

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



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.

1. Introduction	1
1.1. Human genetic variation	2
1.2. Genetics of complex traits in humans	4
1.2.1. Approaches to genetic mapping	4
1.2.2. Allelic spectrum of genetic variants and missing heritability	5
1.2.3. Consensus and challenges in genome-wide association studies	6
1.3. Primary structure of chromatin.....	8
1.4. Determinants of chromatin accessibility	8
1.4.1. DNA sequence-dependent nucleosome positioning.....	9
1.4.2. ATP-dependent chromatin remodelling	9
1.4.3. Histone variants and modifications	10
1.4.4. Competitive protein binding.....	12
1.5. Transcriptional regulation	13
1.6. Transcriptional regulatory elements.....	14
1.6.1. Promoters	15
1.6.2. Enhancers.....	17
1.6.3. Silencers.....	18
1.6.4. Insulators.....	18
1.7. Methods for mapping gene regulatory elements	19
1.7.1. Endonuclease digestion	20
1.7.2. DNA methylation footprinting.....	21
1.7.3. Formaldehyde-assisted isolation of regulatory elements (FAIRE)	21
1.7.4. Chromatin immunoprecipitation (ChIP)	21
1.8. Gene regulatory elements in human disease	23
1.9. Haematopoietic system and genetics of haematological traits	26
1.10. Thesis aims and objectives	27
2. Materials and methods	29
2.1. Culture and preparation of cell lines.....	30
2.2. Isolation, culture and preparation of primary cells.....	31
2.3. Formaldehyde-assisted isolation of regulatory elements (FAIRE).....	34
2.4. Detection and analysis using DNA tiling microarrays.....	36

2.5.	Detection and analysis using high-throughput next-generation sequencing technology	39
2.6.	Annotation of NDRs and statistical analyses.....	42
2.7.	Gene expression analysis during <i>in vitro</i> differentiation of cord blood-derived HPCs	44
2.8.	H3K4me1 and H3K4me3 ChIP-seq.....	44
2.9.	Sanger sequencing of selected NDRs.....	44
2.10.	Transcription factor binding site prediction	45
2.11.	Electrophoretic mobility shift assay (EMSA)	46
2.12.	Expression QTL analysis.....	49
2.13.	Whole-genome gene expression profiling of <i>Pik3cg</i> ^{-/-} mice	50
2.14.	Protein-protein interaction network	52
2.15.	Exome sequencing of individuals with TAR syndrome	52
2.16.	Sanger sequencing of the <i>RBM8A</i> locus	53
2.17.	Genotyping of the 5'-UTR and intronic SNPs at the <i>RBM8A</i> locus.....	54
2.18.	Sequencing of megakaryocyte RNA.....	54
2.19.	<i>RBM8A</i> promoter activity by luciferase reporter assay.....	55
2.20.	Y14 protein expression analysis in platelet extracts.....	55
3.	Maps of open chromatin guide the functional follow-up of genome-wide association signals at haematological trait loci	57
3.1.	Introduction	58
3.2.	Optimisation of formaldehyde-assisted isolation of regulatory elements (FAIRE).....	58
3.3.	Design of an oligonucleotide tiling microarray for NDR mapping.....	60
3.4.	Open chromatin profiles in a megakaryocytic and an erythroblastoid cell line	62
3.5.	Characterisation of open chromatin regions in relation to gene annotations and cell types	64
3.6.	Seven sequence variants associated with haematological and cardiovascular-related quantitative traits are located in NDRs	66
3.7.	Discussion	70
4.	Maps of open chromatin highlight cell type-specific patterns of regulatory sequence variation at haematological trait loci.....	72
4.1.	Introduction	73
4.2.	Functional characterisation of open chromatin profiles in human myeloid cells.....	73
4.3.	Cell type-specific enrichment of haematological trait-associated SNPs in NDRs	81
4.4.	Identification of candidate functional SNPs at platelet QTLs	87
4.5.	Discussion	92

5. Functional follow-up of the platelet volume and function locus at chromosome 7q22.3	95
5.1. Introduction	96
5.2. Identification of rs342293 as the only likely putative functional candidate at the 7q22.3 locus	96
5.3. The alleles of rs342293 differentially bind the transcription factor EVI1	100
5.4. The SNP rs342293 is associated with <i>PIK3CG</i> transcript levels in platelets and macrophages.....	103
5.5. Gene expression profiles in whole blood of <i>Pik3cg</i> ^{-/-} mice.....	104
5.6. Canonical pathway enrichment analysis and protein-protein interaction network	105
5.7. Discussion	106
6. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in <i>RBM8A</i> causes TAR syndrome.....	110
6.1. Introduction	111
6.1.1. Exome sequencing as a tool for gene discovery in rare diseases	111
6.1.2. Genetics of thrombocytopenia with absent radii (TAR) syndrome.....	112
6.2. Most TAR cases have a low-frequency regulatory variant and a rare null allele at the <i>RBM8A</i> locus	113
6.3. The effect of the regulatory SNPs on transcription factor binding, <i>RBM8A</i> promoter activity and protein expression in platelets	117
6.4. Discussion	121
7. Conclusions and outlook.....	124
7.1. Identifying candidate functional variants using maps of open chromatin	125
7.2. Translating candidate functional variants using gene regulatory network approaches.....	127
8. Appendix	130
References.....	166

LIST OF FIGURES.

1. Introduction	1
Figure 1-1. Human genetic variation, haplotype structure and linkage disequilibrium	3
Figure 1-2. Spectrum of genetic variants with respect to allele frequency and genetic effect size	5
Figure 1-3. Chromatin structure and transcriptional regulatory elements	14
Figure 1-4. Experimental methods for mapping gene regulatory elements	22
Figure 1-5. Integration of multiple data sets for annotating GWA loci	24
Figure 1-6. Simplified scheme of lineage determination in human haematopoietic hierarchies	26
Figure 1-7. Translation of genetic signals into molecular mechanism and biological understanding	28
2. Materials and methods	29
Figure 2-1. Characterisation of primary human MOs, MKs and EBs	33
3. Maps of open chromatin guide the functional follow-up of genome-wide association signals at haematological trait loci	57
Figure 3-1. Electropherograms of time course experiments to monitor the distribution of DNA fragment lengths after sonication and formaldehyde fixation of chromatin	59
Figure 3-2. Number of peaks in FAIRE-chip data sets.....	62
Figure 3-3. Location of open chromatin sites with respect to the closest TSS at the selected GWA loci.....	65
Figure 3-4. Average number of FAIRE peaks in lineage-specific genes in MK and EB cells.....	65
Figure 3-5. Open chromatin profiles at selected genetic loci in MK and EB cells displayed as UCSC Genome Browser custom tracks.....	67
Figure 3-6. Gene expression profiles in differentiated human blood cells	70
4. Maps of open chromatin highlight cell type-specific patterns of regulatory sequence variation at haematological trait loci	72
Figure 4-1. Overview of the study design.....	73
Figure 4-2. Overlap of NDRs across primary cells and their representative cell lines	74
Figure 4-3. Pearson's correlation of peak score between independent FAIRE experiments	77
Figure 4-4. Characterisation of open chromatin profiles of primary human megakaryocytes (MKs), erythroblasts (EBs) and monocytes (MOs)	78

Figure 4-5. Distance of NDRs to the closest TSSs.....	79
Figure 4-6. Overlap of H3K4me3 (promoter) and H3K4me1 (enhancer) histone marks with NDRs identified in MKs and EBs with respect to NDR score and distance to the closest TSS.....	79
Figure 4-7. Bootstrapped quantile-quantile distributions	81
Figure 4-8. Enrichment of associations with platelet and erythrocyte phenotypes in NDRs across quantitative haematological traits, cell types and NDR classes.....	83
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	85
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	86
Figure 4-11. Estimated false discovery rates (FDRs) for mean red cell volume-associated SNPs in cell type-specific NDRs	87
Figure 4-12. Electrophoretic mobility shift assays (EMSAs) for platelet candidate functional SNPs	88
Figure 4-13. Functional follow-up of the <i>ABCC4</i> platelet count locus	92
5. Functional follow-up of the platelet volume and function locus at chromosome 7q22.3	95
Figure 5-1. Functional follow-up of the 7q22.3 locus associated with platelet volume and function	98
Figure 5-2. The effect of rs342293C>G on transcription factor binding.....	101
Figure 5-3. Gel shift assays in CHRF-288-11 nuclear protein extracts using GATA1 and RUNX1 antibodies.....	102
Figure 5-4. GATA1 and RUNX1 ChIP-seq profiles at the MK-specific open chromatin region at chromosome 7q22.3 in primary megakaryocytes.....	102
Figure 5-5. Gene expression profiles in differentiated human blood cells	103
Figure 5-6. Association of rs342293 genotypes with <i>PIK3CG</i> transcript levels in platelets and macrophages.....	104
Figure 5-7. Protein-protein interaction network centred on <i>PIK3CG</i>	106
Figure 5-8. Model of the mechanism by which rs342293 may affect platelet volume	109
6. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in <i>RBM8A</i> causes TAR syndrome.....	110
Figure 6-1. Skeletal abnormalities in TAR cases.....	113
Figure 6-2. Low-frequency non-coding SNPs and a rare null mutation at the <i>RBM8A</i> locus.....	116

Figure 6-3. The effect of the regulatory 5'-UTR and intronic SNPs on transcription factor binding.....	118
Figure 6-4. Luciferase reporter assays in cell lines representative of MKs (CHRF-288-11 and DAMI) and osteoblasts (MC3T3).....	119
Figure 6-5. Immunoblot staining for Y14, the protein encoded by <i>RBM8A</i> , and densitometry analysis.....	120
7. Conclusions and outlook.....	124
Figure 7-1. Example of the systematic genome annotation using chromatin maps across different cell types.....	128
Figure 7-2. Integrative genomics and molecular networks.....	129

LIST OF TABLES.

1. Introduction	1
Table 1-1. Overview of different types of histone modifications	11
Table 1-2. Overview of different functional types of gene promoters in vertebrates.....	16
2. Materials and methods	29
Table 2-1. Overview of experimental parameters applied in FAIRE experiments	36
Table 2-2. DNA quantity of FAIRE and reference samples before and after labelling with cyanine dyes	37
Table 2-3. Overview of sequencing statistics.....	40
Table 2-4. Overview of FAIRE peak statistics	41
Table 2-5. Overview of the number of FAIRE peaks for each intensity bin.....	41
Table 2-6. Sanger sequencing primer pairs for two sequence-tagged sites at chr7q22.3.....	45
Table 2-7. EMSA probes.....	46
Table 2-8. Overview of the experimental setup for EMSA and supershift experiments	49
Table 2-9. Genotyping and gene expression platforms used for eQTL analyses.....	50
Table 2-10. Assessment of quantity and quality of total RNA and biotin-labelled cRNA.....	51
Table 2-11. Primer pairs used for Sanger sequencing of the <i>RBM8A</i> locus.....	53
3. Maps of open chromatin guide the functional follow-up of genome-wide association signals at haematological trait loci	57
Table 3-1. Summary of the genetic loci included on the custom DNA tiling array.....	61
Table 3-2. Comparison of the FAIRE peak density between the ENCODE data set and the data sets presented here	63
Table 3-3. Characterisation of open chromatin regions in relation to gene annotations.....	64
Table 3-4. Seven sequence variants associated with haematological and cardiovascular-related traits are located in NDRs	66
4. Maps of open chromatin highlight cell type-specific patterns of regulatory sequence variation at haematological trait loci	72
Table 4-1. Trends of up- or downregulation of genes close to FAIRE peaks during haematopoietic differentiation	80
Table 4-2. Pearson's correlation coefficients between erythrocyte traits	84

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.....	91
5. Functional follow-up of the platelet volume and function locus at chromosome 7q22.3.....	95
Table 5-1. Resequencing of the MK-specific open chromatin region at chromosome 7q22.3.....	100
6. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in <i>RBM8A</i> causes TAR syndrome.....	110
Table 6-1. Genotyping of the 5'-UTR and intronic SNPs at the <i>RBM8A</i> locus in 7,504 healthy individuals of the Cambridge BioResource and association with platelet count.....	115
8. Appendix.....	130
Table 8-1. Genetic loci selected for the high-density DNA tiling array.....	131
Table 8-2. Ontology analysis of genes flanking FAIRE peaks using GREAT.....	136
Table 8-3. SNPs associated with platelet and erythrocyte phenotypes located in open chromatin in primary human MKs, EBs and MOs.....	142
Table 8-4. <i>In silico</i> transcription factor binding site predictions.....	153
Table 8-5. Investigation of the functional role of platelet volume-associated variants at chromosome 7q22.3.....	156
Table 8-6. Expression QTL associations at the <i>PIK3CG</i> gene locus in platelets, macrophages, monocytes, B cells (LCLs), adipose and skin.....	158
Table 8-7. Functional ontology classification of differentially expressed genes between <i>Pik3cg</i> ^{-/-} and wild type mice.....	159
Table 8-8. Genotype and phenotype information for TAR cases and unaffected parents.....	161

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	<i>cis</i> -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