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.

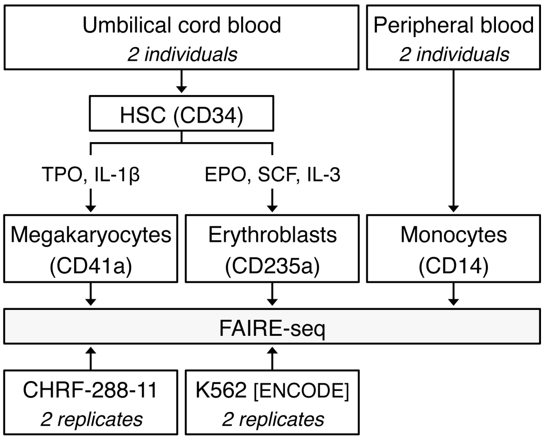
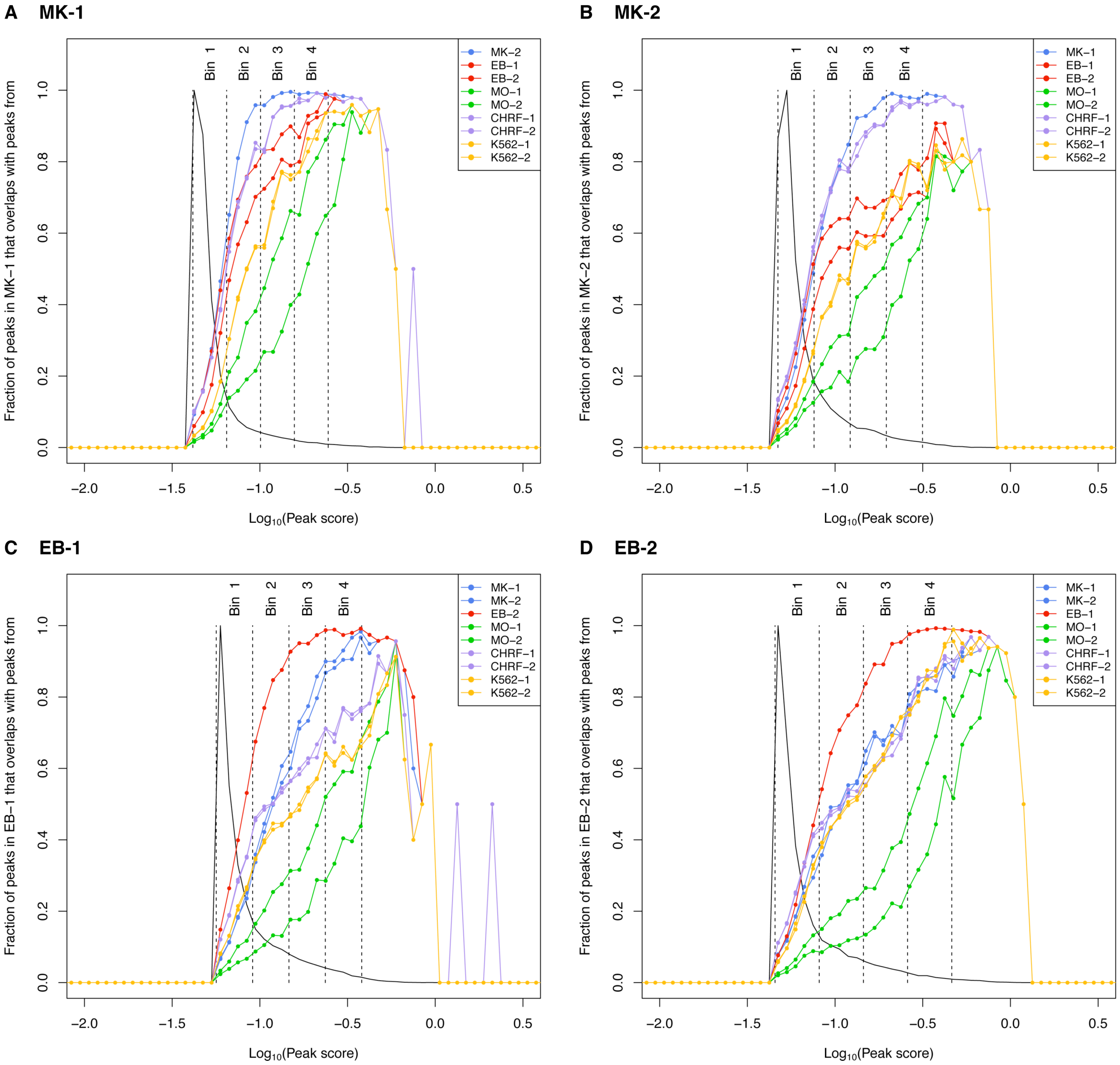
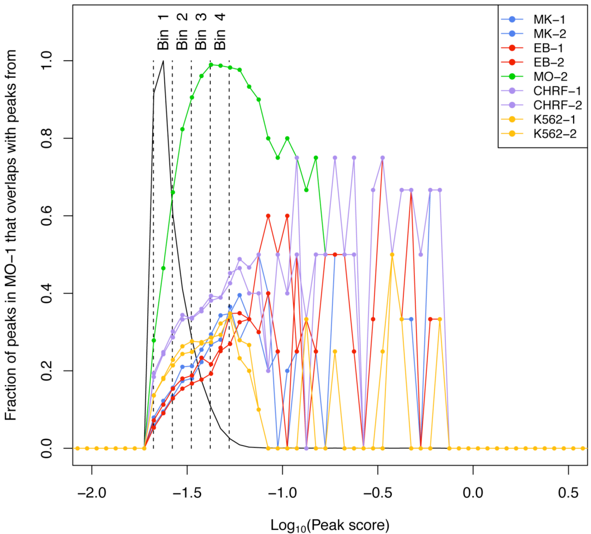


Figure 4-1. Overview of the study design.

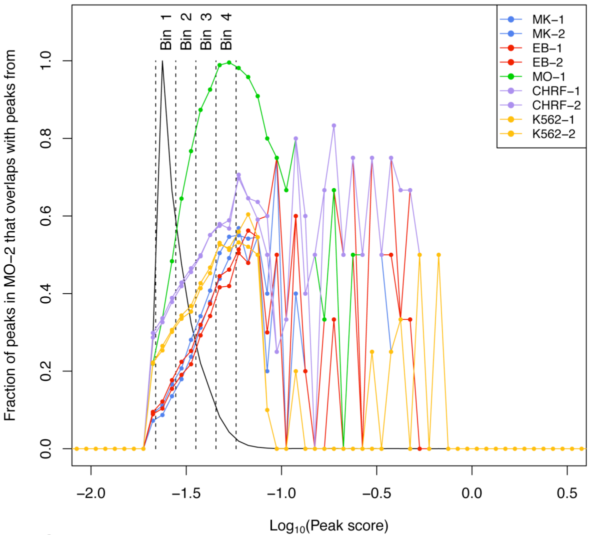
I determined FAIRE peaks marking NDRs using a Gaussian kernel density estimator implemented in the software F-Seq (Boyle, Guinney, et al., 2008). In order to reduce false positive peak calls, regions of collapsed repeats were removed, as described in Pickrell et al., 2011 (**Section 2.5**). For each 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**.



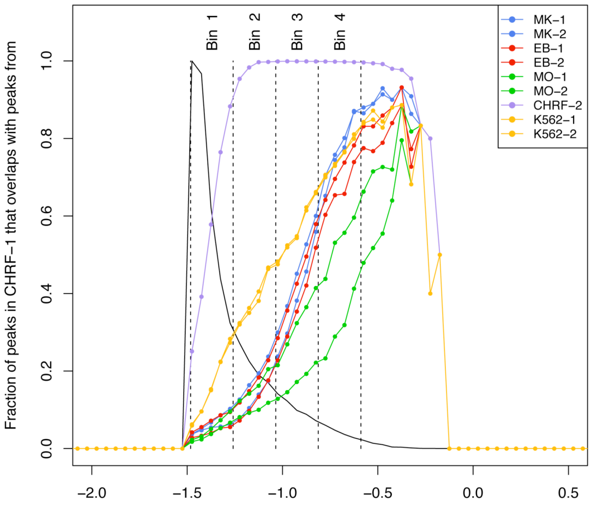
E MO-1



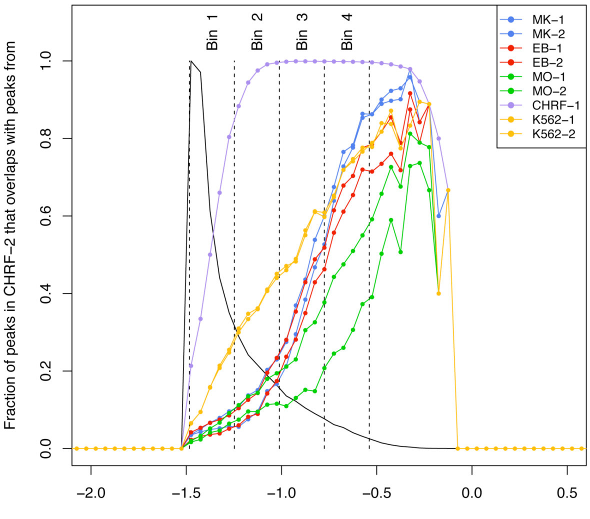
F MO-2



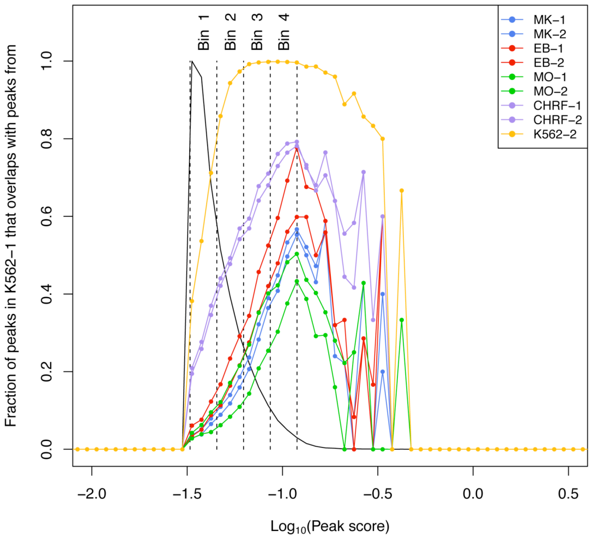
G CHRF-1



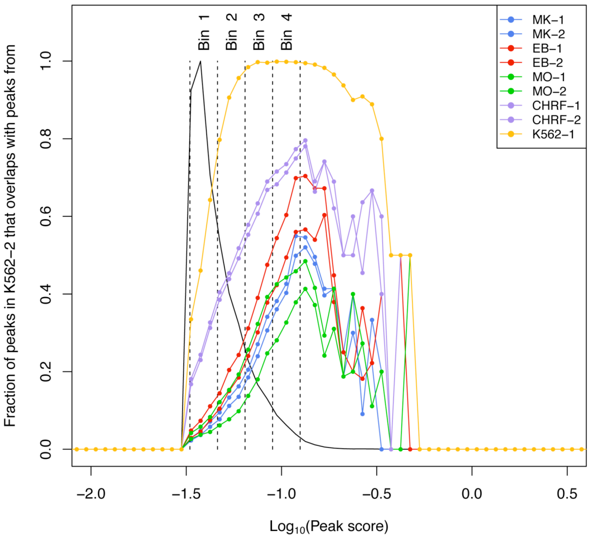
H CHRF-2



I K562-1



J K562-2



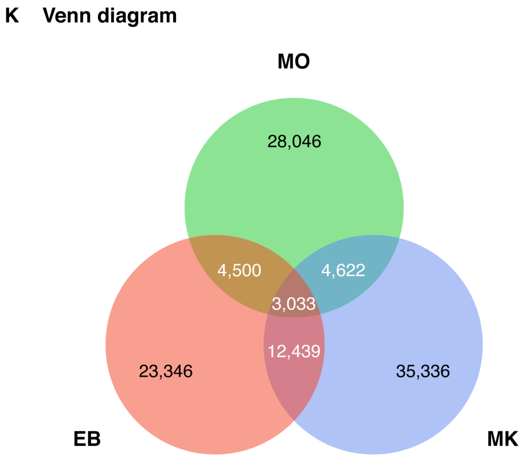


Figure 4-2. Overlap of NDRs across primary cells and their representative cell lines. The peak score distribution and 'intensity bins' for all studied cell types are shown (A–J). The top and bottom one percentile of the peak score distribution was excluded. The remaining peaks were stratified according 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.

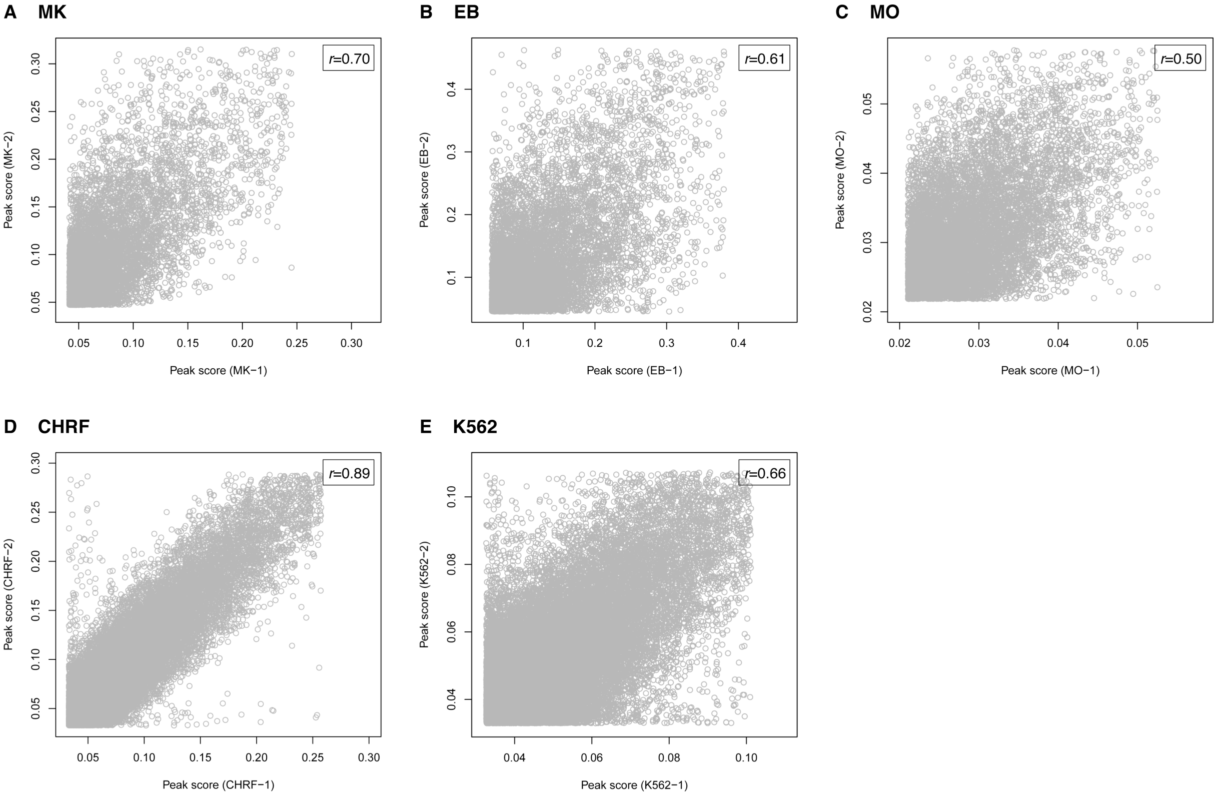


Figure 4-3. Pearson's correlation of peak score between independent FAIRE experiments. Correlation coefficients were calculated for scores of overlapping peaks between biological replicates (cell lines) or individuals (primary cells). Peaks unique to replicates/individuals were not included in these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

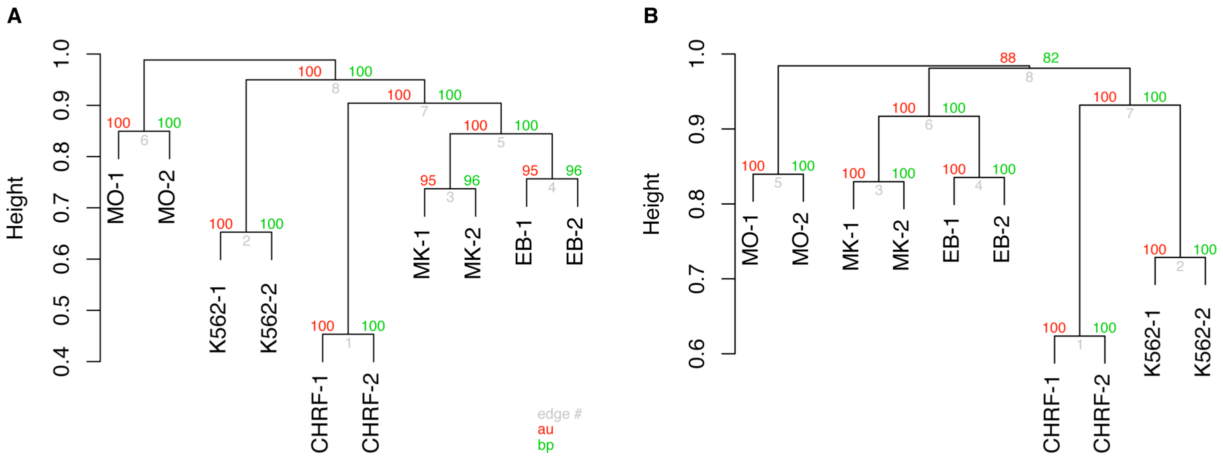


Figure 4-4. Characterisation of open chromatin profiles of primary human megakaryocytes (MKs), erythroblasts (EBs) and monocytes (MOs). Dendrogram of the hierarchical clustering of the overlap of FAIRE-derived NDRs across primary cells and immortalised cell lines. For assessing the relationship of the 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**).

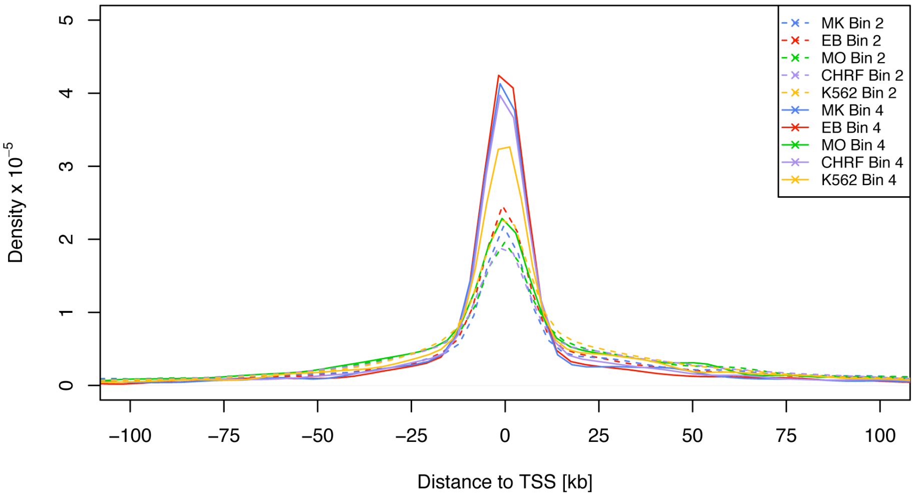


Figure 4-5. Distance of NDRs to the closest TSSs. The genomic distances between FAIRE peaks and transcription start sites were exported from GREAT (McLean et al., 2010). The density graph was plotted with a bandwidth of 5,000.

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.

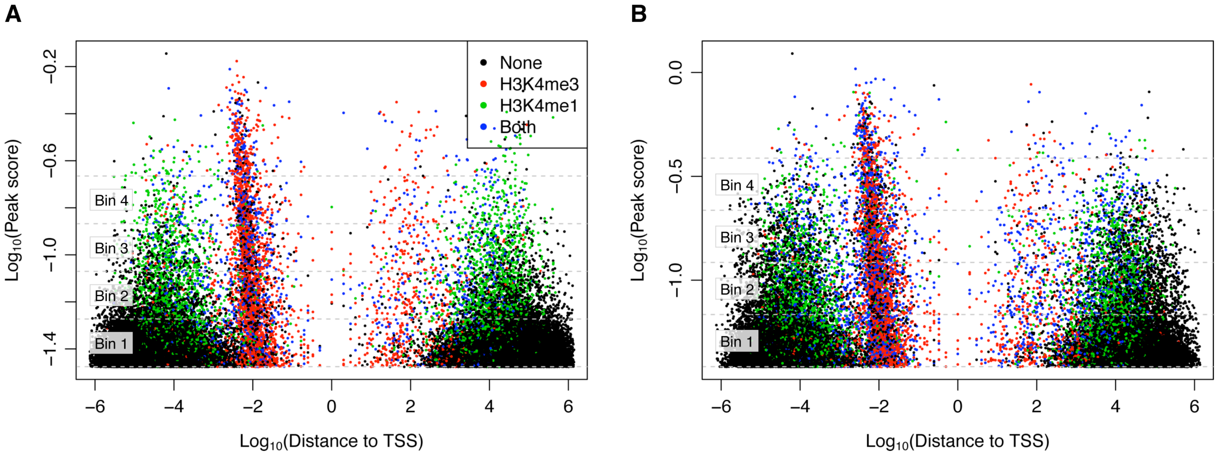


Figure 4-6. Overlap of H3K4me3 (promoter) and H3K4me1 (enhancer) histone marks with NDRs identified in (A) MKs and (B) EBs with respect to NDR score and distance to the closest TSS. The peak bins are indicated with a dashed grey line. In MKs, we identified 79,049 and 17,402 regions of enrichment of H3K4me1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

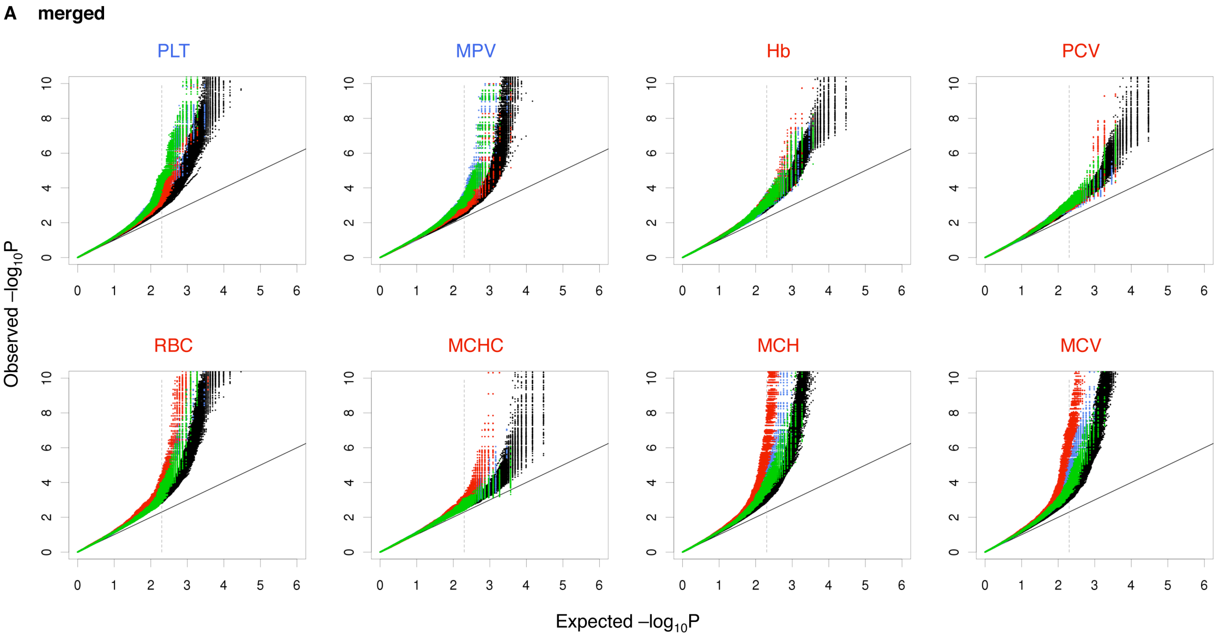
We then examined if cell type-restricted NDRs mark lineage-specific elements involved in regulation of expression of genes relevant to blood cell lineage commitment. We assessed the expression levels of the single closest gene to each lineage-specific NDR, interrogated over several time points during *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.

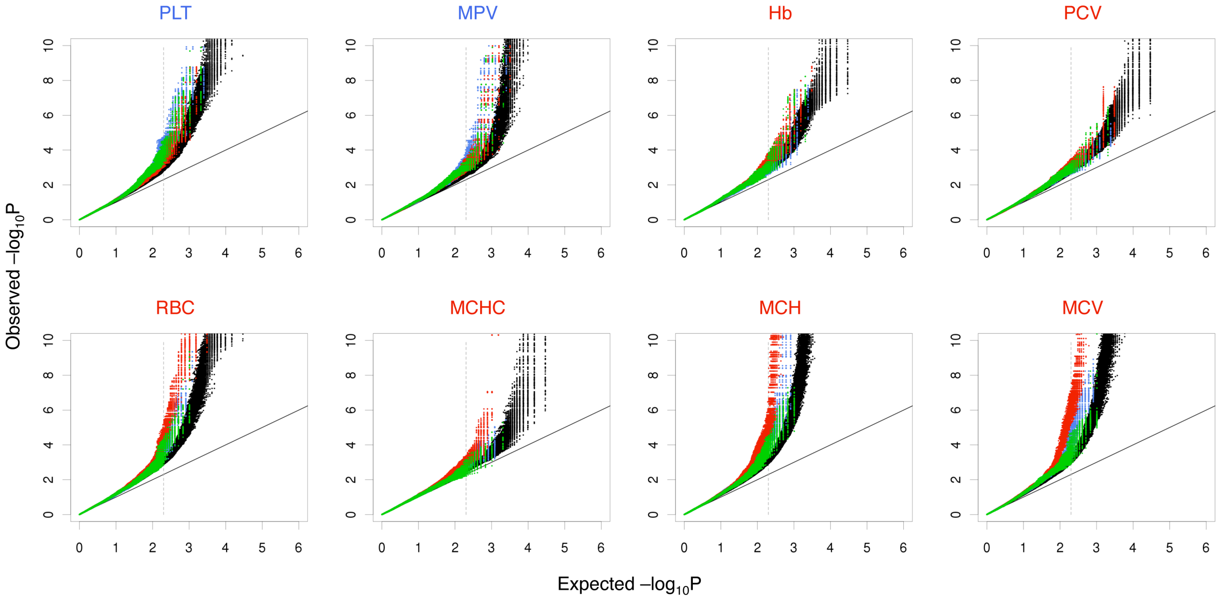
Cell type(s) containing NDRs	Number genes ↓	Number genes ↑	*P*-val (two-tailed chi-squared test)	Fold enrichment
(A) Differentiation into megakaryocytes				
EB	3,490	3,909	7.64x10^{-02}	0.98
MK	2,777	4,282	2.01x10^{-30}	1.13
MK/EB	1,294	1,631	3.90x10^{-02}	1.04
MK/MO	274	457	2.65x10^{-06}	1.16
MK/MO/EB	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.



B intersected



C cell type-specific

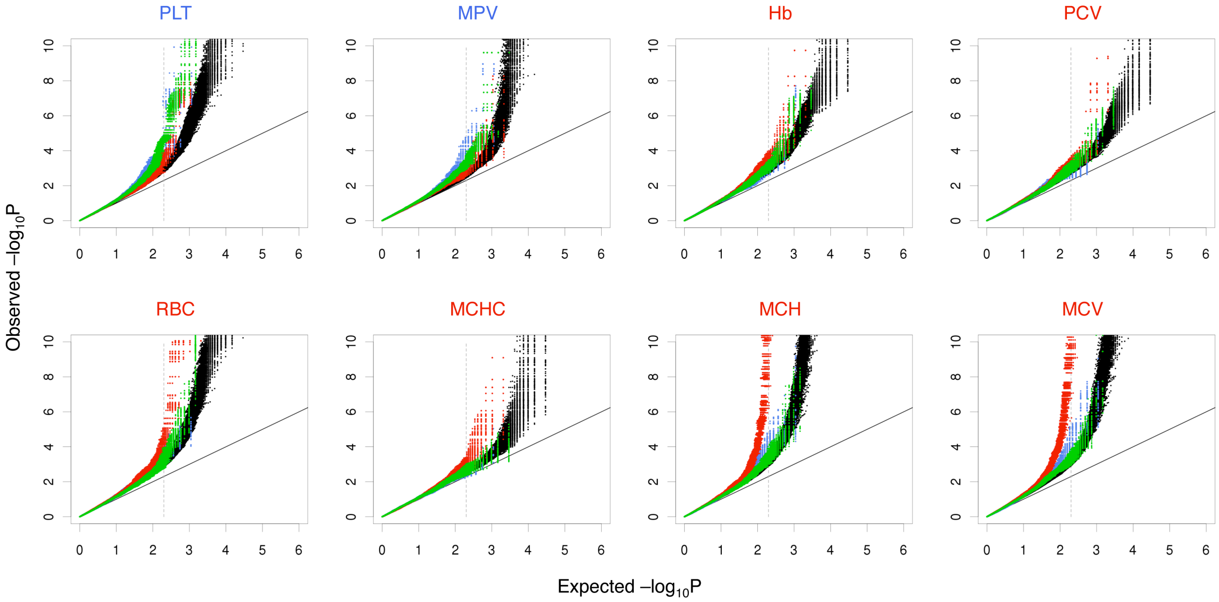


Figure 4-7. Bootstrapped quantile-quantile distributions. Data points in black represent the distribution of *P*-values for all 2.6x10^6 imputed SNPs. Enrichment values in the primary cell types shown in **Figure 4-8** are relative to this distribution. Enrichments are shown for three classes of NDRs: (**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.

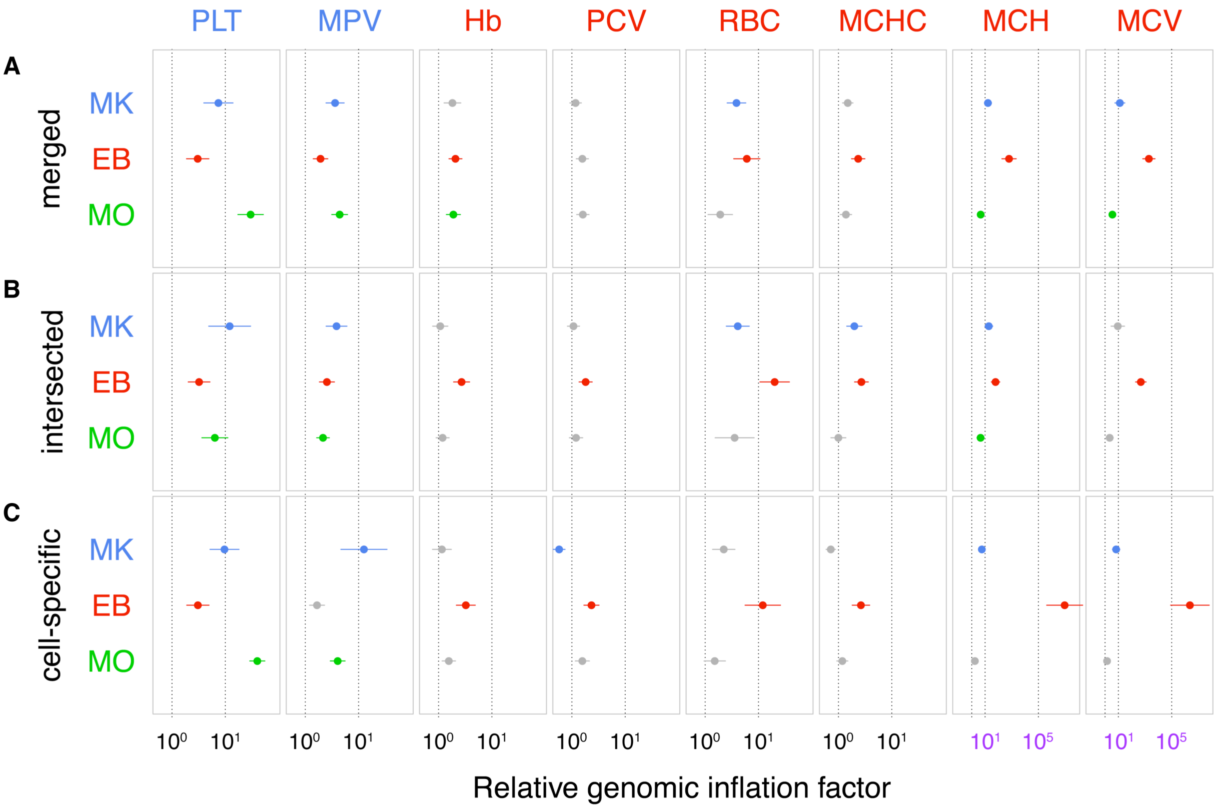


Figure 4-8. Enrichment of associations with platelet and erythrocyte phenotypes in NDRs across quantitative haematological traits, cell types and NDR classes. For each trait, the enrichment of associations (genomic inflation) in MKs, EBs and MOs are indicated. Enrichments are shown for three classes 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 2.6×10^6 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, $P=4.62\text{x}10^{-2}$–$9.94\text{x}10^{-2}$, Benjamini-Hochberg corrected for multiple testing; n=7 genes) and 'cell-to-cell signalling and interaction' ($P=4.62\text{x}10^{-2}$–$9.94\text{x}10^{-2}$; 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 ($P<5\text{x}10^{-8}$ 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 ($P<5\text{x}10^{-6}$) 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 ($P<5\text{x}10^{-6}$) 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.

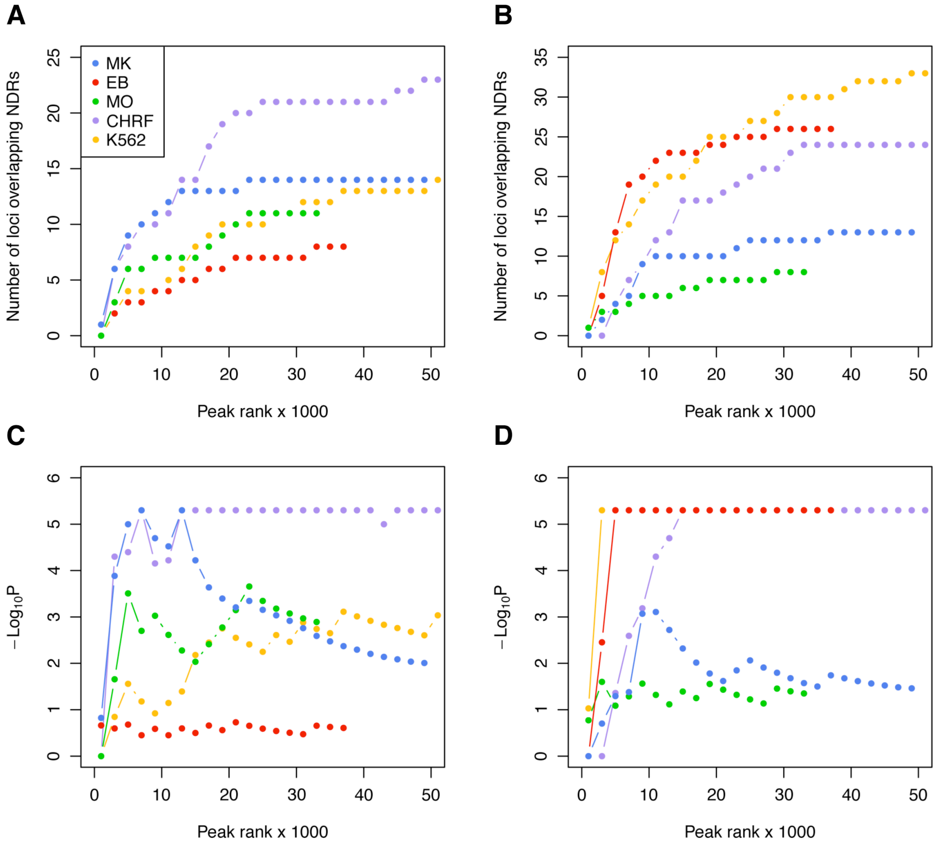


Figure 4-9. Genome-wide significant signals associated with platelet and erythrocyte phenotypes at sites of open chromatin in primary cells and immortalised cell lines. Number of GWA loci harbouring (A) platelet- and (B) erythrocyte trait-associated SNPs in NDRs in MKs, EBs, MOs, CHRF-288-11 and K562 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.

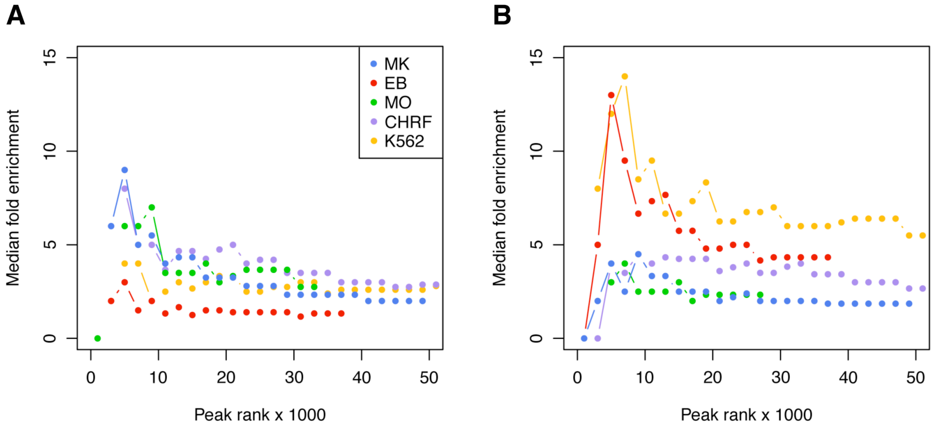


Figure 4-10. Fold enrichment of the number of loci with at least one overlap with an NDR compared to the median of 100,000 random sets of loci. The enrichment is shown at (A) platelet and (B) erythrocyte loci. The extent of overlap with NDRs in each cell type at the association loci was assessed by comparison 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.

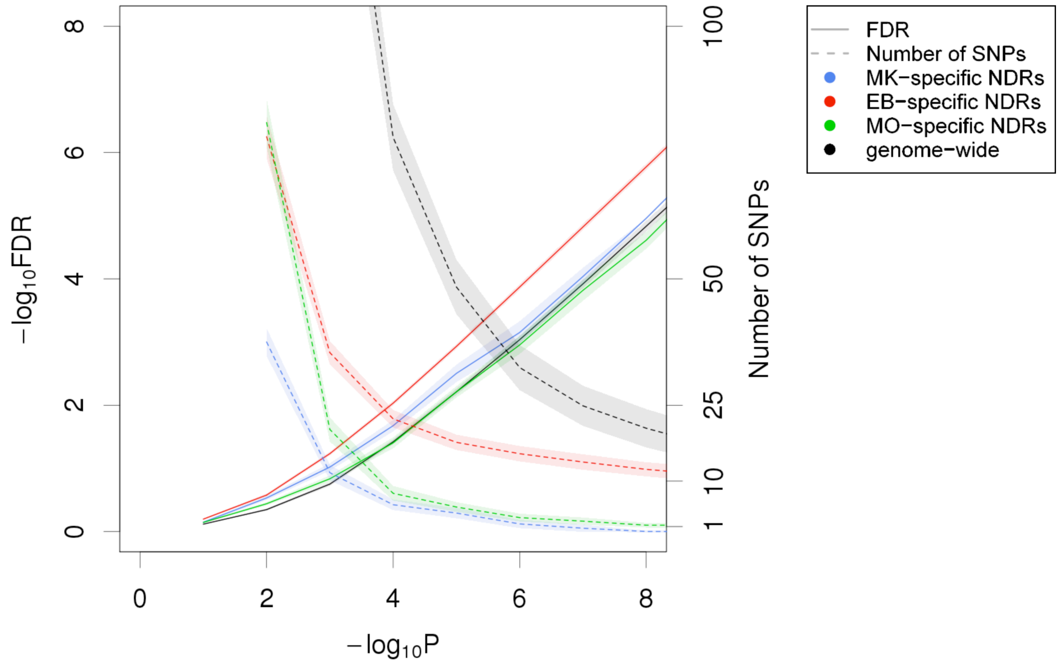
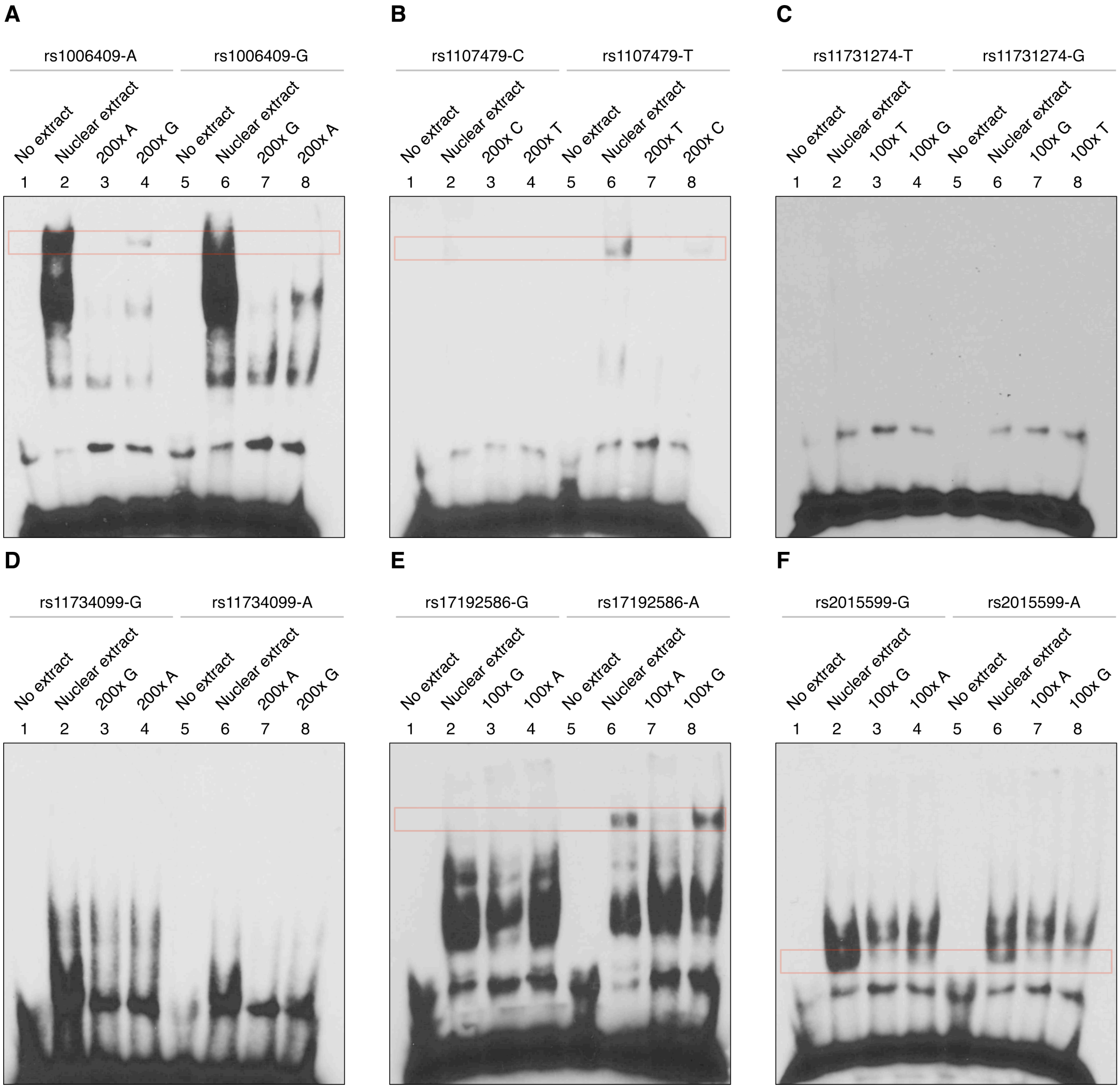


Figure 4-11. Estimated false discovery rates (FDRs) for mean red cell volume-associated SNPs in cell type-specific NDRs. The FDR for SNPs (solid lines) in cell type-specific NDRs (left y-axis) was estimated as a function of the significance level (x-axis). The number of SNPs (dashed lines) associated 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**).



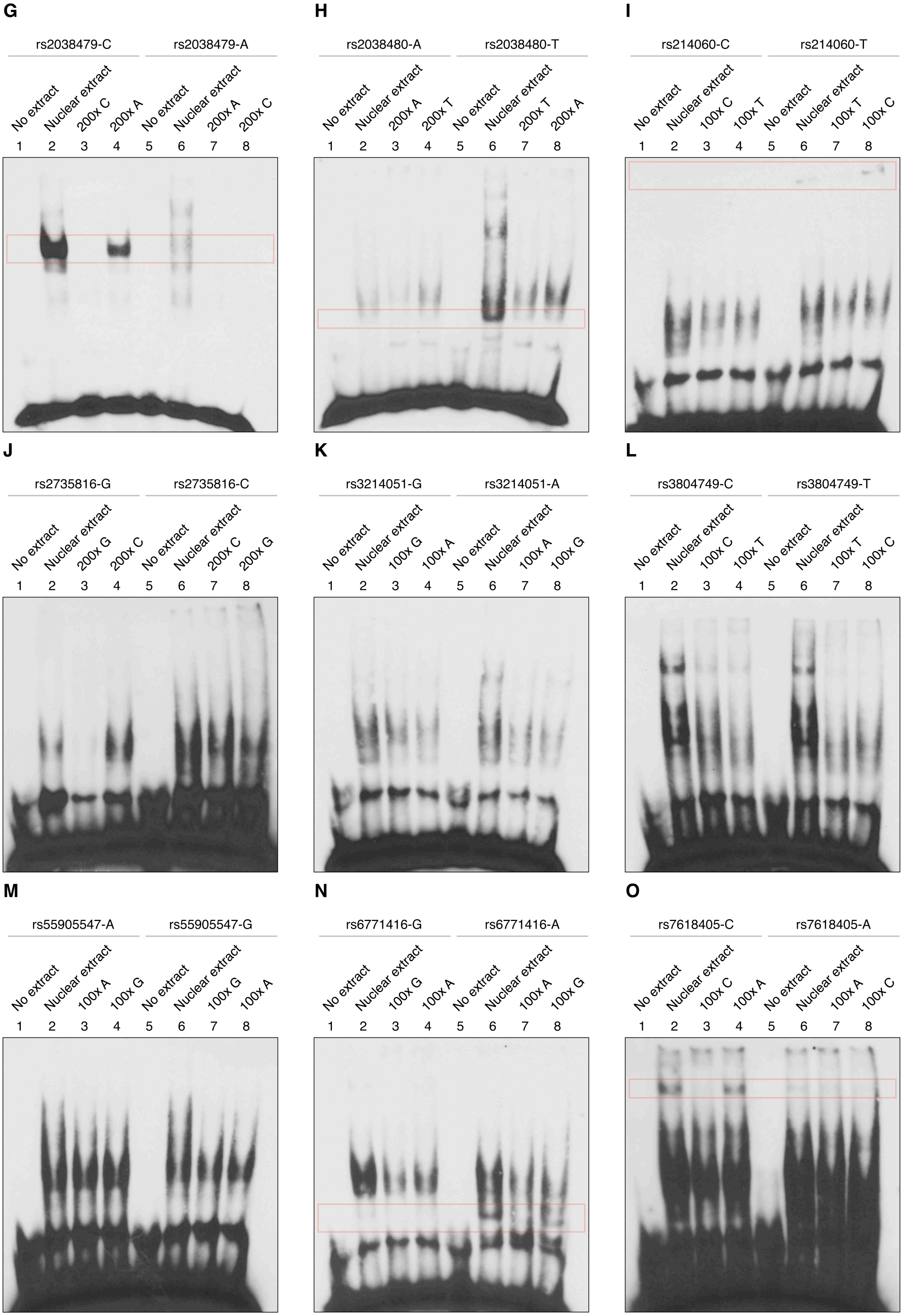

G
rs2038479-C
rs2038479-A
No extract
Nuclear extract
200x C
200x A
No extract
Nuclear extract
200x A
200x C
1 2 3 4 5 6 7 8
H
rs2038480-A
rs2038480-T
No extract
Nuclear extract
200x A
200x T
No extract
Nuclear extract
200x T
200x A
1 2 3 4 5 6 7 8
I
rs214060-C
rs214060-T
No extract
Nuclear extract
100x 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.

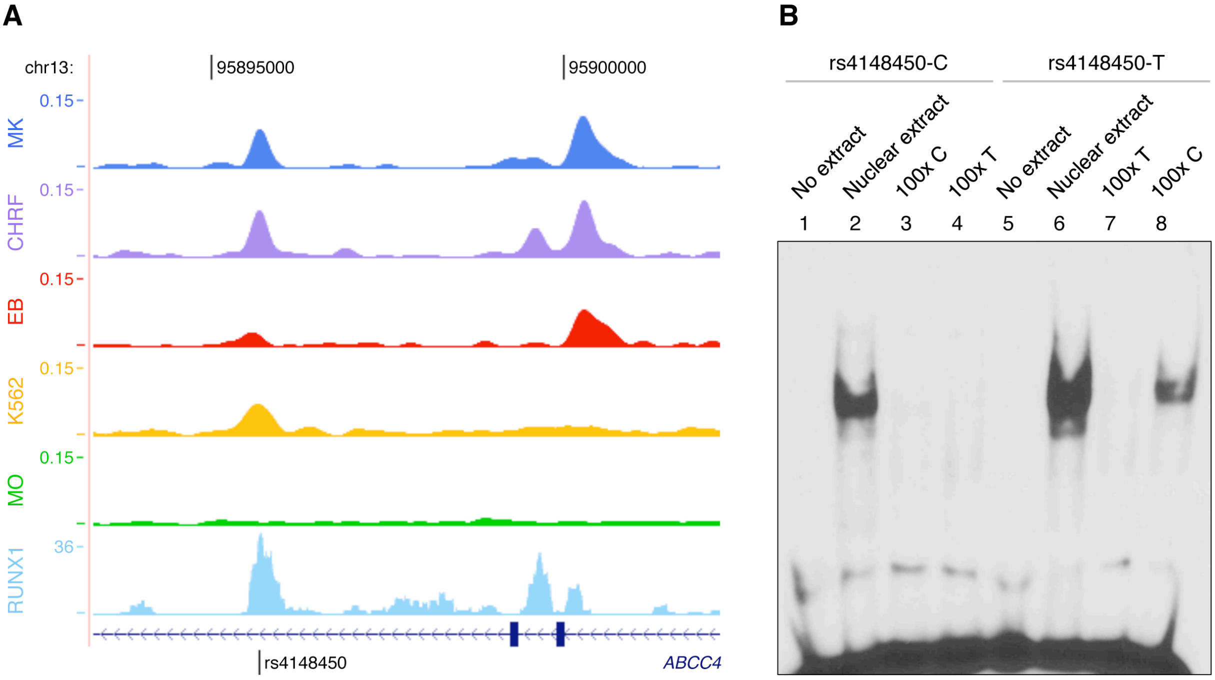


Figure 4-13. Functional follow-up of the *ABCC4* platelet count locus. (A) Coverage profiles of FAIRE-seq data in MKs, EBs, MOs, CHRF-288-11 and K562 cells. An open chromatin region found only in MKs and CHRF-288-11 cells contained the platelet count-associated SNP rs4148450 (r^2=1.0 with GWA lead SNP 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.