The role of regulatory variation in sculpting gene expression across human populations and cell types

Antigone Dimas

Darwin College

University of Cambridge

August 2009

This dissertation is submitted for the degree of Doctor of Philosophy

DECLARATION

This dissertation describes my work undertaken in the group of Dr Manolis Dermitzakis, at the Wellcome Trust Sanger Institute, in fulfilment of the requirements for the degree of Doctor of Philosophy, at Darwin College, University of Cambridge. This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where indicated in the text. The work described here has not been submitted for a degree, diploma or any other qualification at any other university or institution. I confirm that this dissertation does not exceed the page limit specified by the Biology Degree Committee.

Antigone Dimas
Cambridge, August 2009

ABSTRACT

Genetic variants that influence expression levels of genes have a key role in shaping phenotypes. From cell type definition during development, to sculpting higher level traits, within and across populations, in health and disease, the importance of regulatory variation is emerging rapidly. The goal of this thesis was to identify genetic variants that shape gene expression levels (expression quantitative trait loci or eQTLs) across different human populations and cell types. Three general aspects of regulatory variation were addressed: a) impact of interactions between regulatory (eQTLs) and protein-coding variants (non-synonymous SNPs or nsSNPs) on gene expression in cis and trans, b) fine-scale architecture of the cis regulatory landscape, c) cell type specificity of eQTLs. To do this, I performed association of transcript levels (as a proxy to gene expression) with SNP genotypes and identified eQTLs using two resources: a) the HapMap Project for which expression was quantified in lymphoblastoid cell lines (LCLs) of geographically diverse populations and b) the GenCord Project for which expression was quantified in fibroblasts, LCLs and T-cells of a single population of European descent.

HapMap was used to explore a specific model of epistasis between eQTLs and nsSNPs, in which the functional impact of nsSNPs is modulated by regulatory variants nearby. From a total of 8,233 nsSNPs interrogated, 1,502 (18.2%) were found to be differentially expressed (DE), with important implications for protein diversity in the cell. Modification in cis also had an impact on gene expression in trans with a subset of DE nsSNPs being associated with expression variation of other genes in the genome.

To explore the architecture of the cis regulatory landscape and given the need to identify functional variants, I designed a framework to dissect and fine-map regulatory

variation. Using HapMap, and upon correction for the correlated structure of variants in the genome, it was found that over 19% of genes have multiple cis eQTLs, but also that single eQTLs can regulate the expression of multiple genes. The multidimensionality and complex architecture of cis regulation was further highlighted by showing that interactions between genetic variants in cis influence gene expression levels.

Cell type specificity of regulatory variation was addressed using GenCord and it was found that over 83% of independent cis eQTLs were unique to a single cell type. Importantly, LCL eQTLs replicated well across studies with over 80% of HapMap eQTLs replicating in GenCord, an observation that demonstrates the usefulness and stability of large collections of LCLs. GenCord cell type-specific cis eQTLs were found to span a wide range of distances from the transcription start site (TSS) of genes mirroring the distribution of known enhancer elements. Furthermore, a correlation between number of cis eQTLs identified for a given gene and number of transcripts was detected.

Given the role of gene expression in shaping phenotypic variation in health and disease, elucidating the nature of regulatory variation is crucial. Especially in the case of disease, integrating regulatory information with the results of genome-wide disease association studies is a promising way forward and will help unravel mechanisms leading to disease pathogenesis.

PUBLICATIONS

Publications arising from the work described in this thesis:

International Headache Genetics Consortium (including **A.S. Dimas** and E.T. Dermitzakis). Genome-wide analysis of migraine identifies a common variant on 8q22.1 modulating glial glutamate transport. [*submitted*].

Nica, A.C., S.B. Montgomery, **A.S. Dimas**, B.E. Stranger, C. Beazley, I. Barroso, E.T. Dermitzakis. Causal regulatory effects for complex trait associations. 2009. [*under review*]

Ritchie, M.E., M.S. Forrest, **A.S. Dimas**, C. Daelemans, E.T. Dermitzakis, P. Deloukas, S. Tavare. Data analysis issues for allele-specific expression using Illumina's GoldenGate assay. 2009. [*under review*].

Borel C., S. Deutsch, A. Letourneau, E. Migliavacca, H. Attar, **A.S. Dimas**, M. Gagnebin, C. Gehrig, E. Falconnet, Y. Dupré, S.E. Antonarakis. Identification of *cis*- and *trans*-regulatory variation modulating miRNA expression levels in human fibroblasts. [*under review*].

Dimas, A.S. and Dermitzakis, E.T. Genetic variation of regulatory systems. 2009. **Curr Opin Genet Dev** 19:586-590.

Dimas, A.S., S. Deutsch, B.E. Stranger, S. Montgomery, C. Borel, C. Ingle, C. Beazley, M. Gutierrez Arcelus, H. Attar-Cohen, M. Sekowska, M. Gagnebin, J. Nisbett, P. Deloukas, E. T. Dermitzakis, S. E. Antonarakis. 2009. Most common regulatory variation impacts gene expression in a tissue-dependent manner. **Science** 325: 1246-1250.

Dimas, A.S., B.E. Stranger, C. Beazley, R.D. Finn, C.I. Ingle, M.S. Forrest, M. Ritchie, P. Deloukas, S.Tavaré, E.T. Dermitzakis. 2008. Modifier effects between regulatory and protein-coding variation. **PLoS Genetics** Oct;4(10):e1000244.

Stranger, B.E., A.C. Nica, M.S. Forrest, **A. Dimas**, C.P. Bird, C. Beazley, C.E. Ingle, M. Dunning, P. Flicek, S., Montgomery, S. Tavaré, P. Deloukas, E.T. Dermitzakis. 2007. Population genomics of human gene expression. **Nat Genetics** 39: 1217-1224.

The ENCODE Project Consortium (including B.E. Stranger, E.T. Dermitzakis, **A. Dimas**). 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. **Nature** 447: 799-816.

ACKNOWLEDGEMENTS

In 2004 I was working as a journalist for a Greek newspaper. For various reasons, after studying biology and biological anthropology, I decided to leave science and worked first as a management consultant and then as a journalist. But it was not what I wanted to do and I missed science very much. I was visiting research institutes in Athens to find a group working on human genetic variation, when I came across an advertisement for a two-day symposium on human genetics, complex traits and disease. It was going to take place in Athens, on the following day and I blinked in disbelief when I read through the list of invited speakers. That is where I met Manolis and I would have never guessed that my quest for a PhD in Athens would eventually lead me to his group at the Sanger Institute. Although we had some difficult times, Manolis was a great supervisor and I would like to thank him for his enthusiasm, support and ideas, but also for believing that I could do this after such a long absence from the field. I am very glad that I will continue my work with him at the University of Geneva and am sure that we will work on even more exciting projects. With Manolis I was able to share my fascination for human genetic variation which I am sure has kept us both awake until the early morning hours. Human variation was the reason I studied biology in the first place and I was lucky enough to be taught by Ryk Ward at the Department of Biological Anthropology in Oxford. Ryk was one of the most inspiring people I have met and through him I discovered what fascinates me most in science. I am sure that if it weren't for him, I would not be writing this now. I am also very grateful to Simon Tavaré who was a great co-supervisor. His positive view on work and life, as well as his exciting ideas were very motivating throughout these years. I would also like to thank Panos Deloukas for his guidance and reassuring comments, especially during thesis committee meetings.

Over the years, our group at Sanger grew and every member has contributed to this dissertation in one way or another. I would like to thank Barbara Stranger for her invaluable support, advice and patience. Coming back to science after a five year absence was more difficult than I had anticipated and her help and friendship has been very important. I would also like to thank Claude Beazley for his precious friendship, help and support and although it's a cliché, it's true: without him this thesis would not have been possible. Each member of the group has contributed in a very important way, either practically or through their presence: big thanks go to Catherine Ingle, Alexandra Nica, Stephen Montgomery, James Nisbett, Magdalena Sekowska, Daniel Jeffares, Christine Bird, Tsun-Po Yang, and Maria Gutierrez-Arcelus.

Throughout these years, I had the opportunity to collaborate with some excellent scientists to whom I am indebted: Samuel Deutsch, Cristelle Borel and Stylianos Antonarakis at the University of Geneva Medical School, Mark Dunning, Matthew Ritchie and Doug Speed at the Cambridge Research Institute, Robert Finn, Matthew Forrest, Naomi Hammond, Verneri Anttila and Virpi Leppa at the Sanger Institute. I would also like to thank Ken Weiss for his inspiring articles, discussions and emails over the years and Mark McCarthy for his enthusiasm and interest in this work.

A big thank you goes to the PhD students in my year for the good and challenging times. Alex Bateman, Christina Hedberg-Delouka and Annabel Smith have done a great job in taking care of us, and I am very grateful to them, as I am to Joan Green for compiling Journal Picks and to Andrew King and Frances Martin who have been very patient with my constant book renewals –some were ongoing for over a year.

During these four years I was lucky enough to make some very dear friends. Eleni and Samrah made Cambridge feel like home and coming back to Spitaki was always something to look forward to. Thank you to Bryndis for her great warmth, support and for being on the same wavelength and understanding. B&Bj and Raffaella, made my life in Cambridge exciting and fun and that was not easy. Gareth and Salim have been great friends and housemates through good and difficult times. Nikoleta and Panos managed to make my very difficult first year in Cambridge fun and at times very amusing. Once again I find that only when leaving a place I realise how many good friends I leave behind.

On the Athens side, the biggest thank you goes to my mother for her 100% support and enthusiasm for my decision, and to my brothers Costas and Christos for believing (and knowing) that I worked on something exciting. Thanks also to my father who gave me my first anthropology book.

I am very grateful to Yiannis, who may not know, but his keenness for the PhD idea motivated me to leave Athens - his PhD plant is doing well. Jamal was wonderful company when I was writing up despite his snores and Sofia was a wonderful hostess – finishing my thesis on a Greek island was very inspiring. A great big thanks goes to Stefanos for being such a wonderful friend and for listening to PhD pros and cons possibly up to 100 times in Tynda. This was definitely the best decision I have made so far.

BRUEGHEL'S TWO MONKEYS

This is what I see in my dreams about final exams:
two monkeys, chained to the floor, sit on the windowsill,
the sky behind them flutters,
the sea is taking its bath.

The exam is History of Mankind.
I stammer and hedge.

One monkey stares and listens with mocking disdain,
the other seems to be dreaming away–
but when it's clear I don't know what to say
he prompts me with a gentle
clinking of his chain.

Wisława Szymborska

Translation by Stanisław Barańczak & Clare Cavanagh

ABBREVIATIONS

ACE	angiotensin-converting enzyme
ANOVA	analysis of variance
API	application programme interface
ASE	allele-specific expression
ASW	African ancestry in Southwest USA
BMI	body mass index
bp	base pairs
CD	Crohn disease
cDNA	copy DNA
CEPH	Centre d'Étude du Polymorphisme Humain
CEU	Utah residents with Northern of Western European ancestry
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado, USA
ChIP	chromatin immunoprecipitation
CNV	copy number variant
CRI	Cambridge Research Institute
DE	differentially expressed
ds	double stranded
EBI	European Bioinformatics Institute
EBV	Epstein-Barr virus
ENCODE	encyclopedia of DNA elements
eQTL	expression quantitative trait locus
EST	expressed sequence tag
FDR	false discovery rate
gDNA	genomic DNA
GEO	gene expression omnibus
GIH	Gujarati Indians in Houston, Texas, USA
GO	gene ontology
GWAS	genome-wide association study/studies
HLA	human leukocyte antigen
IBP	insulator binding protein
IVT	in vitro transcription
JPT	Japanese in Tokyo, Japan

Kb	kilobase
LCLs	lymphoblastoid cell lines (EBV-transformed B-cells)
LCR	locus control region
LD	linkage disequilibrium
LR	linear regression
LWK	Luhya in Webuye, Kenya
MAF	minor allele frequency
Mb	megabase
MEX	Mexican ancestry in Los Angeles, California, USA
miRNA	microRNA
MKK	Maasai in Kinyawa, Kenya
MS	multiple sclerosis
M-W	Mann-Whitney test
nAChR	neuronal nicotinic acetyl choline receptor
nsSNP	non-synonymous SNP
OMIM	online mendelian inheritance in man
OPA	oligo pool all
PCA	principal components analysis
PHA	phytohemagglutinin
QTL	quantitative trait locus
RMA	repeated measures ANOVA
RT-PCR	reverse transcriptase polymerase chain reaction
SAM	sentrix array matrix
SMA	spinal muscular atrophy
SNP	single nucleotide polymorphism
SRC	Spearman rank correlation
SSAHA	sequence search and alignment by hashing algorithm
TES	transcription end site
TF	transcription factor
TSI	Toscans in Italy
TSS	transcription start site
UGMS	University of Geneva Medical School
UTR	untranslated region
WTSI	Wellcome Trust Sanger Institute
YRI	Yoruban in Ibadan, Nigeria

TABLE OF CONTENTS