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.

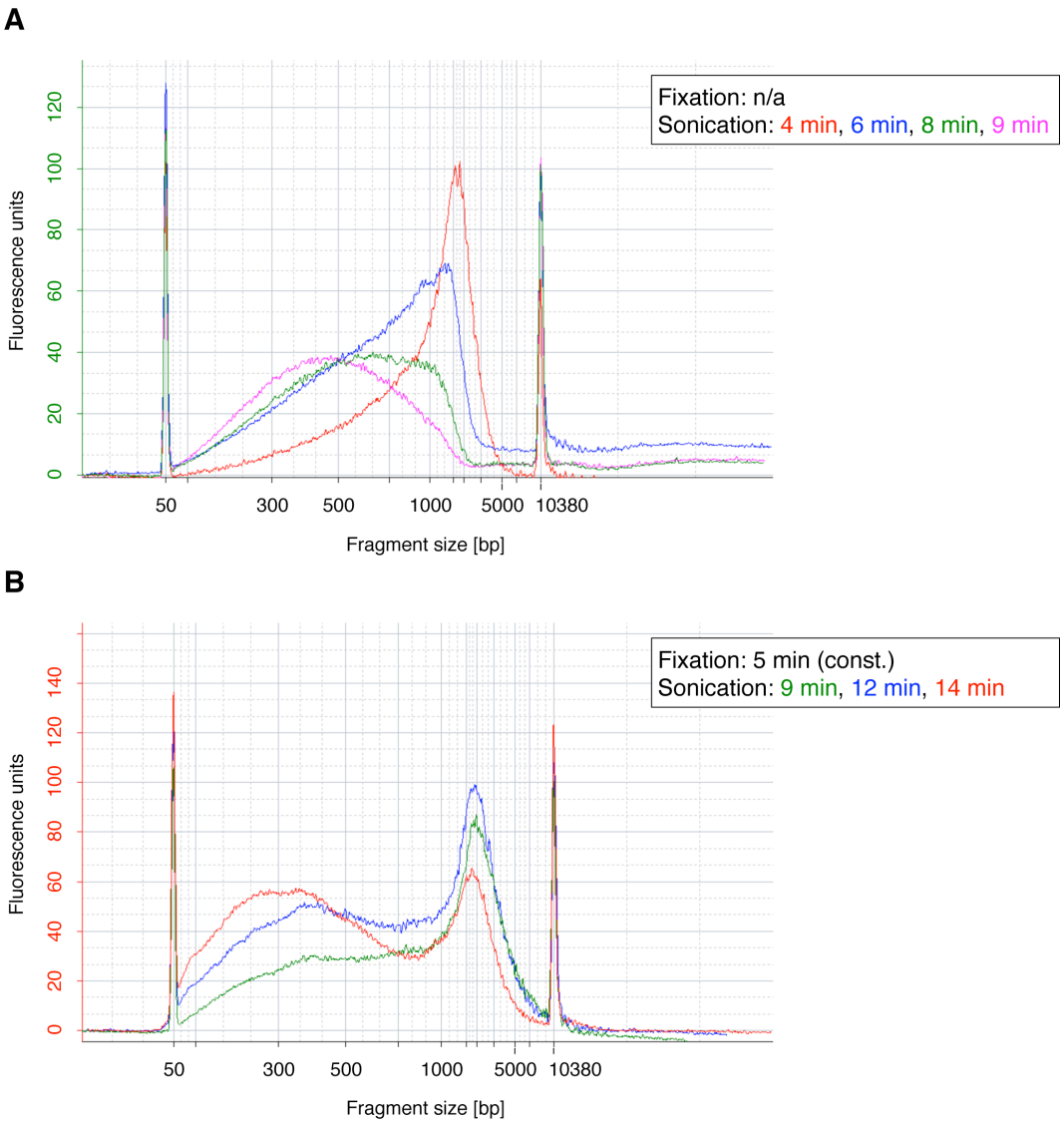
3.2. Optimisation of formaldehyde-assisted isolation of regulatory elements (FAIRE)

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.



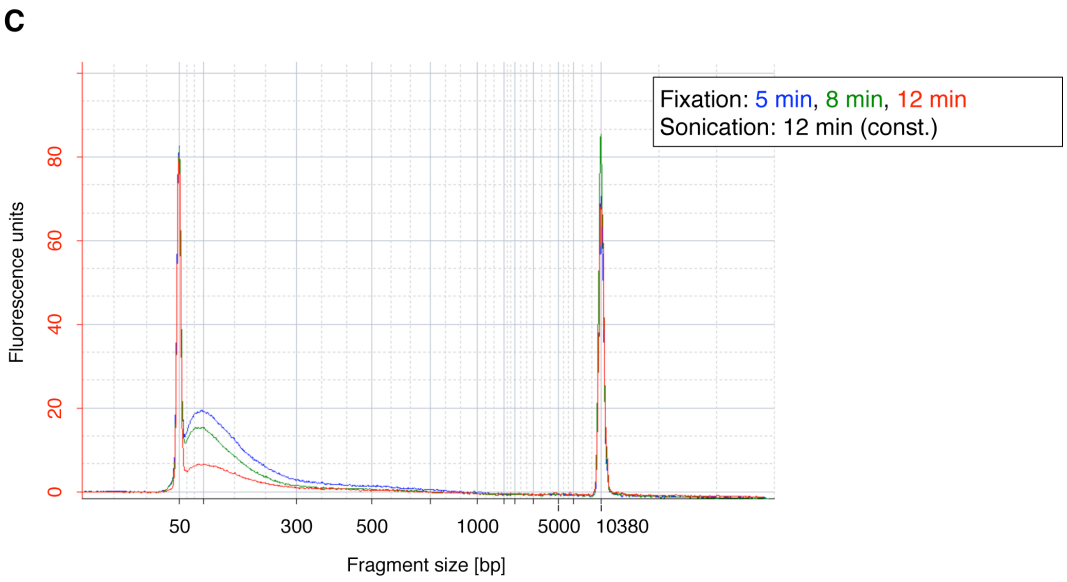


Figure 3-1. Electropherograms of time course experiments to monitor the distribution of DNA fragment lengths after sonication and formaldehyde fixation of chromatin. Shown are the electropherograms of (**A**) uncross-linked chromatin and (**B**) chromatin cross-linked with 1% formaldehyde for 5 min after 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 $P<5x10^{-8}$ (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**).

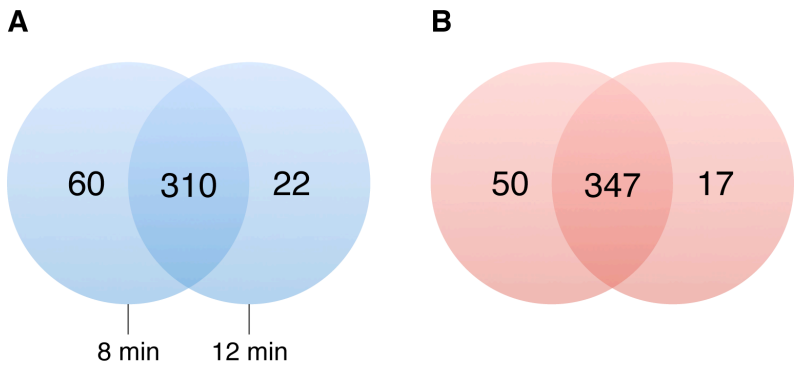


Figure 3-2. Number of peaks in FAIRE-chip data sets. FAIRE was performed with different formaldehyde cross-linking times (8 and 12 min), and enriched regions identified using the software NimbleScan v2.5 [Roche NimbleGen]. Shown is the number of peaks across the entire array content, either present in 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.

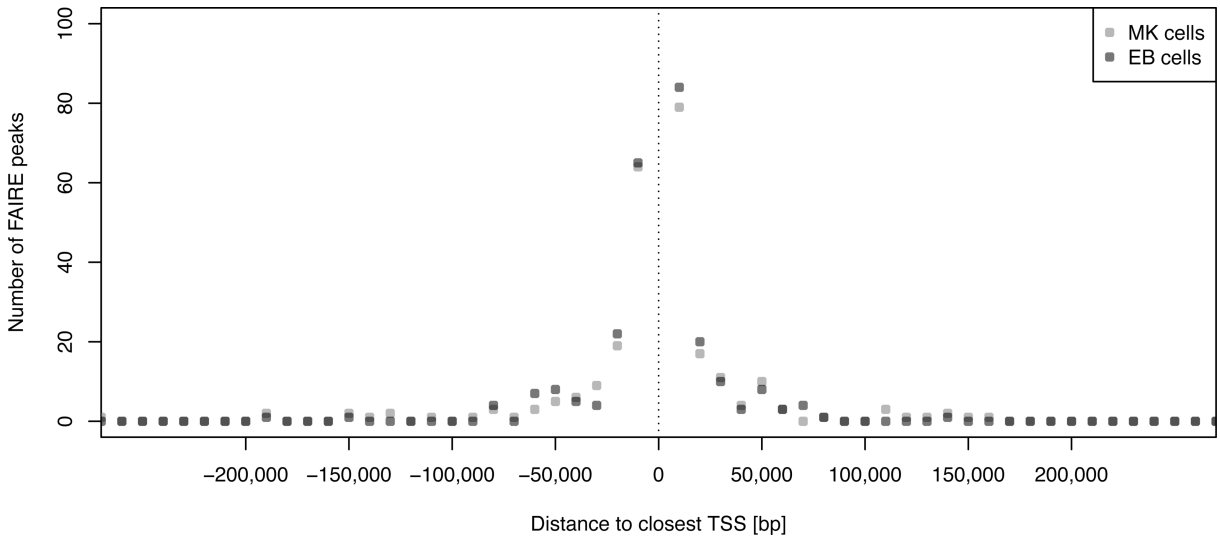


Figure 3-3. Location of open chromatin sites with respect to the closest TSS at the selected GWA loci.

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.

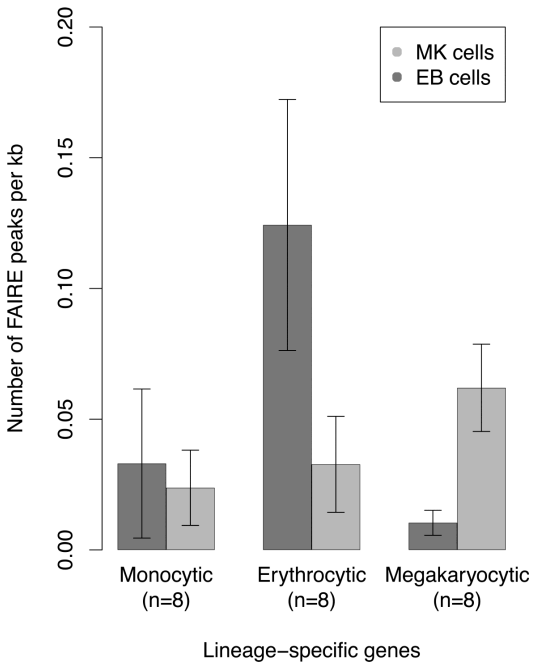


Figure 3-4. Average number of FAIRE peaks in lineage-specific genes in MK and EB cells. The number of open chromatin sites in lineage-specific genes (±2 kb) was averaged and normalised for the length of the gene. Error bars indicate standard error of the mean.

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.

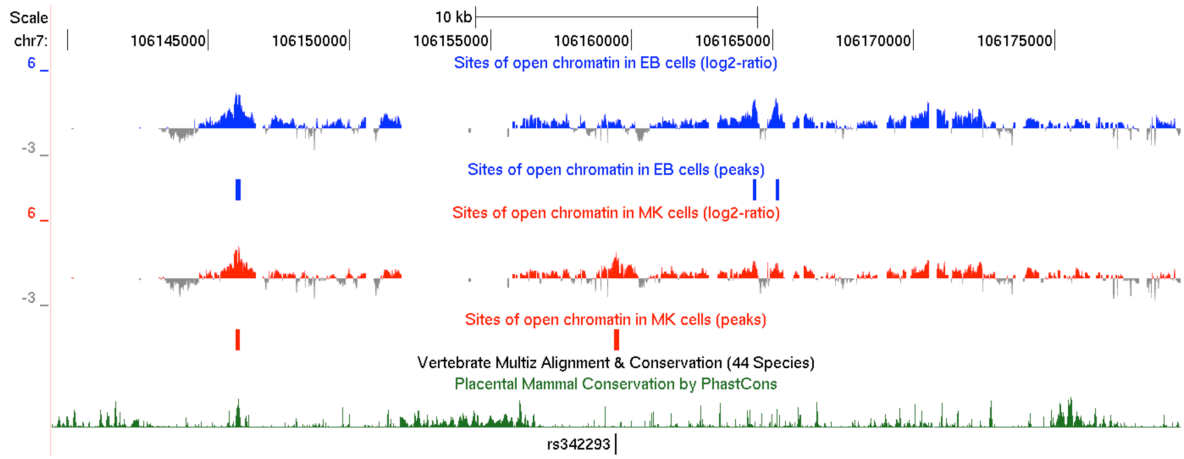
Cell type	Trait	Locus	SNP in open chromatin			GWA index SNP		
			ID	*MAF*	*Annotation*	*ID*	*r^2*	*Distance*
MK	MPV	*FLJ36031-PIK3CG*	rs342293	0.45	Intergenic	rs342293	1.00	(index SNP)
MK	MPV	*DNM3*	rs2038479	0.16	Intronic	rs10914144	0.94	10 kb
EB	MCV	*HBS1L-MYB*	rs7775698	0.22	Intergenic	rs9402686	0.85	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**);

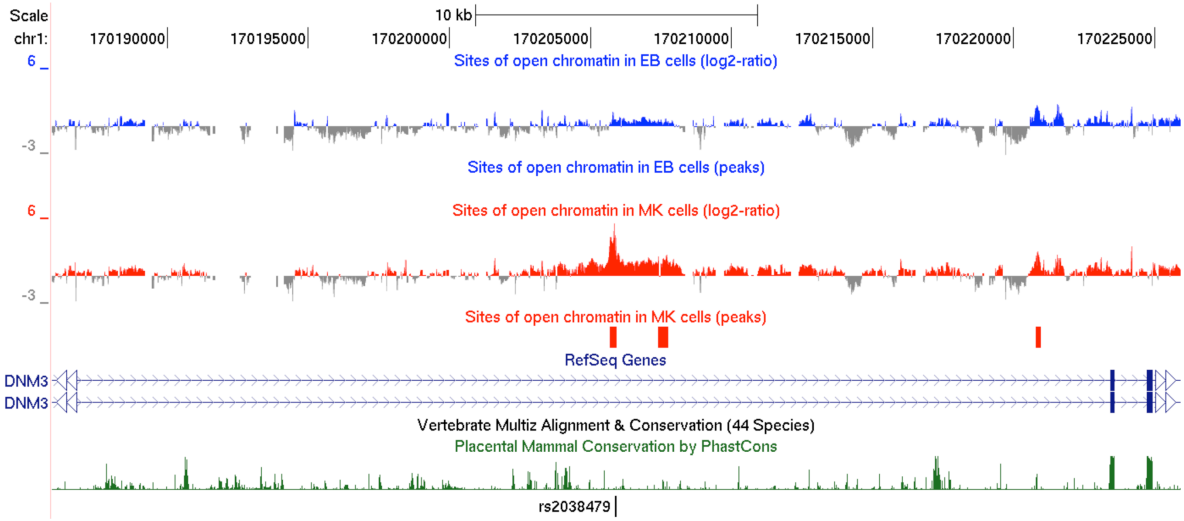
and systolic blood pressure (SBP) in an intronic region of *C10orf32* (*CYP17A1* gene cluster; **Figure 3-5 G**).

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**).

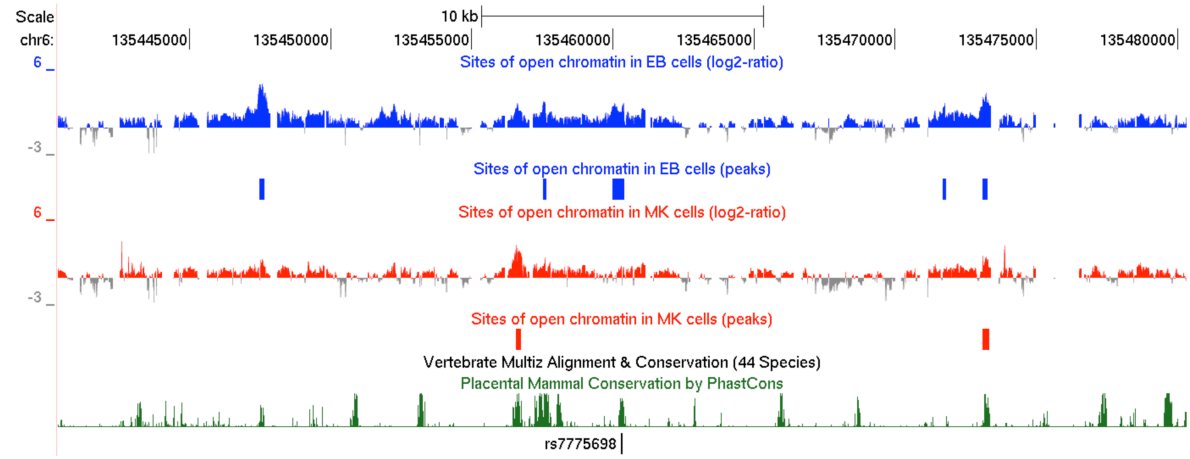
A



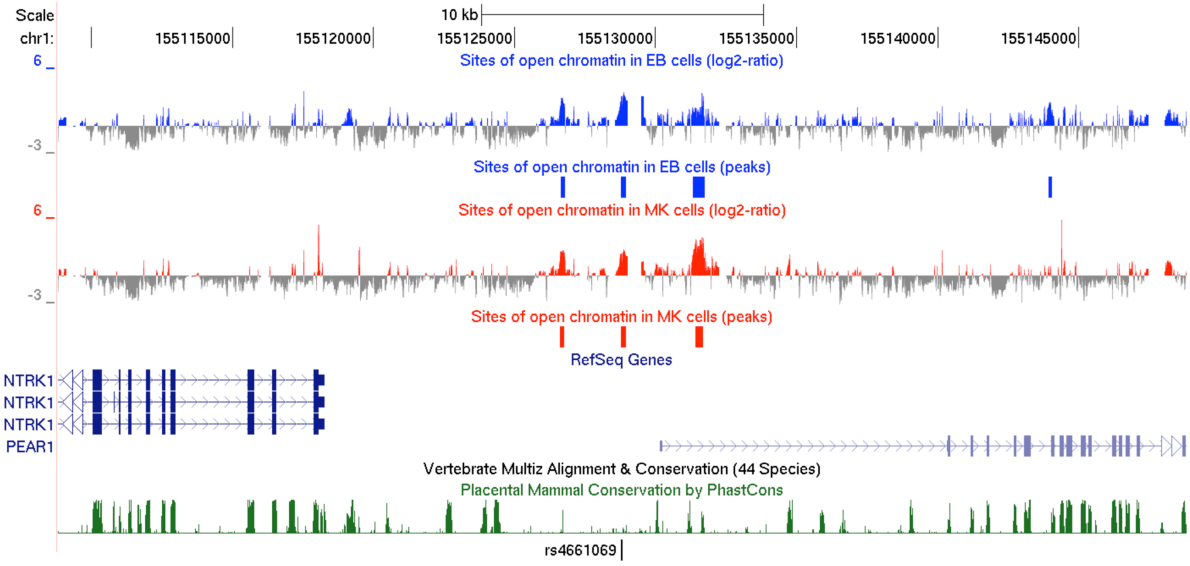
B



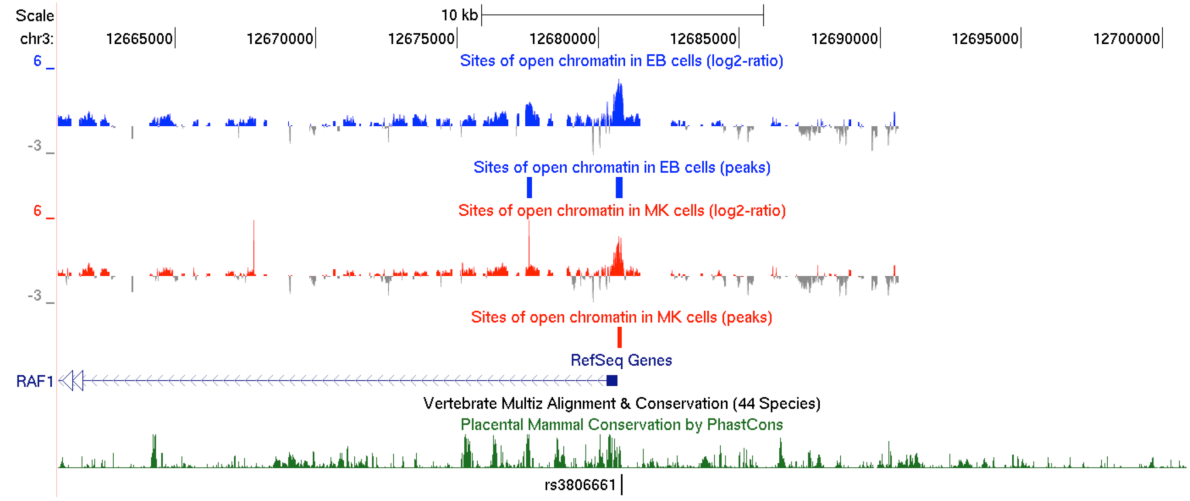
C


Scale
chr6:
10 kb
135445000
135450000
135455000
135460000
135465000
135470000
135475000
135480000
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
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
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

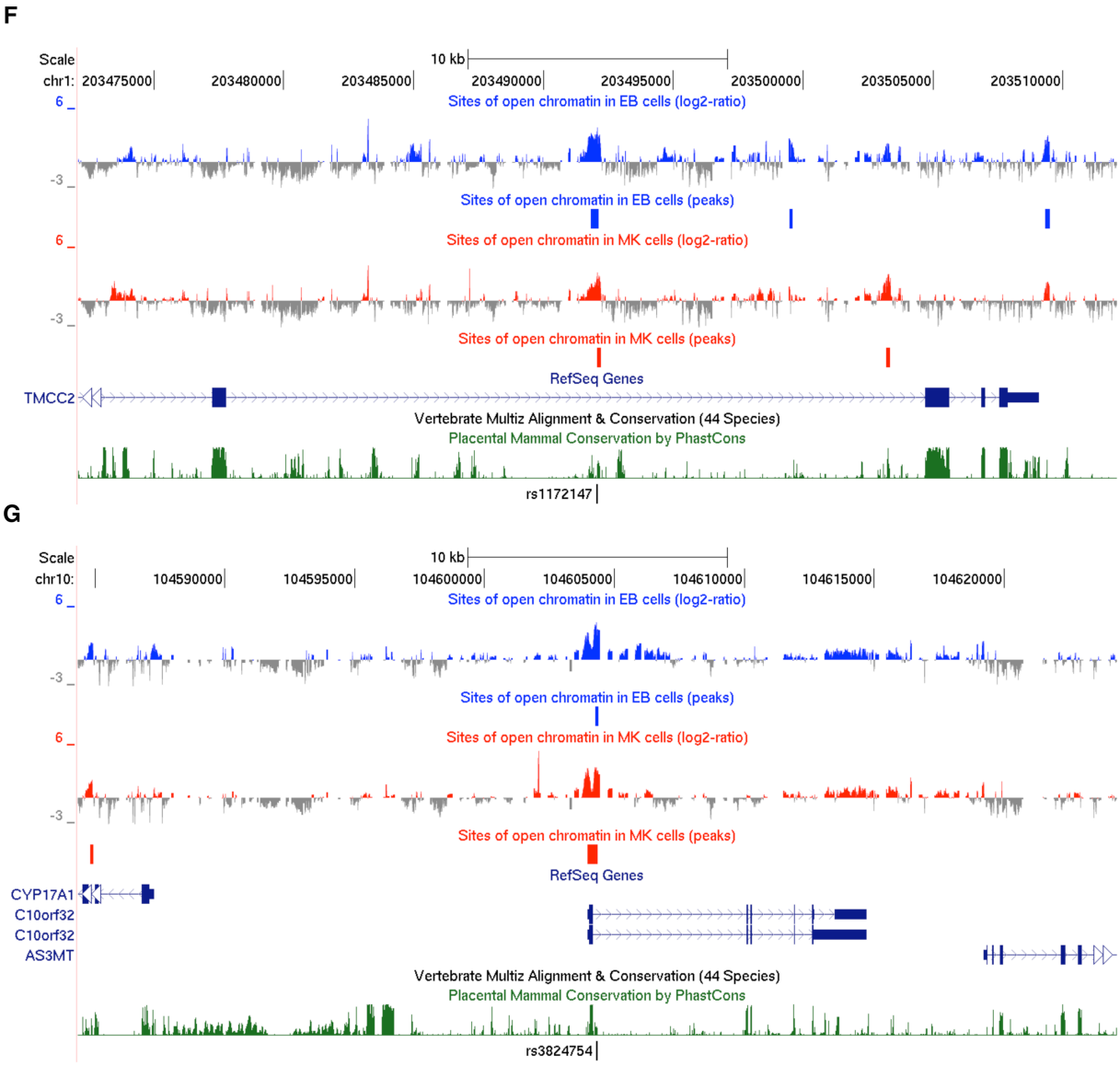


Figure 3-5. Open chromatin profiles at selected genetic loci in MK and EB cells displayed as UCSC Genome Browser custom tracks. (**A**) *FLJ36031-PIK3CG*; (**B**) *DNM3*; (**C**) *HBS1L-MYB*; (**D**) *PEAR1*; (**E**) *RAF1*; (**F**) *TMCC2*; (**G**) *CYP17A1-C10orf32*. Shown are the scaled $\log_2$-ratio 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.

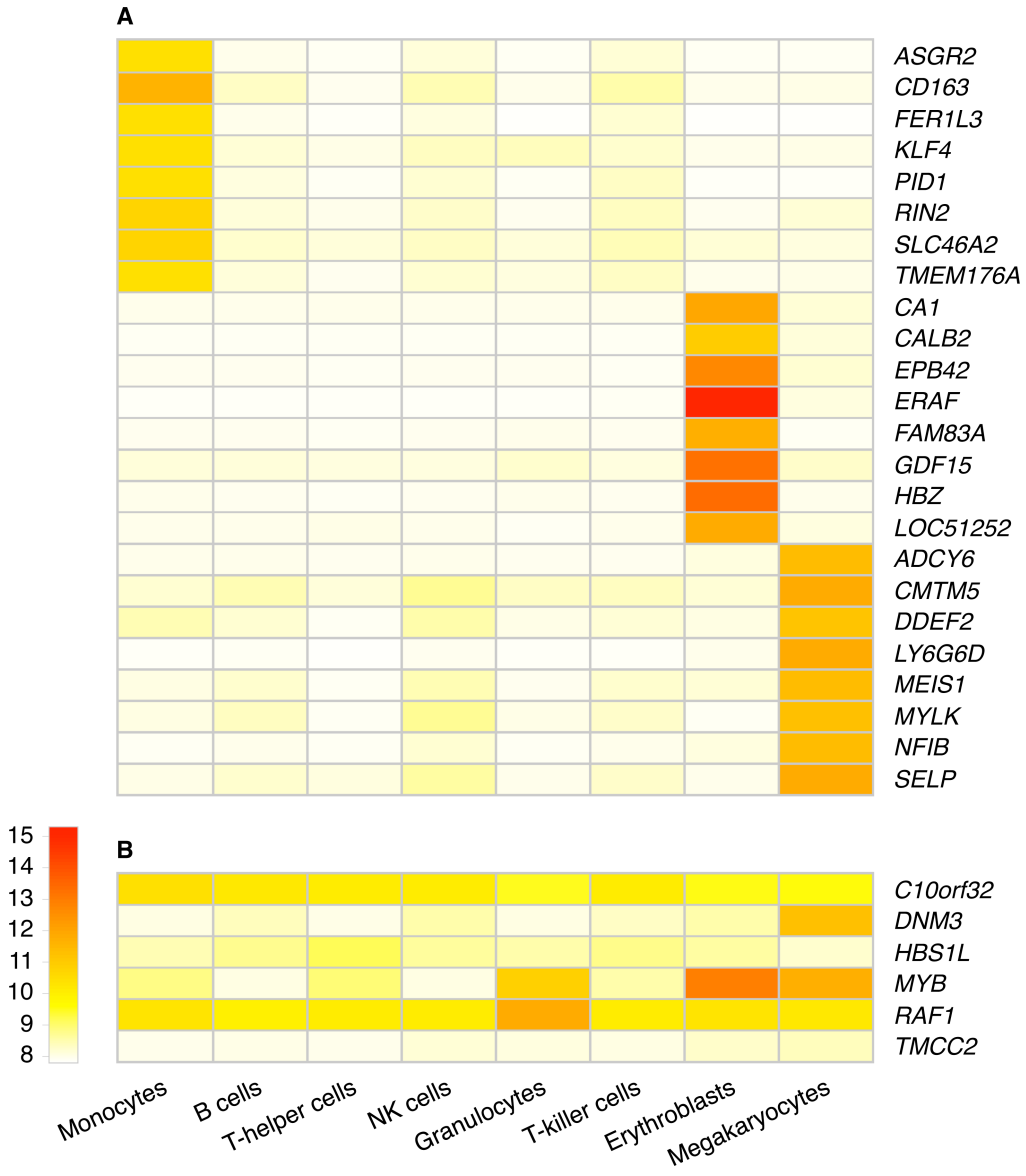


Figure 3-6. Gene expression profiles in differentiated human blood cells. Based on the HaemAtlas data (Watkins et al., 2009), the heat map shows normalised gene expression profiles of (**A**) lineage-specific reference genes selected for the custom DNA tiling array, and (**B**) genes that harbour cell 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 $I\leq 8.7$.

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.