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.

LIST OF FIGURES.

LIST OF TABLES.

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