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell *5*, 111-123.

Chow, C.-M., Georgiou, A., Szutorisz, H., Maia e Silva, A., Pombo, A., Barahona, I., Dargelos, E., Canzonetta, C. & Dillon, N. (2005). Variant histone H3.3 marks promoters of transcriptionally active genes during mammalian cell division. EMBO Rep. *6*, 354-360.

Chu, S.G., Becker, R.C., Berger, P.B., Bhatt, D.L., Eikelboom, J.W., Konkle, B., Mohler, E.R., Reilly, M.P. & Berger, J.S. (2010). Mean platelet volume as a predictor of cardiovascular risk: a systematic review and meta-analysis. J. Thromb. Haemost. *8*, 148-156.

Cirulli, E.T. & Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. *11*, 415-425.

Clapier, C.R. & Cairns, B.R. (2009). The biology of chromatin remodeling complexes. Annu. Rev. Biochem. *78*, 273-304.

Clark, T.G., Andrew, T., Cooper, G.M., Margulies, E.H., Mullikin, J.C. & Balding, D.J. (2007). Functional constraint and small insertions and deletions in the ENCODE regions of the human genome. Genome Biol. *8*, R180.

Coffey, A.J., Kokocinski, F., Calafato, M.S., Scott, C.E., Palta, P., Drury, E., Joyce, C.J., LeProust, E.M., Harrow, J., Hunt, S., et al. (2011). The GENCODE exome: sequencing the complete human exome. Eur. J. Hum. Genet. *19*, 827-831.

Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R. & Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science *305*, 869-872.

Cohen, J.C., Boerwinkle, E., Thomas H. Mosley, J. & Hobbs, H.H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N. Engl. J. Med. *354*, 1264-1272.

Collins, F.S., Guyer, M.S. & Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. Science *278*, 1580-1581.

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

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

Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., Green, E.D., Wolfsberg, T.G. & Collins, F.S. (2004). Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc. Natl. Acad. Sci. USA *101*, 992-997.

Crawford, G.E., Davis, S., Scacheri, P.C., Renaud, G., Halawi, M.J., Erdos, M.R., Green, R., Meltzer, P.S., Wolfsberg, T.G. & Collins, F.S. (2006). DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat. Methods *3*, 503-509.

Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y., Bernat, J.A., Ginsburg, D., et al. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. *16*, 123-131.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. USA *107*, 21931-21936.

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2011). Reactome: a database of reactions, pathways and biological processes. Nucl. Acids Res. *39*, D691-D697.

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.-Y., Cui, K. & Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. *19*, 24-32.

Cui, K., Zang, C., Roh, T.-Y., Schones, D.E., Childs, R.W., Peng, W. & Zhao, K. (2009). Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell *4*, 80-93.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. (2001). High-resolution haplotype structure in the human genome. Nat. Genet. *29*, 229-232.

Danesh, J., Collins, R., Appleby, P. & Peto, R. (1998). Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease: meta-analyses of prospective studies. JAMA *279*, 1477-1482.

De Gobbi, M., Viprakasit, V., Hughes, J.R., Fisher, C., Buckle, V.J., Ayyub, H., Gibbons, R.J., Vernimmen, D., Yoshinaga, Y., de Jong, P., et al. (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science *312*, 1215-1217.

de Souza, N. (2010). Primer: induced pluripotency. Nat. Methods *7*, 20-21.

Deaton, A.M. & Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010-1022.

Dermitzakis, E.T. (2012). Cellular genomics for complex traits. Nat. Rev. Genet. *13*, 215-220.

Dion, M.F., Altschuler, S.J., Wu, L.F. & Rando, O.J. (2005). Genomic characterization reveals a simple histone H4 acetylation code. Proc. Natl. Acad. Sci. USA *102*, 5501-5506.

Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C.C., Taylor, J., Burnett, E., Gut, I., Farrall, M., et al. (2007). A genome-wide association study of global gene expression. Nat. Genet. *39*, 1202-1207.

Doi, A., Park, I.-H., Wen, B., Murakami, P., Aryee, M.J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., et al. (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat. Genet. *41*, 1350-1353.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., Lander, E.S., et al. (1987). A genetic linkage map of the human genome. Cell *51*, 319-337.

Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. Nature *456*, 728-731.

Dorahy, D.J., Thorne, R.F., Fecondo, J.V. & Burns, G.F. (1997). Stimulation of platelet activation and aggregation by a carboxyl-terminal peptide from thrombospondin binding to the integrin-associated protein receptor. J. Biol. Chem. *272*, 1323-1330.

Dorschner, M.O., Hawrylycz, M., Humbert, R., Wallace, J.C., Shafer, A., Kawamoto, J., Mack, J., Hall, R., Goldy, J., Sabo, P.J., et al. (2004). High-throughput localization of functional elements by quantitative chromatin profiling. Nat. Methods *1*, 219-225.

Du, P., Kibbe, W.A. & Lin, S.M. (2008). lumi: a pipeline for processing Illumina microarray. Bioinformatics *24*, 1547-1548.

Edery, P., Marcaillou, C., Sahbatou, M., Labalme, A., Chastang, J., Touraine, R., Tubacher, E., Senni, F., Bober, M.B., Nampoothiri, S., et al. (2010). Association of TALS developmental disorder with defect in minor splicing component U4atac snRNA. Science *332*, 240-243.

Edwards, A.O., Ritter, R., III, Abel, K.J., Manning, A., Panhuysen, C. & Farrer, L.A. (2005). Complement factor H polymorphism and age-related macular degeneration. Science *308*, 421-424.

Egelhofer, T.A., Minoda, A., Klugman, S., Lee, K., Kolasinska-Zwierz, P., Alekseyenko, A.A., Cheung, M.-S., Day, D.S., Gadel, S., Gorchakov, A.A., et al. (2011). An assessment of histone-modification antibody quality. Nat. Struct. Mol. Biol. *18*, 91-93.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423-428.

Emison, E.S., McCallion, A.S., Kashuk, C.S., Bush, R.T., Grice, E., Lin, S., Portnoy, M.E., Cutler, D.J., Green, E.D. & Chakravarti, A. (2005). A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature *434*, 857-863.

Ensrud, K. & Grimm, R.H., Jr. (1992). The white blood cell count and risk for coronary heart disease. Am. Heart J. *124*, 207-213.

Epstein, D.J. (2009). Cis-regulatory mutations in human disease. Brief. Funct. Genomics Proteomics *8*, 310-316.

Erdmann, J., Grohennig, A., Braund, P.S., König, I.R., Hengstenberg, C., Hall, A.S., Linsel-Nitschke, P., Kathiresan, S., Wright, B., Trégouët, D.-A., et al. (2009). New susceptibility locus for coronary artery disease on chromosome 3q22.3. Nat. Genet. *41*, 280-282.

Ernst, J. & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817-825.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43-49.

Evans, D.M., Frazer, I.H. & Martin, N.G. (1999). Genetic and environmental causes of variation in basal levels of blood cells. Twin Res. *2*, 250-257.

Falet, H., Hoffmeister, K.M., Neujahr, R. & Hartwig, J.H. (2002). Normal Arp2/3 complex activation in platelets lacking WASp. Blood *100*, 2113-2122.

Farnham, P.J. (2009). Insights from genomic profiling of transcription factors. Nat. Rev. Genet. *10*, 605-616.

Fatemi, M., Pao, M.M., Jeong, S., Gal-Yam, E.N., Egger, G., Weisenberger, D.J. & Jones, P.A. (2005). Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. Nucl. Acids Res. *33*, e176.

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I.K., Sharon, E., Lubling, Y., Widom, J. & Segal, E. (2008). Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. PLoS Comput. Biol. *4*, e1000216.

Firmann, M., Mayor, V., Vidal, P.M., Bochud, M., Pécoud, A., Hayoz, D., Paccaud, F., Preisig, M., Song, K.S., Yuan, X., et al. (2008). The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC Cardiovasc. Disord. *8*, 6.

Forsberg, E.C., Downs, K.M. & Bresnick, E.H. (2000). Direct interaction of NF-E2 with hypersensitive site 2 of the β-globin locus control region in living cells. Blood *96*, 334-339.

Fragoso, G. & Hager, G.L. (1997). Analysis of in vivo nucleosome positions by determination of nucleosome-linker boundaries in crosslinked chromatin. Methods *11*, 246-252.

Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. *10*, 241-251.

Freedman, M.L., Monteiro, A.N.A., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D., et al. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. Nat. Genet. *43*, 513-518.

Freson, K., Devriendt, K., Matthijs, G., Van Hoof, A., De Vos, R., Thys, C., Minner, K., Hoylaerts, M.F., Vermylen, J. & Van Geet, C. (2001). Platelet characteristics in patients with X-linked macrothrombocytopenia because of a novel GATA1 mutation. Blood *98*, 85-92.

Freson, K., Stolarz, K., Aerts, R., Brand, E., Brand-Herrmann, S.-M., Kawecka-Jaszcz, K., Kuznetsova, T., Tikhonoff, V., Thijs, L., Vermylen, J., et al. (2007). -391 C to G substitution in the regulator of G-protein signalling-2 promoter increases susceptibility to the metabolic syndrome in white European men: consistency between molecular and epidemiological studies. J. Hypertens. *25*, 117-125.

Fugman, D.A., Witte, D.P., Jones, C.L.A., Aronow, B.J. & Lieberman, M.A. (1990). In vitro establishment and characterization of a human megakaryoblastic cell line. Blood *75*, 1252-1261.

Furniss, D., Lettice, L.A., Taylor, I.B., Critchley, P.S., Giele, H., Hill, R.E. & Wilkie, A.O.M. (2008). A variant in the sonic hedgehog regulatory sequence (ZRS) is associated with triphalangeal thumb and deregulates expression in the developing limb. Hum. Mol. Genet. *17*, 2417-2423.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225-2229.

Garner, C., Tatu, T., Reittie, J.E., Littlewood, T., Darley, J., Cervino, S., Farrall, M., Kelly, P., Spector, T.D. & Thein, S.L. (2000). Genetic influences on F cells and other hematologic variables: a twin heritability study. Blood *95*, 342-346.

Gaszner, M. & Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nat. Rev. Genet. *7*, 703-713.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. Nat. Genet. *42*, 255-259.

Geddis, A.E. (2006). Inherited thrombocytopenia: congenital amegakaryocytic thrombocytopenia and thrombocytopenia with absent radii. Sem. Hematol. *43*, 196-203.

Geissmann, F., Manz, M.G., Jung, S., Sieweke, M.H., Merad, M. & Ley, K. (2010). Development of monocytes, macrophages, and dendritic cells. Science *327*, 656-661.

Gibson, G. (2012). Rare and common variants: twenty arguments. Nat. Rev. Genet. *13*, 135-145.

Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. Nature *480*, 201-208.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. & Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. *17*, 877-885.

Giresi, P.G. & Lieb, J.D. (2009). Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods *48*, 233-239.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. *27*, 182-189.

Goasguen, J.E., Bennett, J.M., Bain, B.J., Vallespi, T., Brunning, R. & Mufti, G.J. (2009). Morphological evaluation of monocytes and their precursors. Haematologica *94*, 994-997.

Gottschling, D.E. (1992). Telomere-proximal DNA in Saccharomyces cerevisiae is refractory to methyltransferase activity in vivo. Proc. Natl. Acad. Sci. USA *89*, 4062-4065.

Grant, S.F.A., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., et al. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat. Genet. *38*, 320-323.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A. & Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA *108*, 11983-11988.

Greenhalgh, K.L., Howell, R.T., Bottani, A., Ancliff, P.J., Brunner, H.G., Verschuuren-Bemelmans, C.C., Vernon, E., Brown, K.W. & Newbury-Ecob, R.A. (2002). Thrombocytopenia-absent radius syndrome: a clinical genetic study. J. Med. Genet. *39*, 876-881.

Gretarsdottir, S., Baas, A.F., Thorleifsson, G., Holm, H., den Heijer, M., de Vries, J.P., Kranendonk, S.E., Zeebregts, C.J., van Sterkenburg, S.M., Geelkerken, R.H., et al. (2010). Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm. Nat. Genet. *42*, 692-697.

Grice, E.A., Rochelle, E.S., Green, E.D., Chakravarti, A. & McCallion, A.S. (2005). Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. Hum. Mol. Genet. *14*, 3837-3845.

Gudbjartsson, D.F., Bjornsdottir, U.S., Halapi, E., Helgadottir, A., Sulem, P., Jonsdottir, G.M., Thorleifsson, G., Helgadottir, H., Steinthorsdottir, V., Stefansson, H., et al. (2009). Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nat. Genet. *41*, 342-347.

Guey, L.T., Kravic, J., Melander, O., Burtt, N.P., Laramie, J.M., Lyssenko, V., Jonsson, A., Lindholm, E., Tuomi, T., Isomaa, B., et al. (2011). Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet. Epidemiol. *35*, 236-246.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y., et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature *306*, 234-238.

Hachet, O. & Ephrussi, A. (2001). Drosophila Y14 shuttles to the posterior of the oocyte and is required for oskar mRNA transport. Curr. Biol. *11*, 1666-1674.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. Nat. Struct. Mol. Biol. *11*, 394-403.

Haiman, C.A., Patterson, N., Freedman, M.L., Myers, S.R., Pike, M.C., Waliszewska, A., Neubauer, J., Tandon, A., Schirmer, C., McDonald, G.J., et al. (2007). Multiple regions within 8q24 independently affect risk for prostate cancer. Nat. Genet. *39*, 638-644.

Haines, J.L., Hauser, M.A., Schmidt, S., Scott, W.K., Olson, L.M., Gallins, P., Spencer, K.L., Kwan, S.Y., Noureddine, M., Gilbert, J.R., et al. (2005). Complement factor H variant increases the risk of age-related macular degeneration. Science *308*, 419-421.

Hall, J.G., Levin, J., Kuhn, J.P., Ottenheimer, E.J., van Berkum, K.A. & McKusick, V.A. (1969). Thrombocytopenia with absent radius (TAR). Medicine (Baltimore) *48*, 411-439.

Haremaki, T., Sridharan, J., Dvora, S. & Weinstein, D.C. (2010). Regulation of vertebrate embryogenesis by the exon junction complex core component Eif4a3. Dev. Dynam. *239*, 1977-1987.

Harismendy, O. & Frazer, K.A. (2009). Elucidating the role of 8q24 in colorectal cancer. Nat. Genet. *41*, 868-869.

Harismendy, O., Notani, D., Song, X., Rahim, N.G., Tanasa, B., Heintzman, N., Ren, B., Fu, X.-D., Topol, E.J., Rosenfeld, M.G., et al. (2011). 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. Nature *470*, 264-268.

Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M. & Tilghman, S.M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature *405*, 486-489.

Hassan, A.H., Neely, K.E. & Workman, J.L. (2001). Histone acetyltransferase complexes stabilize SWI/SNF binding to promoter nucleosomes. Cell *104*, 817-827.

Hawkins, P.T. & Stephens, L.R. (2007). PI3Kγ is a key regulator of inflammatory responses and cardiovascular homeostasis. Science *318*, 64-66.

Hawkins, R.D., Hon, G.C. & Ren, B. (2010). Next-generation genomics: an integrative approach. Nat. Rev. Genet. *11*, 476-486.

He, H., Liyanarachchi, S., Akagi, K., Nagy, R., Li, J., Dietrich, R.C., Li, W., Sebastian, N., Wen, B., Xin, B., et al. (2011). Mutations in U4atac snRNA, a component of the minor spliceosome, in the developmental disorder MOPD I. Science *332*, 238-240.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311-318.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., Jonasdottir, A., Sigurdsson, A., Baker, A., Palsson, A., et al. (2007). A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science *316*, 1491-1493.

Helgadottir, A., Thorleifsson, G., Magnusson, K.P., Grétarsdottir, S., Steinthorsdottir, V., Manolescu, A., Jones, G.T., Rinkel, G.J.E., Blankensteijn, J.D., Ronkainen, A., et al. (2008). The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nat. Genet. *40*, 217-224.

Henikoff, S. (2008). Nucleosome destabilization in the epigenetic regulation of gene expression. Nat. Rev. Genet. *9*, 15-26.

Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., et al. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat. Methods *6*, 283-289.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. & Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362-9367.

Hirsch, E., Bosco, O., Tropel, P., Laffargue, M., Calvez, R., Altruda, F., Wymann, M.P. & Montrucchio, G. (2001). Resistance to thromboembolism in PI3Kγ-deficient mice. FASEB J. *15*, 2019-2021.

Hoffman, M., Blum, A., Baruch, R., Kaplan, E. & Benjamin, M. (2004). Leukocytes and coronary heart disease. Atherosclerosis *172*, 1-6.

Hon, G., Wang, W. & Ren, B. (2009). Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput. Biol. *5*, e1000566.

Horn, P.J. & Peterson, C.L. (2002). Chromatin higher order folding–wrapping up transcription. Science *297*, 1824-1827.

Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. (2010). Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. *6*, e1001154.

Ioshikhes, I.P. & Zhang, M.Q. (2000). Large-scale human promoter mapping using CpG islands. Nat. Genet. *26*, 61-63.

Jedlitschky, G., Tirschmann, K., Lubenow, L.E., Nieuwenhuis, H.K., Akkerman, J.W.N., Greinacher, A. & Kroemer, H.K. (2004). The nucleotide transporter MRP4 (ABCC4) is highly expressed in human platelets and present in dense granules, indicating a role in mediator storage. Blood *104*, 3603-3610.

Jedlitschky, G., Cattaneo, M., Lubenow, L.E., Rosskopf, D., Lecchi, A., Artoni, A., Motta, G., Nießen, J., Kroemer, H.K. & Greinacher, A. (2010). Role of MRP4 (ABCC4) in platelet adenine nucleotide-storage. Am. J. Pathol. *176*, 1097-1103.

Jenuwein, T. & Allis, C.D. (2001). Translating the histone code. Science *293*, 1074-1080.

Jiang, C. & Pugh, B.F. (2009). Nucleosome positioning and gene regulation: advances through genomics. Nat. Rev. Genet. *10*, 161-172.

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K. & Felsenfeld, G. (2009). H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. Nat. Genet. *41*, 941-945.

Johansen, C.T., Wang, J., Lanktree, M.B., Cao, H., McIntyre, A.D., Ban, M.R., Martins, R.A., Kennedy, B.A., Hassell, R.G., Visser, M.E., et al. (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat. Genet. *42*, 684-687.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.-H., Biddie, S.C., Johnson, T.A., Hager, G.L. & Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet. *43*, 264-268.

Johnson, A.D., Yanek, L.R., Chen, M.-H., Faraday, N., Larson, M.G., Tofler, G., Lin, S.J., Kraja, A.T., Province, M.A., Yang, Q., et al. (2010). Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists. Nat. Genet. *42*, 608-613.

Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Jones, C.I., Bray, S., Garner, S.F., Stephens, J., de Bono, B., Angenent, W.G.J., Bentley, D., Burns, P., Coffey, A., Deloukas, P., et al. (2009). A functional genomics approach reveals novel quantitative trait loci associated with platelet signalling pathways. Blood *114*, 1405-1416.

Jones, P.L., Veenstra, G.C.J., Wade, P.A., Vermaak, D., Kass, S.U., Landsberger, N., Strouboulis, J. & Wolffe, A.P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat. Genet. *19*, 187-191.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.

Kalisky, T. & Quake, S.R. (2011). Single-cell genomics. Nat. Methods *8*, 311-314.

Kamath, S., Blann, A.D. & Lip, G.Y.H. (2001). Platelet activation: assessment and quantification. Eur. Heart J. *22*, 1561-1571.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., et al. (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362-366.

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N.P., Rieder, M.J., Cooper, G.M., Roos, C., Voight, B.F., Havulinna, A.S., et al. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat. Genet. *40*, 189-197.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmond, T.A., Wu, Y., Green, R.D. & Ren, B. (2005). A high-resolution map of active promoters in the human genome. Nature *436*, 876-880.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V. & Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231-1245.

Kim, V.N., Kataoka, N. & Dreyfuss, G. (2001). Role of the nonsense-mediated decay factor hUpf3 in the splicing-dependent exon-exon junction complex. Science *293*, 1832-1836.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. Science *308*, 385-389.

Kleinjan, D.A. & van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. Am. J. Hum. Genet. *76*, 8-32.

Klopocki, E., Schulze, H., Strauß, G., Ott, C.-E., Hall, J., Trotier, F., Fleischhauer, S., Greenhalgh, L., Newbury-Ecob, R.A., Neumann, L.M., et al. (2007). Complex inheritance pattern resembling autosomal recessive inheritance involving a microdeletion in thrombocytopenia-absent radius syndrome. Am. J. Hum. Genet. *80*, 232-240.

Kobor, M.S., Venkatasubrahmanyam, S., Meneghini, M.D., Gin, J.W., Jennings, J.L., Link, A.J., Madhani, H.D. & Rine, J. (2004). A protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. PLoS Biol. *2*, e131.

Koeffler, H.P. & Golde, D.W. (1980). Human myeloid leukemia cell lines: a review. Blood *56*, 344-350.

Konev, A.Y., Tribus, M., Park, S.Y., Podhraski, V., Lim, C.Y., Emelyanov, A.V., Vershilova, E., Pirrotta, V., Kadonaga, J.T., Lusser, A., et al. (2007). CHD1 motor protein is required for deposition of histone variant H3.3 into chromatin in vivo. Science *317*, 1087-1090.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693-705.

Kruglyak, L. & Nickerson, D.A. (2001). Variation is the spice of life. Nat. Genet. *27*, 234-236.

Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA *106*, 3871-3876.

Kurokawa, M., Mitani, K., Imai, Y., Ogawa, S., Yazaki, Y. & Hirai, H. (1998). The t(3;21) fusion product, AML1/Evi-1, interacts with Smad3 and blocks transforming growth factor-β-mediated growth inhibition of myeloid cells. Blood *92*, 4003-4012.

Kurokawa, M., Mitani, K., Irie, K., Matsuyama, T., Takahashi, T., Chiba, S., Yazaki, Y., Matsumoto, K. & Hirai, H. (1998). The oncoprotein Evi-1 represses TGF-β signalling by inhibiting Smad3. Nature *394*, 92-96.

Lander, E.S. (1996). The new genomics: global views of biology. Science *274*, 536-539.

Lander, E.S. (2011). Initial impact of the sequencing of the human genome. Nature *470*, 187-197.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832-838.

Lawrence, M., Gentleman, R. & Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. Bioinformatics *25*, 1841-1842.

Lawrence, R., Day-Williams, A.G., Mott, R., Broxholme, J., Cardon, L.R. & Zeggini, E. (2009). GLIDERS–a web-based search engine for genome-wide linkage disequilibrium between HapMap SNPs. BMC Bioinformatics *10*, 367.

Le Hir, H., Gatfield, D., Izaurralde, E. & Moore, M.J. (2001). The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. EMBO J. *20*, 4987-4997.

Lenhard, B., Sandelin, A. & Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat. Rev. Genet. *13*, 233-245.

Lettice, L.A., Horikoshi, T., Heaney, S.J.H., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N., et al. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc. Natl. Acad. Sci. USA *99*, 7548-7553.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E. & de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725-1735.

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. Curr. Biol. *20*, R754-R763.

Levy, D., Ehret, G.B., Rice, K., Verwoert, G.C., Launer, L.J., Dehghan, A., Glazer, N.L., Morrison, A.C., Johnson, A.D., Aspelund, T., et al. (2009). Genome-wide association study of blood pressure and hypertension. Nat. Genet. *41*, 677-687.

Li, G., Levitus, M., Bustamante, C. & Widom, J. (2005). Rapid spontaneous accessibility of nucleosomal DNA. Nat. Struct. Mol. Biol. *12*, 46-53.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, Q., Peterson, K.R., Fang, X. & Stamatoyannopoulos, G. (2002). Locus control regions. Blood *100*, 3077-3086.

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

Liang, G., Lin, J.C.Y., Wei, V., Yoo, C., Cheng, J.C., Nguyen, C.T., Weisenberger, D.J., Egger, G., Takai, D., Gonzales, F.A., et al. (2004). Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proc. Natl. Acad. Sci. USA *101*, 7357-7362.

Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., Vermeire, S., Dewit, O., de Vos, M., Dixon, A., et al. (2007). Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet. *3*, e58.

Lin, Y.-S., Carey, M., Ptashne, M. & Green, M.R. (1990). How different eukaryotic transcriptional activators can cooperate promiscuously. Nature *345*, 359-361.

Lomvardas, S. & Thanos, D. (2001). Nucleosome sliding via TBP DNA binding in vivo. Cell *106*, 685-696.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M. & Frazer, K.A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science *288*, 136-140.

Lowe, C.E., Cooper, J.D., Brusko, T., Walker, N.M., Smyth, D.J., Bailey, R., Bourget, K., Plagnol, V., Field, S., Atkinson, M., et al. (2007). Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. Nat. Genet. *39*, 1074-1082.

Lozzio, C.B. & Lozzio, B.B. (1975). Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. Blood *45*, 321-334.

Lucas, I., Palakodeti, A., Jiang, C., Young, D.J., Jiang, N., Fernald, A.A. & Le Beau, M.M. (2007). High-throughput mapping of origins of replication in human cells. EMBO Rep. *8*, 770-777.

Ludlow, L.B., Schick, B.P., Budarf, M.L., Driscoll, D.A., Zackai, E.H., Cohen, A. & Konkle, B.A. (1996). Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibβ promoter resulting in the Bernard-Soulier syndrome. J. Biol. Chem. *271*, 22076-22080.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F. & Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature *389*, 251-260.

Lunter, G. & Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. *21*, 936-939.

Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C.Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D.A., et al. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N. Engl. J. Med. *362*, 1181-1191.

Lykke-Andersen, J., Shu, M.-D. & Steitz, J.A. (2001). Communication of the position of exon-exon junctions to the mRNA surveillance machinery by the protein RNPS1. Science *293*, 1836-1839.

Macaulay, I.C., Tijssen, M.R., Thijssen-Timmer, D.C., Gusnanto, A., Steward, M., Burns, P., Langford, C.F., Ellis, P.D., Dudbridge, F., Zwaginga, J.-J., et al. (2007). Comparative gene expression profiling of in vitro differentiated megakaryocytes and erythroblasts identifies novel activatory and inhibitory platelet membrane proteins. Blood *109*, 3260-3269.

Maher, B. (2008). Personal genomes: the case of the missing heritability. Nature *456*, 18-21.

Malik, S. & Roeder, R.G. (2010). The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. Nat. Rev. Genet. *11*, 761-772.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747-753.

Maston, G.A., Evans, S.K. & Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet. *7*, 29-59.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., et al. (2009). Reactome knowledgebase of human biological pathways and processes. Nucl. Acids Res. *37*, D619-D622.

May, D., Blow, M.J., Kaplan, T., McCulley, D.J., Jensen, B.C., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., et al. (2012). Large-scale discovery of enhancers from human heart tissue. Nat. Genet. *44*, 89-93.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A. & Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356-369.

McCarthy, M.I. & Hirschhorn, J.N. (2008). Genome-wide association studies: potential next steps on a genetic journey. Hum. Mol. Genet. *17*, R156-R165.

McDonald, T.P. & Sullivan, P.S. (1993). Megakaryocytic and erythrocytic cell lines share a common precursor cell. Exp. Hematol. *21*, 1316-1320.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297-1303.

McKittrick, E., Gafken, P.R., Ahmad, K. & Henikoff, S. (2004). Histone H3.3 is enriched in covalent modifications associated with active chromatin. Proc. Natl. Acad. Sci. USA *101*, 1525-1530.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M. & Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495-501.

McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D.R., Hinds, D.A., Pennacchio, L.A., Tybjaerg-Hansen, A., Folsom, A.R., et al. (2007). A common allele on chromosome 9 associated with coronary heart disease. Science *316*, 1488-1491.

Meisinger, C., Prokisch, H., Gieger, C., Soranzo, N., Mehta, D., Rosskopf, D., Lichtner, P., Klopp, N., Stephens, J., Watkins, N.A., et al. (2009). A genome-wide association study identifies three loci associated with mean platelet volume. Am. J. Hum. Genet. *84*, 66-71.

Metzker, M.L. (2010). Sequencing technologies–the next generation. Nat. Rev. Genet. *11*, 31-46.

Michelson, A.D. (2010). Antiplatelet therapies for the treatment of cardiovascular disease. Nat. Rev. Drug Discov. *9*, 154-169.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-561.

Miranda-Saavedra, D. & Göttgens, B. (2008). Transcriptional regulatory networks in haematopoiesis. Curr. Opin. Genet. Dev. *18*, 530-535.

Mito, Y., Henikoff, J.G. & Henikoff, S. (2007). Histone replacement marks the boundaries of cis-regulatory domains. Science *315*, 1408-1411.

Mizuguchi, G., Shen, X., Landry, J., Wu, W.-H., Sen, S. & Wu, C. (2004). ATP-driven exchange of histone H2AZ variant catalyzed by SWR1 chromatin remodeling complex. Science *303*, 343-348.

Moffatt, M.F., Kabesch, M., Liang, L., Dixon, A.L., Strachan, D., Heath, S., Depner, M., von Berg, A., Bufe, A., Rietschel, E., et al. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature *448*, 470-473.

Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pagès, H. & Gentleman, R. (2009). ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. Bioinformatics *25*, 2607-2608.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714-719.

Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. Nat. Genet. *40*, 1124-1129.

Nagy, P.L., Cleary, M.L., Brown, P.O. & Lieb, J.D. (2003). Genome-wide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. Proc. Natl. Acad. Sci. USA *100*, 6364-6369.

Narlikar, L. & Ovcharenko, I. (2009). Identifying regulatory elements in eukaryotic genomes. Brief. Funct. Genomics Proteomics *8*, 215-230.

Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science *324*, 387-389.

Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M.D., Bochud, M., Coin, L., Najjar, S.S., Zhao, J.H., Heath, S.C., Eyheramendy, S., et al. (2009). Genome-wide association study identifies eight loci associated with blood pressure. Nat. Genet. *41*, 666-676.

Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat. Genet. *42*, 790-793.

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. *42*, 30-35.

Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I. & Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. *6*, e1000895.

Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., et al. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. *7*, e1002003.

Nicholson, P., Yepiskoposyan, H., Metze, S., Zamudio Orozco, R., Kleinschmidt, N. & Mühlemann, O. (2010). Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors. Cell. Mol. Life Sci. *67*, 677-700.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E. & Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.

Nikopoulos, K., Gilissen, C., Hoischen, A., van Nouhuys, C.E., Boonstra, F.N., Blokland, E.A.W., Arts, P., Wieskamp, N., Strom, T.M., Ayuso, C., et al. (2010). Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy. Am. J. Hum. Genet. *86*, 240-247.

Nobrega, M.A., Ovcharenko, I., Afzal, V. & Rubin, E.M. (2003). Scanning human gene deserts for long-range enhancers. Science *302*, 413.

O'Donnell, C.J., Kavousi, M., Smith, A.V., Kardia, S.L., Feitosa, M.F., Hwang, S.J., Sun, Y.V., Province, M.A., Aspelund, T., Dehghan, A., et al. (2011). Genome-wide association study for coronary artery calcification with follow-up in myocardial infarction. Circulation *124*, 2855-2864.

Ogbourne, S. & Antalis, T.M. (1998). Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. Biochem. J. *331*, 1-14.

Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J. & Zwick, M.E. (2007). Microarray-based genomic selection for high-throughput resequencing. Nat. Methods *4*, 907-909.

Ong, C.-T. & Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. *12*, 283-293.

Orford, K., Kharchenko, P., Lai, W., Dao, M.C., Worhunsky, D.J., Ferro, A., Janzen, V., Park, P.J. & Scadden, D.T. (2008). Differential H3K4 methylation identifies developmentally poised hematopoietic genes. Dev. Cell *14*, 798-809.

Org, E., Eyheramendy, S., Juhanson, P., Gieger, C., Lichtner, P., Klopp, N., Veldre, G., Döring, A., Viigimaa, M., Sõber, S., et al. (2009). Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two European populations. Hum. Mol. Genet. *18*, 2288-2296.

Orkin, S.H. (2000). Diversification of haematopoietic stem cells to specific lineages. Nat. Rev. Genet. *1*, 57-64.

Orkin, S.H. & Zon, L.I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. Cell *132*, 631-644.

Palacios, I.M., Gatfield, D., St Johnston, D. & Izaurralde, E. (2004). An eIF4AIII-containing complex required for mRNA localization and nonsense-mediated mRNA decay. Nature *427*, 753-757.

Pang, L., Weiss, M.J. & Poncz, M. (2005). Megakaryocyte biology and related disorders. J. Clin. Invest. *115*, 3332-3338.

Panne, D. (2008). The enhanceosome. Curr. Opin. Struct. Biol. *18*, 236-242.

Panzenböck, B., Bartunek, P., Mapara, M.Y. & Zenke, M. (1998). Growth and differentiation of human stem cell factor/erythropoietin-dependent erythroid progenitor cells in vitro. Blood *92*, 3658-3668.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669-680.

Pasquet, J.-M., Gross, B.S., Gratacap, M.-P., Quek, L., Pasquet, S., Payrastre, B., van Willigen, G., Mountford, J.C. & Watson, S.P. (2000). Thrombopoietin potentiates collagen receptor signaling in platelets through a phosphatidylinositol 3-kinase-dependent pathway. Blood *95*, 3429-3434.

Paul, D.S., Nisbet, J.P., Yang, T.-P., Meacham, S., Rendon, A., Hautaviita, K., Tallila, J., White, J., Tijssen, M.R., Sivapalaratnam, S., et al. (2011). Maps of open chromatin guide the functional follow-up of genome-wide association signals: application to hematological traits. PLoS Genet. *7*, e1002139.

Pe'er, I., de Bakker, P.I.W., Maller, J., Yelensky, R., Altshuler, D. & Daly, M.J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat. Genet. *38*, 663-667.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499-502.

Pennacchio, L.A. & Visel, A. (2010). Limits of sequence and functional conservation. Nat. Genet. *42*, 557-558.

Pennisi, E. (2011). Disease risk links to gene regulation. Science *332*, 1031.

Perkins, A.S., Mercer, J.A., Jenkins, N.A. & Copeland, N.G. (1991). Patterns of Evi-1 expression in embryonic and adult tissues suggest that Evi-1 plays an important regulatory role in mouse development. Development *111*, 479-487.

Phillips, J.E. & Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194-1211.

Pickrell, J.K., Gaffney, D.J., Gilad, Y. & Pritchard, J.K. (2011). False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. Bioinformatics *27*, 2144-2146.

Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y. & Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. *21*, 447-455.

Plomin, R., Haworth, C.M.A. & Davis, O.S.P. (2009). Common disorders are quantitative traits. Nat. Rev. Genet. *10*, 872-878.

Pomerantz, M.M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M.P., Doddapaneni, H., Beckwith, C.A., Chan, J.A., Hills, A., Davis, M., et al. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat. Genet. *41*, 882-884.

Ponjavic, J., Lenhard, B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. & Sandelin, A. (2006). Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. Genome Biol. *7*, R78.

Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F., et al. (2007). Multiplex amplification of large sets of human exons. Nat. Methods *4*, 931-936.

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E.M., Couronne, O. & Pennacchio, L.A. (2006). Close sequence comparisons are sufficient to identify human cis-regulatory elements. Genome Res. *16*, 855-863.

Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. *69*, 124-137.

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

Rach, E.A., Winter, D.R., Benjamin, A.M., Corcoran, D.L., Ni, T., Zhu, J. & Ohler, U. (2011). Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. PLoS Genet. *7*, e1001274.

Raisner, R.M., Hartley, P.D., Meneghini, M.D., Bao, M.Z., Liu, C.L., Schreiber, S.L., Rando, O.J. & Madhani, H.D. (2005). Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. Cell *123*, 233-248.

Recillas-Targa, F., Pikaart, M.J., Burgess-Beusse, B., Bell, A.C., Litt, M.D., West, A.G., Gaszner, M. & Felsenfeld, G. (2002). Position-effect protection and enhancer blocking by the chicken β-globin insulator are separable activities. Proc. Natl. Acad. Sci. USA *99*, 6883-6888.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., et al. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199-204.

Reich, D.E. & Lander, E.S. (2001). On the allelic spectrum of human disease. Trends Genet. *17*, 502-510.

Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516-1517.

Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., et al. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. *43*, 1066-1073.

Roach, J.C., Glusman, G., Smit, A.F.A., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science *328*, 636-639.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods *4*, 651-657.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. & Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24-26.

Roider, H.G., Lenhard, B., Kanhere, A., Haas, S.A. & Vingron, M. (2009). CpG-depleted promoters harbor tissue-specific transcription factor binding signals–implications for motif overrepresentation analyses. Nucl. Acids Res. *37*, 6305-6315.

Roth, G.J. (1991). Developing relationships: arterial platelet adhesion, glycoprotein Ib, and leucine-rich glycoproteins. Blood *77*, 5-19.

Rotival, M., Zeller, T., Wild, P.S., Maouche, S., Szymczak, S., Schillert, A., Castagné, R., Deiseroth, A., Proust, C., Brocheton, J., et al. (2011). Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. PLoS Genet. *7*, e1002367.

Rozen, S. & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *In:* Methods in Molecular Biology: Bioinformatics–Methods and Protocols. Humana Press.

Rückle, T., Schwarz, M.K. & Rommel, C. (2006). PI3Kγ inhibition: towards an 'aspirin of the 21st century'? Nat. Rev. Drug Discov. *5*, 903-918.

Sabo, P.J., Hawrylycz, M., Wallace, J.C., Humbert, R., Yu, M., Shafer, A., Kawamoto, J., Hall, R., Mack, J., Dorschner, M.O., et al. (2004). Discovery of functional noncoding elements by digital analysis of chromatin structure. Proc. Natl. Acad. Sci. USA *101*, 16837-16842.

Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., et al. (2006). Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat. Methods *3*, 511-518.

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M. & Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development *132*, 797-803.

Saito, H. (1997). Megakaryocytic cell lines. Baillière Clin. Haem. *10*, 47-63.

Sakamaki, S., Hirayama, Y., Matsunaga, T., Kuroda, H., Kusakabe, T., Akiyama, T., Konuma, Y., Sasaki, K., Tsuji, N., Okamoto, T., et al. (1999). Transforming growth factor-β1 (TGF-β1) induces thrombopoietin from bone marrow stromal cells, which stimulates the expression of TGF-β receptor on megakaryocytes and, in turn, renders them susceptible to suppression by TGF-β itself with high specificity. Blood *94*, 1961-1970.

Salicioni, A.M., Xi, M., Vanderveer, L.A., Balsara, B., Testa, J.R., Dunbrack, R.L., Jr. & Godwin, A.K. (2000). Identification and structural analysis of human RBM8A and RBM8B: two highly conserved RNA-binding motif proteins that interact with OVCA1, a candidate tumor suppressor. Genomics *69*, 54-62.

Samani, N.J., Erdmann, J., Hall, A.S., Hengstenberg, C., Mangino, M., Mayer, B., Dixon, R.J., Meitinger, T., Braund, P., Wichmann, H.-E., et al. (2007). Genomewide association analysis of coronary artery disease. N. Engl. J. Med. *357*, 443-453.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. & Hume, D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat. Rev. Genet. *8*, 424-436.

Sanseau, P., Agarwal, P., Barnes, M.R., Pastinen, T., Richards, J.B., Cardon, L.R. & Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. Nat. Biotechnol. *30*, 317-320.

Sasaki, T., Irie-Sasaki, J., Jones, R.G., Oliveira-dos-Santos, A.J., Stanford, W.L., Bolon, B., Wakeham, A., Itie, A., Bouchard, D., Kozieradzki, I., et al. (2000). Function of PI3Kγ in thymocyte development, T cell activation, and neutrophil migration. Science *287*, 1040-1046.

Schadt, E.E. (2009). Molecular networks as sensors and drivers of common human diseases. Nature *461*, 218-223.

Schneider, R., Bannister, A.J., Myers, F.A., Thorne, A.W., Crane-Robinson, C. & Kouzarides, T. (2004). Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. Nat. Cell Biol. *6*, 73-77.

Schoenwaelder, S.M., Ono, A., Sturgeon, S., Chan, S.M., Mangin, P., Maxwell, M.J., Turnbull, S., Mulchandani, M., Anderson, K., Kauffenstein, G., et al. (2007). Identification of a unique co-operative phosphoinositide 3-kinase signaling mechanism regulating integrin αIIbβ3 adhesive function in platelets. J. Biol. Chem. *282*, 28648-28658.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G. & Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887-898.

Schones, D.E. & Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. Nat. Rev. Genet. *9*, 179-191.

Schotta, G., Lachner, M., Sarma, K., Ebert, A., Sengupta, R., Reuter, G., Reinberg, D. & Jenuwein, T. (2004). A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. Genes Dev. *18*, 1251-1262.

Schunkert, H., Götz, A., Braund, P., McGinnis, R., Trégouët, D.-A., Mangino, M., Linsel-Nitschke, P., Cambien, F., Hengstenberg, C., Stark, K., et al. (2008). Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary artery disease. Circulation *117*, 1675-1684.

Schunkert, H., König, I.R., Kathiresan, S., Reilly, M.P., Assimes, T.L., Holm, H., Preuss, M., Stewart, A.F.R., Barbalic, M., Gieger, C., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat. Genet. *43*, 333-338.

Schuster, S.C., Miller, W., Ratan, A., Tomsho, L.P., Giardine, B., Kasson, L.R., Harris, R.S., Petersen, D.C., Zhao, F., Qi, J., et al. (2010). Complete Khoisan and Bantu genomes from southern Africa. Nature *463*, 943-947.

Schwartz, B.E. & Ahmad, K. (2005). Transcriptional activation triggers deposition and removal of the histone variant H3.3. Genes Dev. *19*, 804-814.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z. & Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772-778.

Senyuk, V., Sinha, K.K., Li, D., Rinaldi, C.R., Yanamandra, S. & Nucifora, G. (2007). Repression of RUNX1 activity by EVI1: a new role of EVI1 in leukemogenesis. Cancer Res. *67*, 5658-5666.

Sexton, T., Bantignies, F. & Cavalli, G. (2009). Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation. Sem. Cell Dev. Biol. *20*, 849-855.

Shaw, S. & Oliver, R.A.M. (1959). Congenital hypoplastic thrombocytopenia with skeletal deformities in siblings. Blood *14*, 374-377.

Shimizu, S., Nagasawa, T., Katoh, O., Komatsu, N., Yokota, J. & Morishita, K. (2002). EVI1 is expressed in megakaryocyte cell lineage and enforced expression of EVI1 in UT-7/GM cells induces megakaryocyte differentiation. Biochem. Biophys. Res. Commun. *292*, 609-616.

Shogren-Knaak, M., Ishii, H., Sun, J.-M., Pazin, M.J., Davie, J.R. & Peterson, C.L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science *311*, 844-847.

Sieberts, S.K. & Schadt, E.E. (2007). Moving toward a system genetics view of disease. Mamm. Genome *18*, 389-401.

Simon, J.M., Giresi, P.G., Davis, I.J. & Lieb, J.D. (2012). Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nat. Protoc. *7*, 256-267.

Singh, J. & Klar, A.J.S. (1992). Active genes in budding yeast display enhanced in vivo accessibility to foreign DNA methylases: a novel in vivo probe for chromatin structure of yeast. Genes Dev. *6*, 186-196.

Slavka, G., Perkmann, T., Haslacher, H., Greisenegger, S., Marsik, C., Wagner, O.F. & Endler, G. (2011). Mean platelet volume may represent a predictive parameter for overall vascular mortality and ischemic heart disease. Arterioscler. Thromb. Vasc. Biol. *31*, 1215-1218.

Smith, C.L. & Peterson, C.L. (2005). ATP-dependent chromatin remodeling. Curr. Top. Dev. Biol. *65*, 115-148.

Song, L., Zhang, Z., Grasfeder, L.L., Boyle, A.P., Giresi, P.G., Lee, B.-K., Sheffield, N.C., Gräf, S., Huss, M., Keefe, D., et al. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. *21*, 1757-1767.

Soranzo, N., Rendon, A., Gieger, C., Jones, C.I., Watkins, N.A., Menzel, S., Döring, A., Stephens, J., Prokisch, H., Erber, W., et al. (2009). A novel variant on chromosome 7q22.3 associated with mean platelet volume, counts and function. Blood *113*, 3831-3837.

Soranzo, N., Spector, T.D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat. Genet. *41*, 1182-1190.

Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R. & Flavell, R.A. (2005). Interchromosomal associations between alternatively expressed loci. Nature *435*, 637-645.

Stein, A., Takasuka, T.E. & Collings, C.K. (2010). Are nucleosome positions in vivo primarily determined by histone-DNA sequence preferences? Nucl. Acids Res. *38*, 709-719.

Stitziel, N.O., Kiezun, A. & Sunyaev, S. (2011). Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol. *12*, 227.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545-15550.

Suganuma, T., Gutiérrez, J.L., Li, B., Florens, L., Swanson, S.K., Washburn, M.P., Abmayr, S.M. & Workman, J.L. (2008). ATAC is a double histone acetyltransferase complex that stimulates nucleosome sliding. Nat. Struct. Mol. Biol. *15*, 364-372.

Suzuki, R. & Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics *22*, 1540-1542.

Tabilio, A., Pelicci, P.G., Vinci, G., Mannoni, P., Civin, C.I., Vainchenker, W., Testa, U., Lipinski, M., Rochant, H. & Breton-Gorius, J. (1983). Myeloid and megakaryocytic properties of K-562 cell lines. Cancer Res. *43*, 4569-4574.

Talbert, P.B. & Henikoff, S. (2010). Histone variants–ancient wrap artists of the epigenome. Nat. Rev. Mol. Cell Biol. *11*, 264-275.

Tarpey, P.S., Raymond, F.L., Nguyen, L.S., Rodriguez, J., Hackett, A., Vandeleur, L., Smith, R., Shoubridge, C., Edkins, S., Stevens, C., et al. (2007). Mutations in UPF3B, a member of the nonsense-mediated mRNA decay complex, cause syndromic and nonsyndromic mental retardation. Nat. Genet. *39*, 1127-1133.

Taverna, S.D., Li, H., Ruthenburg, A.J., Allis, C.D. & Patel, D.J. (2007). How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. Nat. Struct. Mol. Biol. *14*, 1025-1040.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707-713.

Thomas-Chollier, M., Hufton, A., Heinig, M., O'Keeffe, S., El Masri, N., Roider, H.G., Manke, T. & Vingron, M. (2011). Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. Nat. Protoc. *6*, 1860-1869.

Tijssen, M.R., Cvejic, A., Joshi, A., Hannah, R.L., Ferreira, R., Forrai, A., Bellissimo, D.C., Oram, S.H., Smethurst, P.A., Wilson, N.K., et al. (2011). Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. Dev. Cell *20*, 597-609.

Todd, J.A., Walker, N.M., Cooper, J.D., Smyth, D.J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S.F., Payne, F., et al. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat. Genet. *39*, 857-864.

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W., et al. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat. Genet. *39*, 984-988.

Trégouët, D.-A., König, I.R., Erdmann, J., Munteanu, A., Braund, P.S., Hall, A.S., Grohennig, A., Linsel-Nitschke, P., Perret, C., DeSuremain, M., et al. (2009). Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. Nat. Genet. *41*, 283-285.

Trip, M.D., Cats, V.M., van Capelle, F.J. & Vreeken, J. (1990). Platelet hyperreactivity and prognosis in survivors of myocardial infarction. N. Engl. J. Med. *322*, 1549-1554.

Trynka, G., Hunt, K.A., Bockett, N.A., Romanos, J., Mistry, V., Szperl, A., Bakker, S.F., Bardella, M.T., Bhaw-Rosun, L., Castillejo, G., et al. (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat. Genet. *43*, 1193-1201.

Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., Björklund, M., Wei, G., Yan, J., Niittymäki, I., et al. (2009). The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat. Genet. *41*, 885-890.

Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z. & Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. Nature *474*, 516-520.

van der Harst, P., Zhang, W., Leach, I.M., Rendon, A., Verweij, N., Sehmi, J., Paul, D.S., Elling, U., Allayee, H., Li, X., et al. (2012). 75 genetic loci influencing the human red blood cell. Nature, in press.

van Steensel, B. & Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. Nat. Biotechnol. *28*, 1089-1095.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. *10*, 252-263.

Verlaan, D.J., Berlivet, S., Hunninghake, G.M., Madore, A.-M., Larivière, M., Moussette, S., Grundberg, E., Kwan, T., Ouimet, M., Ge, B., et al. (2009). Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. Am. J. Hum. Genet. *85*, 377-393.

Vilar, J.M.G. & Saiz, L. (2005). DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise. Curr. Opin. Genet. Dev. *15*, 136-144.

Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E.M. & Pennacchio, L.A. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat. Genet. *40*, 158-160.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854-858.

Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. (2012). Five years of GWAS discovery. Am. J. Hum. Genet. *90*, 7-24.

Visvader, J. & Adams, J.M. (1993). Megakaryocytic differentiation induced in 416B myeloid cells by GATA-2 and GATA-3 transgenes or 5-azacytidine is tightly coupled to GATA-1 expression. Blood *82*, 1493-1501.

Visvader, J.E., Elefanty, A.G., Strasser, A. & Adams, J.M. (1992). GATA-1 but not SCL induces megakaryocytic differentiation in an early myeloid line. EMBO J. *11*, 4557-4564.

Visvader, J.E., Crossley, M., Hill, J., Orkin, S.H. & Adams, J.M. (1995). The C-terminal zinc finger of GATA-1 or GATA-2 is sufficient to induce megakaryocytic differentiation of an early myeloid cell line. Mol. Cell. Biol. *15*, 634-641.

Volpi, L., Roversi, G., Colombo, E.A., Leijsten, N., Concolino, D., Calabria, A., Mencarelli, M.A., Fimiani, M., Macciardi, F., Pfundt, R., et al. (2009). Targeted next-generation sequencing appoints c16orf57 as clericuzio-type poikiloderma with neutropenia gene. Am. J. Hum. Genet. *86*, 72-76.

Wang, X. & Hayes, J.J. (2008). Acetylation mimics within individual core histone tail domains indicate distinct roles in regulating the stability of higher-order chromatin structure. Mol. Cell. Biol. *28*, 227-236.

Wang, Y., O'Connell, J.R., McArdle, P.F., Wade, J.B., Dorff, S.E., Shah, S.J., Shi, X., Pan, L., Rampersaud, E., Shen, H., et al. (2009). Whole-genome association study identifies STK39 as a hypertension susceptibility gene. Proc. Natl. Acad. Sci. USA *106*, 226-231.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M.Q., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet. *40*, 897-903.

Watkins, N.A., Gusnanto, A., de Bono, B., De, S., Miranda-Saavedra, D., Hardie, D.L., Angenent, W.G.J., Attwood, A.P., Ellis, P.D., Erber, W., et al. (2009). A HaemAtlas: characterizing gene expression in differentiated human blood cells. Blood *113*, e1-e9.

Weischenfeldt, J., Damgaard, I., Bryder, D., Theilgaard-Mönch, K., Thoren, L.A., Nielsen, F.C., Jacobsen, S.E.W., Nerlov, C. & Porse, B.T. (2008). NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. Genes Dev. *22*, 1381-1396.

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature *451*, 796-801.

Whitehouse, I., Rando, O.J., Delrow, J. & Tsukiyama, T. (2007). Chromatin remodelling at promoters suppresses antisense transcription. Nature *450*, 1031-1035.

Wiegand, H.L., Lu, S. & Cullen, B.R. (2003). Exon junction complexes mediate the enhancing effect of splicing on mRNA expression. Proc. Natl. Acad. Sci. USA *100*, 11327-11332.

Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., Heath, S.C., Timpson, N.J., Najjar, S.S., Stringham, H.M., et al. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat. Genet. *40*, 161-169.

Wirbelauer, C., Bell, O. & Schübeler, D. (2005). Variant histone H3.3 is deposited at sites of nucleosomal displacement throughout transcribed genes while active histone modifications show a promoter-proximal bias. Genes Dev. *19*, 1761-1766.

Witte, D.P., Harris, R.E., Jenski, L.J. & Lampkin, B.C. (1986). Megakaryoblastic leukemia in an infant: establishment of a megakaryocytic tumor cell line in athymic nude mice. Cancer *58*, 238-244.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al. (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. *3*, e7.

Workman, J.L. & Kingston, R.E. (1992). Nucleosome core displacement in vitro via a metastable transcription factor-nucleosome complex. Science *258*, 1780-1784.

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. *8*, 206-216.

Wu, C., Bingham, P.M., Livak, K.J., Holmgren, R. & Elgin, S.C.R. (1979). The chromatin structure of specific genes: I. evidence for higher order domains of defined DNA sequence. Cell *16*, 797-806.

Wu, C. (1980). The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature *286*, 854-860.

Wu, T.D. & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics *26*, 873-881.

Wymann, M.P., Zvelebil, M. & Laffargue, M. (2003). Phosphoinositide 3-kinase signalling–which way to target? Trends Pharmacol. Sci. *24*, 366-376.

Wysocka, J., Swigut, T., Milne, T.A., Dou, Y., Zhang, X., Burlingame, A.L., Roeder, R.G., Brivanlou, A.H. & Allis, C.D. (2005). WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. Cell *121*, 859-872.

Yang, T.-P., Beazley, C., Montgomery, S.B., Dimas, A.S., Gutierrez-Arcelus, M., Stranger, B.E., Deloukas, P. & Dermitzakis, E.T. (2010). Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. Bioinformatics *26*, 2474-2476.

Yeager, M., Xiao, N., Hayes, R.B., BouVard, P., Desany, B., Burdett, L., Orr, N., Matthews, C., Qi, L., Crenshaw, A., et al. (2008). Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. Hum. Genet. *124*, 161-170.

Yoshimura, K., Nakamura, H., Trapnell, B.C., Dalemans, W., Pavirani, A., Lecocq, J.-P. & Crystal, R.G. (1991). The cystic fibrosis gene has a 'housekeeping'-type promoter and is expressed at low levels in cells of epithelial origin. J. Biol. Chem. *266*, 9140-9144.

Zanke, B.W., Greenwood, C.M.T., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S.M., Prendergast, J., Olschwang, S., Chiang, T., Crowdy, E., et al. (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat. Genet. *39*, 989-994.

Zaret, K.S. & Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Genes Dev. *25*, 2227-2241.

Zeuner, A., Signore, M., Martinetti, D., Bartucci, M., Peschle, C. & De Maria, R. (2007). Chemotherapy-induced thrombocytopenia derives from the selective death of megakaryocyte progenitors and can be rescued by stem cell factor. Cancer Res. *67*, 4767-4773.

Zhang, Z. & Pugh, B.F. (2011). High-resolution genome-wide mapping of the primary structure of chromatin. Cell *144*, 175-186.

Zhang, Z., Wippo, C.J., Wal, M., Ward, E., Korber, P. & Pugh, B.F. (2011). A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. Science *332*, 977-980.

Zhao, M., Song, B., Pu, J., Wada, T., Reid, B., Tai, G., Wang, F., Guo, A., Walczysko, P., Gu, Y., et al. (2006). Electrical signals control wound healing through phosphatidylinositol-3-OH kinase-γ and PTEN. Nature *442*, 457-460.

Zhou, V.W., Goren, A. & Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nat. Rev. Genet. *12*, 7-18.