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

1 INTRODUCTION

In this chapter I will:

- Define gene expression as the transfer of information from DNA to mRNA and then into protein.
- Explain that gene expression is a complex, quantitative trait with naturally occurring variation in its patterns and levels. These patterns have a key role in defining and maintaining cell types, and in shaping higher level phenotypes in the normal and disease ranges.
- Give a brief overview of the process of gene expression.
- Outline that this process can be regulated at many levels, the most important of which is transcription initiation which involves the action of cis and trans regulatory elements.
- Argue that a component of variation in expression levels is heritable and arises as a consequence of genetic variation in cis and trans-acting regulatory elements.
- Outline strategies employed to uncover genetic variants responsible for expression variation.
- State the aims of this thesis.

1.1 WHAT IS GENE EXPRESSION?

The process by which a gene gives rise to a functional product is called gene expression (Lewin 2008). Gene expression enables the phenotypic manifestation of genes, results in the production of a protein or a functional RNA molecule (e.g. rRNA, tRNA, microRNA) and is necessary for cells to operate.

Gene expression is a complex trait shaped by genetic (Monks, Leonardson et al. 2004; Morley, Molony et al. 2004; Cheung, Spielman et al. 2005; Stranger, Forrest et al. 2007; Stranger, Nica et al. 2007), epigenetic (Eckhardt, Lewin et al. 2006; Petronis 2006), and environmental (Gibson 2008; Idaghdour, Storey et al. 2008) factors. Interactions between genetic factors (Brem, Storey et al. 2005; Dimas, Stranger et al. 2008), as well as those between genetic factors and the environment (Gibson 2008) also affect the expression levels of genes. As a result, this phenotype exhibits continuous variation among individuals and has the properties of a quantitative trait (Dermitzakis 2008).

In this thesis I address protein-encoding genes whose expression can be divided in two stages: transfer of information from DNA to RNA (transcription) and transfer of information from RNA to protein (translation). A number of mechanisms control these processes, including chromatin condensation, alternative splicing, DNA methylation, transcription initiation, mRNA stability, translational control, post-translational control and protein degradation. In eukaryotic cells the most common point of control is transcription initiation (Stranger and Dermitzakis 2005). In the following sections I discuss the role of gene expression in shaping phenotypes, I give a brief overview of this biological process and outline how it is controlled.

1.2 GENE EXPRESSION DEFINES PHENOTYPES

The idea that gene expression levels play a role in shaping phenotypes is not new. In her doctoral thesis, Marie-Claire King revolutionized evolutionary biology by proving, through the comparative study of proteins, that human and chimpanzee genomes are 99% identical. In their landmark paper, King and Wilson (1975, pg 107) stated that human and chimpanzee macromolecules "are so alike that regulatory mutations may account for the biological differences between these species". In the following sections I discuss that naturally occurring variation in expression levels is widespread, that

expression patterns define and maintain cell types and sculpt higher level phenotypes in the natural and disease ranges.

1.2.1 Naturally occurring variation in gene expression levels

Developmental biology was one of the first fields that recognised the importance of expression patterns in shaping phenotypes. Spatially and temporally regulated expression is critical for the complex developmental programmes that result in the highly specialised cell types of higher eukaryotes. Following development and differentiation, the control and maintenance of appropriate levels and patterns of gene expression are vital cellular processes. Although expression of genes in the right quantity range, at the right time and in the right place is largely responsible for normal functioning of cells, biological systems can also display remarkable robustness. In most species studied (including yeast, fruit flies, mice, and humans) ample tolerance of naturally occurring variation in expression levels has been detected (Hartman, Garvik et al. 2001; Jin, Riley et al. 2001; Brem, Yvert et al. 2002; Schadt, Monks et al. 2003; Stranger and Dermitzakis 2005; Stranger, Forrest et al. 2005; Boone, Bussey et al. 2007; Gibson 2008). In a cross between two strains of yeast for example, profound differences in gene expression were found, with nearly half (2,698 out of 6,215) of all the genes in the genome being differentially expressed (DE) (Brem, Yvert et al. 2002). A study exploring human natural gene expression variation in 16 individuals of European and African descent found that 83% and 17% of genes were DE among individuals and populations respectively (Storey, Madeoy et al. 2007). In another study, three populations of apparently healthy Moroccan Amazigh (Berbers) were found to differ for over a third of their leukocyte transcriptome (Idaghdour, Storey et al. 2008). Taken together these results demonstrate that naturally occurring variation in expression levels is widespread within species.

1.2.2 Gene expression patterns define cell type specificity

Although there is space for expression variation, temporal and spatial regulation of gene expression patterns is critical for defining cell types during development in higher eukaryotes. Furthermore gene expression patterns have a role in the maintenance of cellular and tissue function following differentiation. Some examples are discussed below.

1.2.2.1 Gene expression defines cell type during development

Mammalian skeletal muscles are highly distinctive cells whose differentiation is triggered by expression of specific myogenic proteins including MyoD, Myf5, myogenin and Mrf4. The potency of these proteins was highlighted in a study where their expression in skin fibroblasts triggered muscle differentiation (Alberts 2008). Eye development also illustrates the developmental role of gene expression. In *Drosophila*, mice and humans this process involves highly regulated expression patterns of the gene *Ey* (*Drosophila*) and *Pax-6* (vertebrates). *Ey* expression triggers the formation of a specific cell type, but also of an entire organ, composed of different cell types and arranged in three dimensional space (Alberts 2008).

The critical role of expression patterns even for very closely related genes was demonstrated in an experiment where the transcription factor (TF) gene *SOX10* was deleted in mouse embryos and replaced with *SOX8*, a closely related gene that has overlapping expression patterns (Kellerer, Schreiner et al. 2006). SOX10 has a role in neural crest development and is defective in the human Shah-Waardenburg syndrome. It is essentially expressed in neural crest derivatives that form the peripheral nervous system and in the adult central nervous system (Bondurand, Kobetz et al. 1998). Both genes perform very similar functions and regulate processes such as enteric nervous system development and oligodendrocyte differentiation. Despite their similarities,

SOX8 phenotypic rescue of SOX10 deficiency was variable for different tissues: development of glial cells and neurons in the sensory and sympathetic parts of the peripheral nervous system was almost normal, but melanocyte development was as defective as in SOX10-deficient mice. Furthermore rescuing of defects in enteric nervous system development and oligodendrocyte differentiation was limited. These results highlight the importance of tissue-specific gene expression and demonstrate that the extent of functional equivalence depends on cell type (Kellerer, Schreiner et al. 2006).

1.2.2.2 Cell type-specific patterns of gene expression in differentiated cells

Once cells have undergone differentiation, expression profiles remain critical for maintenance of cellular function. In one of the first studies to explore genome-wide expression signatures, over 1,000 expressed sequence tags (ESTs) were sampled in 30 tissues (Adams, Kerlavage et al. 1995). Substantial tissue specificity of gene expression was detected with only eight genes sharing ESTs across all tissues, and 227 genes being represented in at least 20 tissues. A subsequent study interrogated transcription levels in 46 human and 45 mouse tissues, organs, and cell lines spanning a broad range of biological conditions (Su, Cooke et al. 2002). Only 6% of the genes interrogated were found to be ubiquitously expressed and hierarchical clustering identified groups of genes with specific expression patterns in nearly all tissues examined. Another study explored expression patterns of human orthologue genes from chromosome 21 in mice using RNA *in situ* hybridization and reverse transcriptase polymerase chain reaction (RT-PCR) (Reymond, Marigo et al. 2002). Patterned expression was observed in several tissues including those affected in trisomy 21 phenotypes (central nervous system, heart, gastrointestinal tract, and limbs). Taken together these examples underline that gene expression is a phenotype displaying extensive cell type and tissue specificity.

1.2.3 Gene expression shapes normal range phenotypes

Variation in expression levels is to a large extent responsible for shaping higher level phenotypes in the normal range. In *Drosophila* expression variation in the developmental gene *Svb* underlies trichome pattern differences between species (McGregor, Orgogozo et al. 2007). Expression patterns of the Hox gene *Ubx* outline trichomes on the posterior femur of the second leg (Stern 1998), and male-specific wing pigmentation spots in *D. biarmipes* are a consequence of varying expression levels of the yellow pigmentation gene *y* (Figure 1 a) (Gompel, Prud'homme et al. 2005). In *Geospiza* (Darwin's finches) diverse beak shape and morphology is in part due to expression differences of the gene *Bmp4* (Abzhanov, Protas et al. 2004). Expression patterns in the mesenchyme of upper beaks correlates with beak morphology (Figure 1 b) and when misexpressed in chicken embryos, Bmp4 causes morphological transformations that parallel the beak morphology of the large ground finch *G. magnirostris.*

The predominant differences in branching patterns in domesticated maize (*Zea mays mays*) and its wild ancestors, the teosintes (*Z. mays parviglumis* and *mexicana*) arise in part from expression differences of the gene *tb1* (Clark, Wagler et al. 2006). In *Gasterosteus aculeatus* (sticklebacks), expression variation of the developmental gene *Pitx1* in pelvic and caudal fin precursors results in pelvic reduction and major skeletal changes (Shapiro, Marks et al. 2004). Modified gene expression levels of the prairie vole gene *V1aR* give rise to differences in receptor distribution patterns in the brain. This is thought to affect a range of socio-behavioural traits, including social recognition and investigation, social odour tasks and parental care (Hammock and Young 2005).

Figure 1. Variation in gene expression levels and patterns underlies phenotypic differences. **a)** Like butterflies, different species of *Drosophila* decorate their wings with a great diversity of spots and patterns. Expression of a single gene produces pigmentation patterns and acts as a molecular switch that controls where pigmentation is deployed. This finding explains how expression can be controlled to produce the seemingly endless array of patterns, decoration and body architecture found in animals. Photo by N Gompel and B Prud'homme (from http://www.news.wisc.edu/newsphotos/fruitfly.html). **b)** Expression differences of Bmp4 give rise to beak morphology differences in *Geospiza*. *G. difficilis* is the most basal species of this genus, and the rest of the species form two groups: ground and cactus finches, with distinct beak morphologies. At stage 26 (middle panel) Bmp4 is strongly expressed in the mesenchyme of the upper beak of *G. magnirostris* and at lower levels in *G. fortis* and *G. conirostris*. No Bmp4 was detected in the mesenchyme of *G. difficilis*, *G. fuliginosa*, and *G. scandens*. At stage 29 (right panel) Bmp4 continues to be expressed at high levels in the distal beak mesenchyme of *G. magnirostris*. Broad domains of Bmp4 expression are detectable in *G. fuliginosa* and *G. fortis*. A small domain of Bmp4 expression is also found in the distal mesenchyme of *G. conirostris*, and weaker expression is seen in *G. scandens* and *G. fortis* (red arrows). Adapted from (Abzhanov, Protas et al. 2004).

In humans the ability to digest lactose, the main carbohydrate in milk, declines rapidly after weaning. This is due to decreasing levels of lactase-phlorizin hydrolase, which metabolises lactose and is encoded by the gene *LCT*. Adult expression of LCT results in the ability to digest milk and other dairy products in adulthood (lactase persistence or lactose tolerance) and differences in LCT levels lead to the differences in lactase persistence observed in a number of populations across the world (Tishkoff, Reed et al. 2007) (Figure 2). Overall these examples highlight the role of gene expression in shaping natural range phenotypes.

Figure 2. Lactase persistence differences across human populations are due to differences in adult expression of lactase. a) The degree of lactase persistence is represented by a pie chart for each geographic region (LP: lactase persistence, LIP: lactase intermediate persistence, LNP: lactase non-persistence). **b)** Proportion of compound genotypes of variants that influence levels of LCT (G/C-13907, T/G-13915 and C/G-14010). The pie charts are in the approximate geographic location of the sampled individuals. Adapted from (Tishkoff, Reed et al. 2007).

1.2.4 Gene expression can shape disease phenotypes

Variation in gene expression can have a detrimental impact on cells and tissues if expression profiles are perturbed beyond the range of tolerance (Hartman, Garvik et al. 2001; Stranger and Dermitzakis 2006; Nica and Dermitzakis 2008; Cookson, Liang et al. 2009). Many-fold over-expression of C-MYC can lead to Burkitt's lymphoma (Boxer and Dang 2001), a reduction of APC expression is associated with a pronounced predisposition to hereditary colorectal cancers (Yan, Dobbie et al. 2002) and partial or complete loss of α-globin expression can lead to α-thalassaemia (Weatherall 1998). Subtle changes in gene expression can also contribute to disease phenotypes, as is the case for Type 1 diabetes, whose manifestation depends on the genetic background of individuals (Eaves, Wicker et al. 2002). Type 1 diabetes was one of the first instances where genetic variation driving gene expression was shown to be associated with disease risk (Bennett, Lucassen et al. 1995; Kennedy, German et al. 1995). The insulin-linked polymorphic region (ILPR), mapping 5' of the *INS* gene is composed of a series of tandemly-repeated sequences that contain high affinity binding sites for the TF Pur-1. Allelic variation in these sequences was shown to influence *INS* transcription levels and risk for diabetes. Table 1 (Cookson, Liang et al. 2009) summarises cases from the literature and public databases in which trait and disease phenotypes arise in part due to variation in expression levels.

The link between human tissue-specific gene expression and pathological manifestations has been demonstrated in multiple studies. Lage et al. (2008) mapped 2,000 disease genes to the tissues they affect and identified 1,500 disease-associated complexes. The expression patterns of complex components were analysed and disease genes were found to be over-expressed in the normal tissues where defects eventually cause pathology. For example a complex involved in XY sex reversal was found to be testis-specific and was down-regulated in the ovaries. Tissue specificity of expression

was identified for complexes with a role in Parkingson disease, cardiomyopathies and muscular dystrophies.

Study	Trait	Region	Candidate gene(s)	Transcript affected by SNP	Transcript region	Logarithm of odds (LOD) score
Gudbjartsson *et al.*[102]*	Height	7p22	*GNA12*	*GNA12*	7p22	13
		11q13.2	Intergenic	*CCND1*	11q13	7.4
		7q21.3	*LMTK2*	*C17orf37*	17q21	6.0
				HSD17B8	6	6.4
				NDUFS8	11	6.1
		3p14.3	*PXK*	*RPP14*	3	9.2
Göring *et al.*[15]	High-density lipoprotein cholesterol levels	6q21	*VNN1*	*HDL* (serum)	Multiple sites	8.0
Kathiresan *et al.*[40]	Polygenic dyslipidaemia	20q13	*PLTP*	*PLTP*	20q13	16
		15q22	*LIPC*	*LIPC*	15q22	17
		11q12	*FADS1, FADS2, FADS3*	*FADS1*	11q12	35
				FADS3	11q12	8.0
		9p22	*TTC39B*	*TTC39B*	9p22	7.0
		1p13	*CELSR2, PSRC1, SORT1*	*SORT1*	1p13	270
				PSRC1	1p13	249
				CELSR2	1p13	80
		12q24	*MMAB, MVK*	*MMAB*	12q24	43
		1p31	*ANGPLT3*	*DOCK7*	1p31	27
				ANGPLT3	1p31	11
Libioulle *et al.*[37]	Crohn's disease	5p13	Intergenic	*PTGER4*	5p13	3.0
Barrett *et al.*[36]	Crohn's disease	5q31	*OCTN1, SLC22A4, SLC22A5*	*SLC22A5*	5q31	Unknown
Hom *et al.*[103]*	Systemic lupus erythematosus	8p23.1	*C8orf13, BLK*	*BLK*	8p23.1	20
				C8orf13	8p23.1	28
Harkonason *et al.*[104]*	Type 1 diabetes	12q13	*RAB5B, SUOX, IKZF4*	*RPS26*	12q13	33
		1p31.3	*ANGPTL3*	*DOCK7*	1p31.3	16
Wellcome Trust Case Control Consortium[105]*	Type 1 diabetes	12q13.2	*ERBB3*	*RPS26*	12q13.2	43.2
Todd *et al.*[106]*	Type 1 diabetes	12q13.2	*ERBB3*	*RPS26*	12q13.2	30.3
Plenge *et al.*[107]*	Rheumatoid arthritis	9q34	*TRAF1-C5*	*LOC253039*	9q34	6.3
Thein *et al.*[108]	Fetal haemoglobin F production	6q23.3	Intergenic	*HBS1L*	6q23.3	6.0
Moffatt *et al.*[30]	Childhood asthma	17q21	Intergenic	*ORMDL3*	17	14
Wellcome Trust Case Control Consortium[105]*	Bipolar disorder	16p12	*PALB2, NDUFAB1, DCTN5*	*DCTN5*	16p12	9.2
		6p21	*NR*	*HLA-DQB1*	6p21	8.9
				HLA-DRB4	6p21	11
Di Bernardo *et al.*[109]*	Chronic lymphatic leukaemia	2q37	*SP140*	*SP140*	2q37	8.8

Table 1. Trait and disease phenotypes with an identified gene expression component. Disease-linked associations with significant expression quantitative trait loci (QTLs) from the literature and public databases From (Cookson, Liang et al. 2009).

Tissue specificity in disease pathology has also been addressed from an evolutionary standpoint in a study exploring the relationship between disease genes, tissue specificity, and evolutionary rates (Winter, Goodstadt et al. 2004). Cell type specificity is known to correlate positively with gene evolution rates and ubiquitously expressed, slowly evolving housekeeping genes were found to be under-represented in human disease. Genes with a role in disease on the other hand, had secreted protein products and were highly expressed in tissues such as liver, kidney and lung. This observation is likely due to the effects of purifying selection and may assist in prioritization of candidate genes.

A recent study highlighted the role of a single TF in regulating markedly different cell type-specific programmes (Servitja, Pignatelli et al. 2009). HNF1A controls tissue-specific genetic programmes in pancreatic islets and the liver, and its deficiency causes a severe β-cell phenotype (HNF1A-deficient diabetes), but only subtle abnormalities in other tissues. The final phenotypic outcome of Hnf1a deficiency in mice was highly cell type-specific and resulted from an integrated failure of multiple direct and indirect functions of this gene in pancreatic islets and liver. Due to the breadth of Hnf1a-dependent transcriptional programmes, the authors suggest that correction of defects causing β-cell dysfunction should not focus on restoring individual target gene activity, but should aim at manipulating proteins or pathways acting on the β-cell HNF1A-dependent programme. Taken together these examples outline the multiple effects of gene expression patterns and the role of gene regulation in determining disease risk.

1.3 THE MECHANISM OF GENE EXPRESSION

In eukaryotic cells, protein-encoding genes are transcribed in the nucleus by RNA polymerase II. The RNA transcript produced is the messenger RNA (mRNA) which acts as an intermediary between the gene and the protein product. A further step, known as

translation, is necessary to convert the information carried by the mRNA into a protein (Clark 2005). In the following section I outline the process of gene expression for protein-encoding genes (summarised in Figure 3).

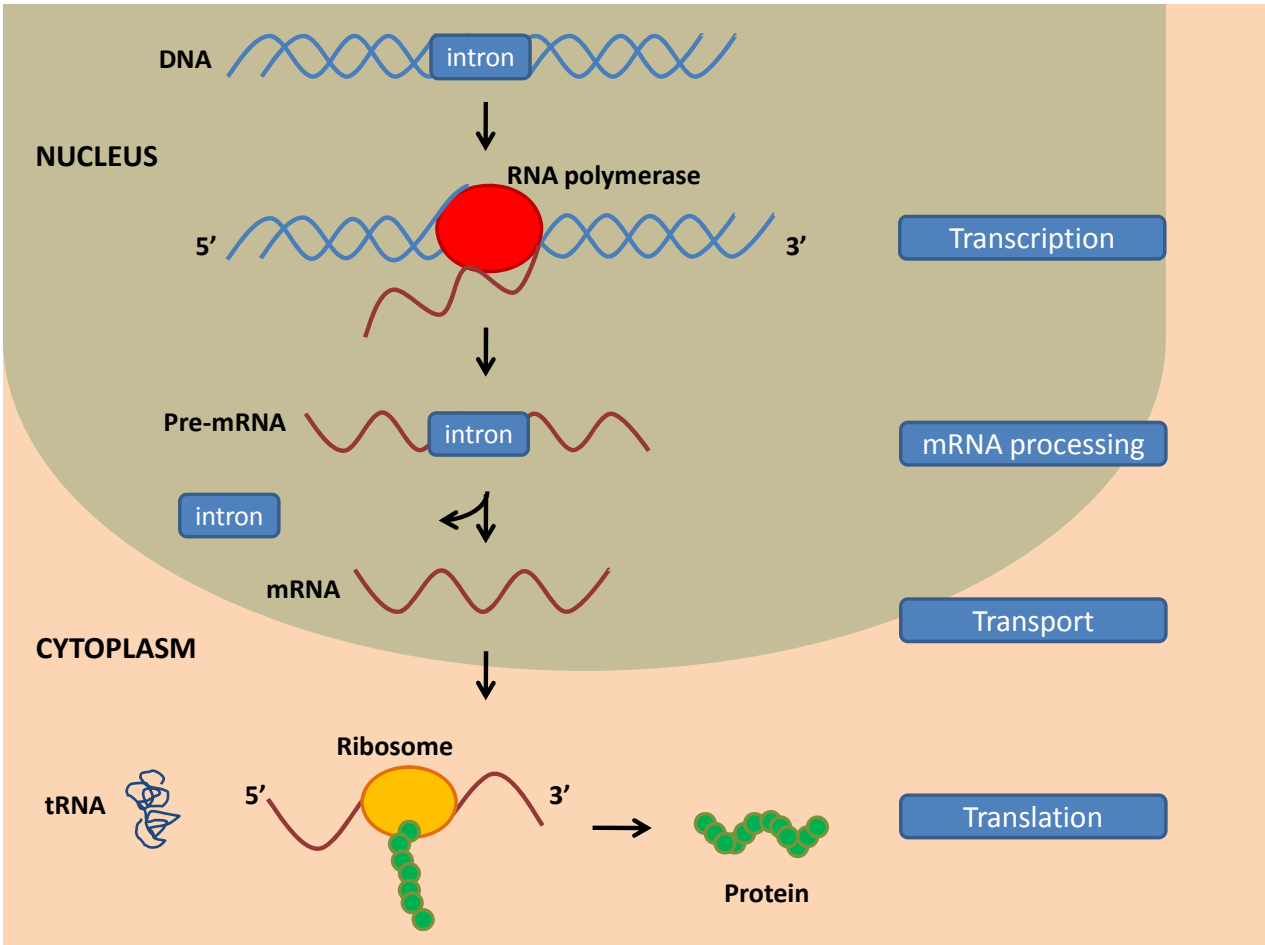


Figure 3. The process of gene expression for protein-encoding genes. RNA is transcribed from a DNA template by RNA polymerase II in the nucleus to produce pre-mRNA (transcription). The pre-mRNA undergoes a series of processing steps including splicing, 5′ capping and 3′ polyadenylation (mRNA processing). Processed mRNA molecules are transferred from the nucleus into the cytoplasm (transport) where they engage with ribosomes and other components of the translational machinery that direct polypeptide synthesis (translation). Adapted from (Clark 2005).

1.3.1 Transcription

RNA polymerase II uses nuclear DNA as a template and ribonucleoside triphosphates to produce pre-mRNA molecules in a 5′ to 3′direction. Chain elongation is achieved by the addition of ribonucleoside monophospate residues to the free hydroxyl group at the 3′ end of the growing pre-mRNA chain (Strachan and Read 2004). This process gives rise to the primary transcript (or pre-mRNA), an RNA molecule complementary to the full sequence of the gene (exons and introns).

1.3.2 mRNA processing

The pre-mRNA undergoes a series of processing steps in the nucleus including splicing, 5′ capping and 3′ polyadenylation. Splicing is mediated by the spliceosome, a large RNA-protein complex, which recognises sequences at exon/intron boundaries (splice junctions). Intronic RNA segments are removed by endonucleolytic cleavage and exonic segments are joined end-to-end (spliced). The end product is a shorter RNA product (mRNA) that contains the information encoding a protein (exons). Alternative splicing can bring together different combinations of exons to produce versions of the polypeptide product (isoforms).

Further mRNA processing involves addition of a methylated nucleoside, 7-methylguanosine (5′ cap) to the first 5′ nucleotide of the RNA transcript, as well as addition of a 3′ polyA tail. 5′ caps and 3′ polyA tails facilitate transfer of mRNA molecules to the cytoplasm, ensure RNA stability, and assist recognition by the translational machinery (Strachan and Read 2004). In some instances in somatic cells, mRNA molecules undergo RNA editing, which results in a coding sequence difference between mRNA and DNA sequence. mRNA editing of *APOB* gene transcripts in the liver for example introduces a stop codon in the mRNA transcript and gives a much

shorter product from the one generated by the unedited mRNA molecule in the intestine (Navaratnam, Bhattacharya et al. 1995; Lewin 2008).

1.3.3 mRNA transport and translation

Following post-transcriptional processing, mRNA molecules migrate from the nucleus to the cytoplasm where they engage with ribosomes and other components of the translational machinery that direct polypeptide synthesis (Strachan and Read 2004). The central part of mRNA molecules encodes the amino acid sequence whereas 5′ and 3′ mRNA ends are untranslated regions (UTRs) (transcribed from the first and terminal exons respectively) with a role in binding and stabilizing mRNA on ribosomes.

Assembly of polypeptides from their constituent amino acids is governed by the triplet genetic code with successive groups of three nucleotides (codons) in the linear mRNA sequence encoding an individual amino acid. Decoding of mRNA is mediated through tRNA molecules that bear specific trinucleotide sequences (anticodons) and covalently bound amino acids. Recognition of the complementary codon on the mRNA ensures that the appropriate amino acid is inserted in the growing polypeptide chain. Translation products are frequently modified, usually through covalent attachment of hydroxyl, phosphoryl, carbohydrate and lipid groups to amino acid side chains. Upon modification, polypeptides may undergo cleavage to generate smaller, mature proteins (e.g. β-globin, plasma proteins, neuropeptides) (Strachan and Read 2004).

1.4 Regulation of gene expression

As discussed, gene expression is influenced by genetic, epigenetic, and environmental factors that give rise to expression differences between species, populations and cell types. Furthermore, interactions between genetic factors (Brem, Storey et al. 2005; Boone, Bussey et al. 2007; Dimas, Stranger et al. 2008) as well as those between genetic

factors and the environment (Gibson 2008) have a key role in shaping expression levels. Although expression regulation involves multiple levels, in eukaryotes the most common point of control is initiation of transcription. For the purposes of this thesis the products of transcription (mRNA levels) were regarded as a proxy to gene expression (see section 2.3).

1.4.1 Transcriptional regulation of gene expression

The simplest model of transcription of protein-coding genes in eukaryotes involves recruitment of RNA polymerase II, which recognises and binds to combinations of short DNA sequences in the proximity of a gene. These sequence elements, are referred to as cis-acting and serve as recognition signals for TFs that engage in gene expression regulation by guiding and activating the polymerase (Strachan and Read 2004). Transcription initiation is influenced by DNA sequences located further away (or on another chromosome) from the gene whose activity is being regulated. These sequences are known as trans-acting and encode proteins that influence transcription levels (e.g. TFs). In this thesis cis regulatory elements are defined as those mapping within a 2 Mb window centred on the probe midpoint or the transcription start site (TSS) of a gene (see section 2.4) and trans-acting regulatory elements are those mapping outside this 2 Mb window or on another chromosome.

Multiple cis and trans elements act in conjunction with each other to control transcription initiation and mRNA levels for a given gene (Stranger, Forrest et al. 2005; Stranger, Nica et al. 2007; Dimas, Deutsch et al. 2009). The identity of regulatory sequences, the TFs present, and their binding affinities all play an important role in transcription initiation. Mutations altering the nucleotide sequence of any of these elements, or the nucleotide sequence of the transcript (affecting its stability), may have substantial effects on mRNA transcript levels (Stranger and Dermitzakis 2005). The

genomic distribution and complexity of cis and trans sequence elements, as well as the architecture of the regulatory landscape is an area of active research and a substantial effort is underway to annotate regulatory elements in the genome. A pilot project in which 1% of the genome was studied (Birney, Stamatoyannopoulos et al. 2007) revealed that the distribution of regulatory sequences is variable, with elements being scattered across the genome. In the following paragraphs I describe well-studied regulatory sequence elements, proteins and RNA molecules.

1.4.1.1 Promoters

Promoters are short sequence cis-acting elements that cluster in the immediate upstream region of a gene's coding sequence, often within 200 base pairs (bp) of the TSS, and control transcription initiation. RNA polymerase and a number of general TFs bind to the promoter region of a gene, which is typically made up of different components, to form the basal transcription complex. Upon binding, the polymerase is activated and RNA synthesis is initiated. Well-studied promoter elements are discussed in Box 1.

The **initiator box (Inr)** is a sequence bound by general TF TFIID at the site of transcription initiation. The first transcribed base of the mRNA is usually an A with a pyrimidine on each side.

The **core promoter sequence** directs the basal transcription complex (RNA polymerase and general TFs) to initiate transcription of a gene. In the absence of additional regulatory elements, it permits constitutive gene expression, but at very low (basal) levels. Core promoters are typically located very close to the TSS, at nucleotide positions -45 to +40, and include the **TATA box**, the **BRE sequence**, and the **DPE**. The TATA box (TATAAA sequence or a variation of it), found approximately at -25 of the TSS, is usually surrounded by GC rich sequences and is recognised by the TATA-binding protein subunit of TFIID. Immediately upstream of the TATA element is the BRE sequence which is recognised by TFIIB. The DPE (downstream promoter element) is located at approximately +30 and is recognised by TFIID.

Non-core promoter elements map immediately upstream of the core promoter, typically spanning the region of -50 to -200 bp and include **GC boxes (Sp1 boxes)** and **CCAAT boxes**. GC boxes are found in multiple copies within 100 bp of the TSS and are bound by the general TF Sp1. CCAAT boxes are strong determinants of promoter efficiency, locate approximately 80 bp upstream of the TSS, and are recognised by TFs CTF and CBF. Both GC and CCAAT boxes function to modulate basal transcription levels of the core promoters, operating as essentially as enhancer sequences.

Box 1. Promoter elements. Adapted from (Strachan and Read 2004).

1.4.1.2 Enhancers

Enhancers are positive control sequence elements, located at variable and often considerable distances from a gene, that increase the basal level of transcription initiated through promoter elements. They are short DNA sequences and may contain several elements recognised by TFs in a ubiquitous or tissue-specific manner (Heintzman, Hon et al. 2009; Visel, Blow et al. 2009). Upon TF binding, the DNA between the enhancer element and the promoter loops out and allows the proteins bound to the enhancer to interact with the basal transcription complex (Strachan and Read 2004). A well-studied enhancer is the locus control region (LCR) located 50-60 kilobases (kb) upstream of the *β-globin* gene whose expression it activates.

1.4.1.3 Silencers

Silencers have similar properties to enhancers, but act to reduce expression levels by inhibiting the transcriptional activity of genes. They have been reported in various positions relative to human genes: close to the promoter, upstream of the TSS, and within introns. Classical silencer elements are position-independent sequences that direct an active transcriptional repression mechanism (Strachan and Read 2004). Negative regulatory elements are position-dependent sequences that exert passive repression of transcription and often act by interfering with activators rather than by obstructing the movement of RNA polymerase.

1.4.1.4 Insulators

Insulators (boundary elements) are regions of DNA spanning a few hundred to a few thousand bases (typically 0.5-3 kb) which block the spreading of agents that affect transcription in a positive or negative manner, and divide chromosomes into regulatory neighbourhoods (Clark 2005). They contain clusters of GC rich sequences that bind multiple copies of zinc-finger proteins known as insulator binding proteins (IBPs). In many cases their action can be countered by methylation of GC sequences.

1.4.1.5 Response elements

Response elements are usually located a short distance upstream of promoter elements (1 kb upstream of the TSS) and are responsible for modification of transcription in response to environmental stimuli. Response elements can respond to specific hormones (e.g. retinoic acid or steroid hormones such as glucocorticoids) or to intracellular second messengers such as cyclic AMP (Strachan and Read 2004).

1.4.1.6 Transcription factors

RNA polymerase II transcribes genes following binding of TFs to specific regulatory DNA within the gene and its vicinity. TFs are typically regarded as trans-acting elements and may bind to the promoter region around genes or to distant enhancer sequences. Activators are TFs that stimulate transcription and repressors are those with antagonistic effects. TFs can be general (e.g. components of the basal transcription complex such as TFIIB or TFIID) or tissue-specific (e.g. HNF1A which controls tissue-specific expression in pancreatic islets and the liver). General TFs are required for transcription from all promoters occupied by RNA polymerase II and their binding results in basal levels of transcription. Specialised TFs modulate basal transcription levels and influence the activity of specific gene sets, usually in a tissue-specific manner.

1.4.1.7 MicroRNA

MicroRNAs (miRNAs) are single stranded, 21–24 nucleotide, regulatory RNA molecules abundant in animals, plants and viruses (Flynt and Lai 2008). They are encoded by genes from whose DNA they are transcribed, but are not translated into protein. Instead each transcript is processed into a short stem-loop structure called a pre-miRNA and finally into a functional miRNA. miRNA molecules are fully or partially complementary to mRNA molecules and their main function is to down-regulate gene expression through partial base pairing with their target mRNAs. Base pairing either inhibits translation of target mRNA molecules or speeds up deadenylation causing mRNA degradation (Williams 2008).

1.4.2 Other mechanisms of gene expression regulation

Although transcription is the primary means of expression regulation, gene activity can be modulated post-transcriptionally, through mechanisms involving mRNA processing,

transport and stability at the mRNA level, as well as translation, processing, targeting and stability at the protein level. Furthermore, expression regulation can be achieved epigenetically through DNA methylation, histone modification and the action of non-coding RNA molecules. A detailed overview of these mechanisms can be found in Genes IX (Lewin 2008).

1.5 GENETIC VARIATION IN GENE EXPRESSION

As described, gene expression is a complex, quantitative trait controlled at many levels and sculpted by numerous factors. In this thesis I address the genetic component of expression variation, or the fraction of transcript level differences that arises as a consequence of genetic variation in DNA sequences. Broadly speaking, genetic variation influencing gene expression can manifest itself in four major ways: gene expression differences among populations, among individuals in a population, among tissues, and in response to environmental factors. In this section I outline a number of landmark studies that have contributed to our understanding of the genetic component of gene expression.

The first series of large-scale studies aiming to uncover regulatory DNA variation focused on model organisms. A genetic component for naturally occurring variation in gene expression was documented in yeast (Brem, Yvert et al. 2002; Steinmetz, Sinha et al. 2002), maize (Schadt, Monks et al. 2003), fruit flies (Jin, Riley et al. 2001; Wittkopp, Haerum et al. 2004), and mice (Sandberg, Yasuda et al. 2000; Cowles, Hirschhorn et al. 2002; Lo, Wang et al. 2003; Schadt, Monks et al. 2003). In humans, familial aggregation of expression profiles was demonstrated by Cheung et al. (2003) who showed that variability in transcript abundance was lower in more closely related individuals. Gene expression heritability estimates for the same individuals showed that approximately 25% of the genes studied had significant heritable variation (Schadt,

Monks et al. 2003). This implied a heritable component of gene expression variation among humans and laid the groundwork for subsequent studies in primates (Enard, Khaitovich et al. 2002) and humans (Cheung, Conlin et al. 2003; Monks, Leonardson et al. 2004; Morley, Molony et al. 2004; Pastinen, Sladek et al. 2004; Stranger, Forrest et al. 2005; Dixon, Liang et al. 2007; Goring, Curran et al. 2007; Stranger, Forrest et al. 2007; Stranger, Nica et al. 2007; Dimas, Deutsch et al. 2009).

To date, most studies interrogating the genetic basis of regulatory variation have explored the effects of single variants on gene expression. Experiments in yeast however have revealed that the inheritance of over half of all transcripts is influenced by interacting locus pairs (Brem, Storey et al. 2005). Interactions between genetic factors have also been shown to occur in humans, for example in studies where the functional impact of coding variants is modified by regulatory variants nearby (Dimas, Stranger et al. 2008; Wang, Cruchaga et al. 2009). However, systematic measures of the extent of genetic interactions are lacking (Flint and Mackay 2009). Furthermore, although numerous large-scale studies have identified loci with a role in expression regulation, in most cases the candidate regions defined are broad and identification of true functional variants is pending. Finally, the bulk majority of studies to date have explored gene expression in a single cell type, usually in Epstein-Barr virus (EBV)-transformed B-cells (lymphoblastoid cell lines or LCLs), as these can be easily obtained from B-cells in blood samples and maintained in the laboratory. The extent of cell type specificity of gene expression (described in section 1.2.2) underscores the need to explore expression systematically in other cell types and catalogue cell type-specific regulatory variation. With the increasing realisation of the role of regulatory variation in shaping phenotypes in health and disease, detection and precise identification of single and interacting variants in multiple populations and cell types is a priority.

1.6 DETECTING REGULATORY VARIATION

Technological advances in the last decade, especially the development of microarray platforms, have made it possible to move from low and medium-throughput quantification of gene expression (e.g. reporter, or allele-specific expression assays (ASE)) to genome-wide quantification of mRNA levels. Transcript abundance for each of thousands of genes can be determined in a single experiment with mRNA intensity values reflecting mRNA levels. mRNA intensity exhibits continuous variation among individuals and mapping gene expression variation is a typical quantitative trait exercise (Stranger and Dermitzakis 2005; Dermitzakis and Stranger 2006). The rationale used to map quantitative trait loci (QTLs) for continuous phenotypes such as weight and height is also employed to detect expression QTLs (eQTLs) (Mackay, Stone et al. 2009). In human populations two approaches have been employed for eQTL mapping: linkage and association mapping (Dermitzakis and Stranger 2006; Gilad, Rifkin et al. 2008; Mackay, Stone et al. 2009).

1.6.1 Linkage mapping

Linkage mapping tracks the transmission of chromosomes through families using pedigrees and requires data on phenotypes and markers for each family member. The aim is to identify markers whose transmission patterns correlate with the phenotype, the implication being that these markers are linked to causal variants driving the phenotype (Gilad, Rifkin et al. 2008). The advantage of linkage mapping is that it requires a relatively low density of markers (<1,000 for microsatellites and slightly higher numbers for single nucleotide polymorphisms (SNPs)). However, it provides coarse-grained (low resolution) localisation, as it depends on the occurrence of recombination events within families for finer mapping (Gilad, Rifkin et al. 2008). Some of the first genome-wide studies on gene expression in humans employed a linkage

approach, using cell lines from individuals of Centre d'Étude du Polymorphisme Humain (CEPH) pedigrees (Monks, Leonardson et al. 2004; Morley, Molony et al. 2004). This approach is powerful when functional variants are rare and there is allelic heterogeneity (different mutations at the same locus that give rise to the same phenotype), as is the case for β-thalassaemia which can be caused by several different mutations in the *β-globin* gene (Dermitzakis and Stranger 2006). If the variants affecting gene activity are of smalle effect size (minor allele frequency (MAF) > 5%), linkage is relatively underpowered and association mapping performs better.

1.6.2 Association mapping

Association mapping identifies markers whose genotypes show a statistical association to the phenotype of interest (in this case mRNA abundance). A statistically significant association for a given marker implies that it is linked to a functional regulatory variant. In its simplest form, association mapping uses samples of unrelated individuals and dense genotyping data (e.g. 500,000 SNPs for a genome-wide study in humans). It is the most powerful method to date for the detection of common variants, provided that the causal sequences are in strong linkage disequilibrium (LD) with the genotyped SNPs (Dermitzakis and Stranger 2006). Additionally, with sufficiently dense genotyping, association mapping is more likely to detect variants with small or medium effect sizes (Gilad, Rifkin et al. 2008). Although this approach rarely detects true functional variants, the resolution provided is much higher compared to linkage, with functional variants mapping within hundreds of kb of associated markers depending on the extent of LD. One potential caveat of association mapping is the occurrence of false positives arising as a consequence of population structure, but this can be resolved using methods that correct for structure (Price, Patterson et al. 2006).

1.7 GENETIC VARIANTS TESTED IN ASSOCIATION STUDIES

An estimated 99.9% of the 6 billion nucleotides making up the human genome is identical across individuals (Sachidanandam, Weissman et al. 2001). The remaining 0.01% that varies between any two randomly chosen individuals consists of variation occurring on different scales and ranges from single base changes to alterations in copy number of larger segments. Genetic variants in the human genome include SNPs, insertion/deletion polymorphisms (indels), retroposon insertions, variation in the number of copies of a tandem repeat (mini and microsatellites), copy number variants (CNVs), inversions and variants that are a combination of some or all the above.

In this thesis genetic variation in the form of SNPs was associated with transcript levels to detect eQTLs. SNPs are the simplest and most common type of genetic variant, constituting roughly 75% of the total variation observed in humans (Levy, Sutton et al. 2007). They are the smallest unit of polymorphism and arise from the exchange of a single base in the DNA sequence (Hartl and Clark 2007). Traditionally, a DNA position is said to be polymorphic when alleles are found at a frequency between 1% and 99% in the population. The human genome is estimated to contain over ten million SNPs, seven million of which are designated as common (MAF ≥ 5% across the entire population) (Kruglyak and Nickerson 2001; Crawford, Akey et al. 2005). The International HapMap Consortium, launched in 2002 aimed to identify and catalogue these variants to quantify the extent of genetic similarities and differences between humans (International HapMap Consortium 2003). Currently, over four million SNPs in 1,301 individuals from eleven geographically distinct populations have been assayed (see section 2.1.1). Depending on their position in the genome, SNPs can be non-coding or coding. For the 1.5% of the genome that encodes proteins, the redundancy of the genetic code means that in some cases specific amino acids can be encoded by multiple codons. Synonymous SNPs are those base substitutions that do not alter the amino acid

sequence, while coding or non-synonymous SNPs (nsSNPs) are those that lead to a change of a single amino acid.

1.8 THESIS AIMS

Studies addressing the genetic component of gene expression have uncovered an abundance of common genetic variation influencing gene expression and have defined a field of intense study over the past few years. It is now well-established that regulatory polymorphisms are widespread in the human genome, with cis and trans-acting loci regulating transcript levels of genes. Most studies to date however have explored the effects of single genetic variants and have interrogated expression in a single cell type. Furthermore, although these studies have made a very important first step in detecting regions harbouring regulatory variants, few have identified precise functional variants. In this thesis I aim to further our understanding of regulatory variation by: a) exploring the effect of interactions between genetic variants on transcript levels (Chapter 3), b) dissecting the fine-scale architecture of the cis regulatory landscape (Chapter 4) and by c) exploring the extent of cell type specificity of regulatory variation (Chapter 5). Uncovering regulatory variation and understanding its function will help elucidate developmental programmes and patterns of cell type specificity and will also shed light on processes determining natural range and disease phenotypes.

2 MATERIALS AND METHODS

In this chapter I will:

- Describe the population samples analysed in this thesis.
- Define the sets of SNPs and genes tested for association.
- Introduce the statistical tests used for association of SNP genotype with mRNA levels, as well methods for significance correction.
- Outline the particulars of the three studies making up this thesis:
  - Impact of eQTL-nsSNP interaction on gene expression in cis and trans (Chapter 3)
  - Fine-scale architecture of the cis regulatory landscape (Chapter 4)
  - Cell type specificity of eQTLs (Chapter 5)

2.1 THE SAMPLES

The population samples studied in this thesis belong to two resources that have been set up to explore human genetic variation: the HapMap Project and the GenCord Project. In the following sections I give a brief outline of these resources.

2.1.1 The HapMap Project

The International HapMap Project was launched in 2002 as a collaborative effort to identify and catalogue genetic similarities and differences in human populations (International HapMap Consortium 2003). The ultimate goal of HapMap was to provide a public resource for medical genetic research by developing a detailed haplotype map (HapMap) of the human genome that would describe common patterns of genetic variation. The core strategy of this project involved genotyping DNA from LCLs

generated from blood samples of individuals belonging to a diverse set of populations. The project is ongoing and is currently in its third phase. (International HapMap Consortium 2003; International HapMap Consortium 2005; International HapMap Consortium 2007).

The aim of HapMap Phase 1 was to genotype at least one common SNP per five kb across the euchromatic portion of the genome of 269 individuals from four geographically distinct populations. The individuals genotyped were: 30 mother–father–adult child trios of northern and western European ancestry living in Utah from the CEPH collection (abbreviated CEU), 45 unrelated Han Chinese individuals in Beijing, China (CHB), 44 unrelated Japanese individuals in Tokyo, Japan (JPT) and 30 trios from the Yoruba in Ibadan, Nigeria (YRI). Approximately 1.3 million SNPs were genotyped per population and a detailed description of this resource was published in 2005 (International HapMap Consortium 2005).

In Phase 2, a further 2.1 million SNPs were genotyped in each of 270 individuals (Phase 1 individuals and an additional sample from the JPT population). The resulting HapMap had a SNP density of approximately one SNP per kb and was estimated to contain approximately 25–35% of all common SNPs (9-10 million SNPs with a MAF ≥ 0.05) in the assembled human genome. A description of this resource was published in 2007 (International HapMap Consortium 2007).

Phase 3 of the HapMap was ongoing at the time of writing and involved additional individuals from the four initial populations, as well as seven additional populations. Over 4 million SNPs were genotyped for 541 individuals of the four initial populations (CEU, CHB, JPT, YRI) and approximately 1.5 million SNPs were genotyped in 760 individuals of seven new populations (90 ASW: African ancestry in Southwest USA; 100 CHD: Chinese in Metropolitan Denver, Colorado, USA; 100 GIH: Gujarati Indians in Houston, Texas, USA; 100 LWK: Luhya in Webuye, Kenya; 90 MEX: Mexican ancestry in Los Angeles, California, USA; 180 MKK: Maasai in Kinyawa, Kenya; 100

TSI: Toscans in Italy). At the time of writing, this study was in preparation for publication.

Table 2 summarizes SNPs and individuals assayed in each of the three phases of the project. Data analysed in this thesis include all four HapMap Phase 2 populations (210 unrelated individuals from CEU, CHB, JPT and YRI) in the study described in Chapter 3. Additional samples for those populations as well as four of the seven new HapMap Phase 3 populations (792 unrelated individuals from CEU, CHB, GWK, JPT, LWK, MEX, MKK and YRI) were analysed in the study described in Chapter 4.

		Populations										
		CEU	CHB	JPT	YRI	ASW	CHD	GIH	LWK	MEX	MKK	TSI
Phase 1	SNPs	1,105,063	1,087,365	1,087,365	1,076,442							
	Individuals	90	45	44	90							
Phase 2	SNPs	3,904,218	3,936,482	3,936,482	3,846,092							
	Individuals	90	45	45	90							
Phase 3	SNPs	4,030,562	4,052,129	4,052,216	3,984,146	1,561,382	1,306,152	1,407,818	1,529,693	1,410,231	1,537,561	1,419,861
	Individuals	180	90	91	180	90	100	100	100	90	180	100

Table 2. Summary of SNPs and individuals assayed in each of the three phases of HapMap. Population descriptors: CEU: Utah residents with Northern and Western European ancestry from the CEPH collection; CHB: Han Chinese in Beijing, China; JPT: Japanese in Tokyo, Japan; YRI: Yoruban in Ibadan, Nigeria; ASW: African ancestry in Southwest USA; CHD: Chinese in Metropolitan Denver, Colorado, USA; GIH: Gujarati Indians in Houston, Texas, USA; LWK: Luhya in Webuye, Kenya; MEX: Mexican ancestry in Los Angeles, California, USA; MKK: Maasai in Kinyawa, Kenya; TSI: Toscans in Italy.

2.1.2 The GenCord Project

The GenCord project is a collection of cell lines derived from umbilical cords of 85 individuals of Western European origin, following appropriate consent and ethical approval (Dimas, Deutsch et al. 2009). The project was conceived as a resource for the identification of QTLs involved in the regulation of cellular phenotypes in primary

fibroblasts, LCLs and primary T-cells. Umbilical cord was chosen because it is readily available and allows the acquisition of multiple cell types for each individual. Sample collection was performed systematically on full term or near full term pregnancies to ensure homogeneity for sample age.

2.1.3 Using HapMap and GenCord to investigate regulatory variation

HapMap and GenCord were used to investigate the impact of genetic variation on expression levels within and across human populations, but also across cell types. Statistical methods were used to associate SNP genotypes with mRNA levels (see sections 2.4 and 2.5), and experimental methods were subsequently employed for the biological verification of a subset of predicted associations (see sections 2.6.5 and 2.8.3).

For expression association studies using the HapMap populations, publicly available genotype data were combined with expression data generated by our group at the Wellcome Trust Sanger Institute (WTSI) for the same set of individuals. This analytical set up made it possible to explore how genetic variation shapes gene expression differences within and across populations, chiefly as a consequence of allele frequency differences. The HapMap Project was launched in 2002 and there have been a number of data release stages over the past few years. As a consequence, analyses performed in this thesis used two different releases of HapMap data: a) Phase 2 data were used to explore the impact of eQTL-nsSNP interactions on gene expression in cis and trans (Chapter 3) and b) Phase 3 data were used to investigate the fine-scale architecture of the cis regulatory landscape (Chapter 4). HapMap genotype data are publicly available at http://www.hapmap.org and expression data generated by our group for these populations are available at: ftp://ftp.sanger.ac.uk/pub/genevar/.

The GenCord study design involves expression quantification in each of three cell types separately, and genotyping using DNA from a single cell type (LCLs). This

analytical set up made it possible to address how genetic variation shapes gene expression differences within a population and across cell types, as a consequence of the cell type-dependent action of genetic variation (Chapter 5). The chief advantage of GenCord is that is allows direct comparisons to be made across cell types, as samples were collected and processed in a systematic way. GenCord was also used to explore the fine-scale architecture of the cis regulatory landscape in a cell type-specific context (Chapter 5). Expression data are available at ftp://ftp.sanger.ac.uk/pub/genevar/ and in the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/ accession number GSE17080).

The datasets used in each of the three studies are outlined in Table 3. In the following sections I describe how SNP genotype (section 2.2) and gene expression (section 2.3) information was obtained or generated, I present the general statistical methods used for detection of variants associated with gene expression (sections 2.4 and 2.5) and outline the specific analyses carried out for each study.

		HapMap Phase 2	HapMap Phase 3	GenCord
eQTL - nsSNP interaction	(Chapter 3)	**X**		
eQTL fine-scale architecture	(Chapter 4)		**X**	**X**
eQTL cell type specificity	(Chapter 5)			**X**

Table 3. Overview of datasets analysed in each of the three studies presented in this thesis. (Note that the eQTL fine-scale architecture study is outlined in Chapter 4 for HapMap, but results using the same strategy are also presented in Chapter 5 for GenCord).

2.2 THE SNPS

HapMap SNP genotypes were generated by the International HapMap Consortium and are publicly available at www.hapmap.org. GenCord SNP genotypes were generated

by our collaborators in the Department of Genetic Medicine and Development, at the University of Geneva Medical School (UGMS).

2.2.1 HapMap Phase 2

Phase 2 of the HapMap involved genotyping of nearly four million SNPs in each of 270 individuals from CEU, CHB, JPT and YRI populations. HapMap version 21 (NCBI Build 35) SNPs were used to interrogate the interaction between functional variants, namely that between cis eQTLs (i.e. tags of regulatory variants) and nsSNPs (protein-coding SNPs). This study is described in Chapter 3.

2.2.1.1 Cis eQTLs

Cis eQTLs were identified in a genome-wide association study (GWAS) by Stranger et al (2007) as Phase 2 HapMap SNPs (mapping in a 2 Mb window centred on the expression probe midpoint) that showed a statistically significant association with mRNA levels at the 0.01 permutation threshold (see section 2.6.1).

2.2.1.2 nsSNPs

nsSNPs are protein-coding variants that result in a single amino acid substitution in the protein product. The strategy used to select nsSNPs for this study is summarized in Figure 4 and involved the following steps: rsIDs and coordinates for all known nsSNPs were downloaded from Biomart (http://www.biomart.org/biomart/martview), dbSNP125 (NCBI Build 36), and HapMart (version 21 NCBI Build 35) (http://hapmart.hapmap.org/BioMart/martview). The distribution of nsSNPs in genes was interrogated for three gene collections created using different annotation methods (Brent 2005; Flicek 2007): CCDS, Ensembl and RefSeq genes. CCDS genes are those genes for which structure has been agreed upon by NCBI, Ensembl, and UCSC. Ensembl genes are to a certain extent annotated automatically, whereas Refseq gene

annotation is largely manual. The Refseq collection was chosen as it represents a set of genes with explicitly linked nucleotide and protein sequence and a large enough number of genes to work with. Refseq gene IDs and coordinates were downloaded from the UCSC genome browser (http://genome.ucsc.edu/) and genes mapping on chromosomes X and Y, as well as those without coordinate information were removed. nsSNPs of all frequencies were subsequently mapped on Refseq genes using nsSNP and gene coordinates.

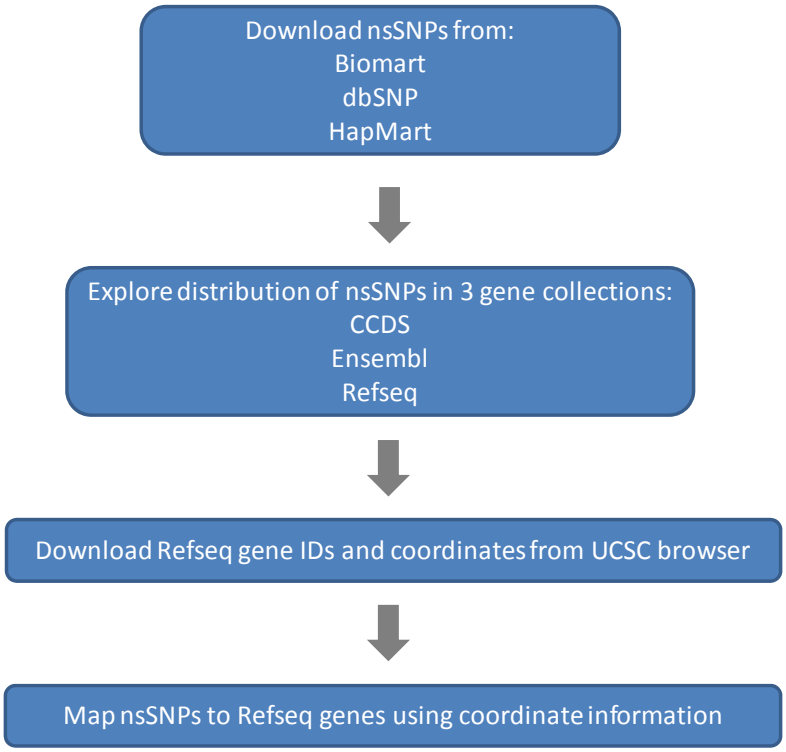


Figure 4. Strategy employed to select nsSNPs for the eQTL-nsSNP interaction study described in Chapter 3.

2.2.2 HapMap Phase 3

In Phase 3 of the HapMap, over 4 million SNPs were genotyped in the initial four populations and 1.5 million SNPs were genotyped in the seven additional populations.

Out of a total of 1,301 individuals genotyped, I used HapMap version 27 (NCBI Build 36) SNPs from 792 individuals from the CEU, CHB, GWK, JPT, LWK, MEX, MKK, and YRI populations to investigate the fine-scale architecture of the regulatory regions around genes, in the study described in Chapter 4.

2.2.3 GenCord

Approximately half a million SNPs were genotyped in the 85 individuals of GenCord. DNA samples were extracted from cord tissue LCLs with the Puregene cell kit (Gentra-Qiagen, Venlo, The Netherlands). Genotyping was performed using the illumina 550K SNP array (illumina, San Diego, California, USA) following the instructions of the manufacturers (Figure 5). This work was carried out by Samuel Deutsch at the UGMS. Principal component analysis (PCA) was performed on the genotype data to detect potential outliers. This analysis was carried out by Stephen Montgomery at the WTSI.

2.3 The Genes

Transcript levels in HapMap LCLs and in the three cell types of GenCord were quantified using gene expression arrays at the WTSI. All data generated are publicly available at http://www.sanger.ac.uk/Software/Genevar. GenCord data are also available on the GEO (section 2.1.3).

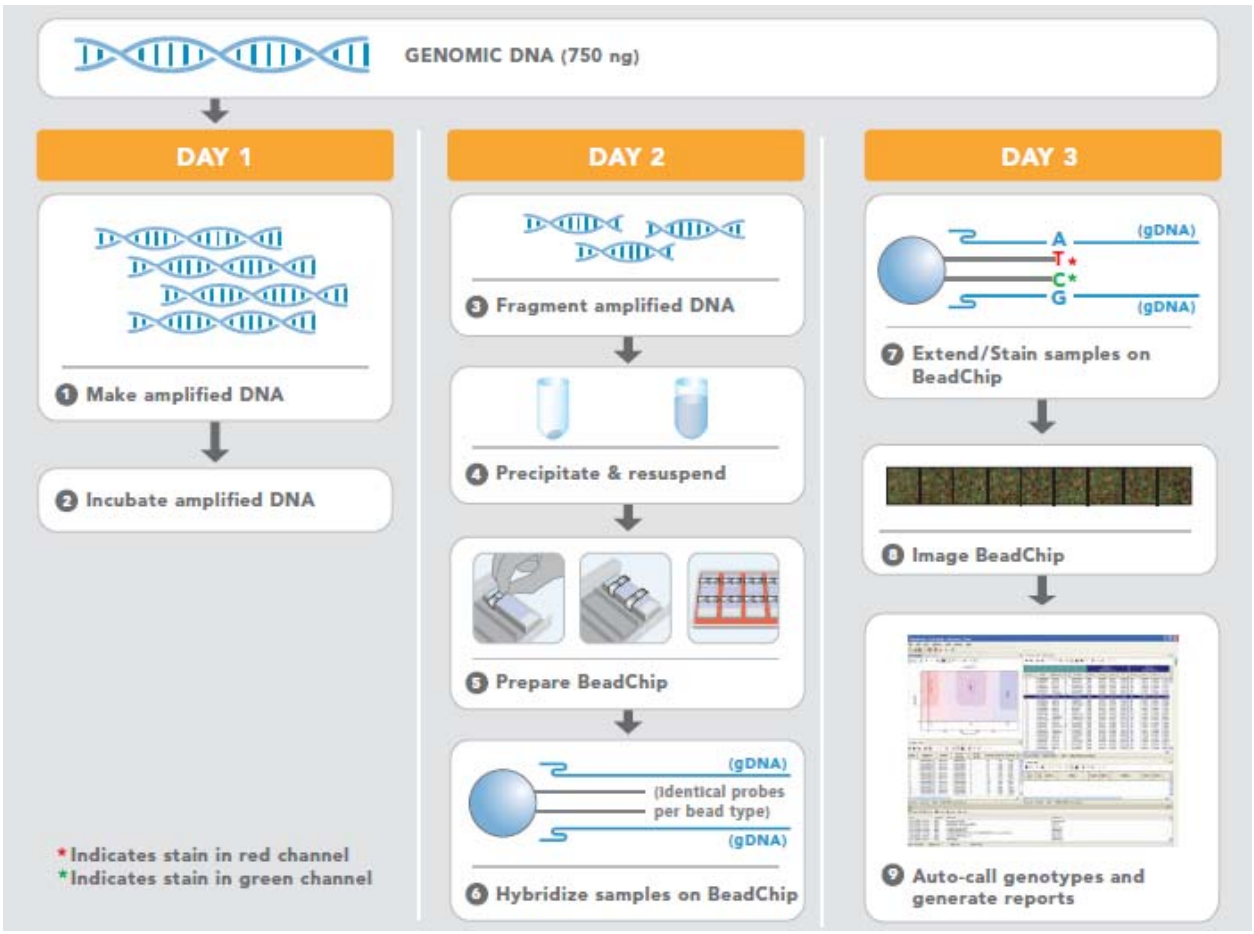


Figure 5. Infinium SNP genotyping assay. The DNA sample used for this assay is isothermally amplified overnight (**Steps 1** and **2**). This amplification has no appreciable allelic partiality. Approximately 750 ng of DNA is used to assay 500,000 SNP loci and the amplified product is fragmented by a controlled enzymatic process (**Step 3**). After alcohol precipitation and resuspension of the DNA (**Step 4**), the BeadChip is prepared for hybridization in the capillary flow-through chamber (**Step 5**); samples are applied to BeadChips and incubated overnight. The amplified and fragmented DNA samples anneal to locus-specific 50-mers (covalently linked to one of over 500,000 bead-types) during the hybridization step (**Step 6**). One bead type corresponds to each allele per SNP locus. After hybridization, allelic specificity is conferred by enzymatic base extension. Products are subsequently fluorescently stained (**Step 7**). The intensities of the beads' fluorescence are detected by the illumina BeadArray Reader (**Step 8**), and are in turn analysed using illumina's software for automated genotype calling (**Step 9**). Figure and assay description from http://www.illumina.com/.

2.3.1 HapMap Phase 2

2.3.1.1 RNA preparation, gene expression quantification and normalization

Total RNA was extracted from LCLs of the 210 unrelated individuals of the HapMap Phase 2 (Coriell, Camden, New Jersey, USA). For each RNA extraction, two one-quarter scale Message Amp II reactions (in vitro transcription reactions or IVTs) (Ambion, Austin, Texas, USA) were performed using 200 ng of total RNA, to produce cRNA. To assay transcript levels, 1.5 μg of the cRNA was hybridized to illumina's commercial whole genome expression array, Sentrix Human-6 v1 Expression BeadChip (Kuhn, Baker et al. 2004). These arrays utilize a bead pool containing ~48,000 unique bead types (one for each of 47,294 transcripts, plus controls), each with several hundred thousand gene-specific 50mer probes attached (Figure 6). Six arrays were run in parallel on a single BeadChip. Each bead type (probe) is present on a single array on average 30 times. Each of the two IVT reactions from the 210 samples was hybridized to two arrays each, so that each cell line had four replicate hybridizations. cRNA was hybridized to arrays, labelled with Cy3-streptavidin (Amersham Biosciences, Little Chalfont, UK) and scanned with a Bead Station (illumina). This work was carried out by Catherine Ingle at the WTSI.

With the illumina bead technology, a single hybridization of RNA from one cell line to an array produced approximately 30 intensity values for each of 47,294 bead types. These background-corrected values for a single bead type were summarized by illumina software and output to the user as a set of 47,294 intensity values for each individual hybridization. In this experiment, each cell line was hybridized to four arrays, resulting in four reported intensity values (as averages of the values from the 30 beads per probe) for each of the 47,294 bead types. To combine information from replicate hybridizations, raw data were read using the *Beadarray R* package (Dunning, Smith et al. 2007) and normalized on a $\log_2$ scale using a quantile normalization method

(Bolstad, Irizarry et al. 2003) across replicates of a single individual, followed by a median normalization method across individuals of a single population. These normalized values (for each probe, across replicates for each individual) were used in subsequent analyses. Normalization was carried out by Mark Dunning and Simon Tavaré at the Cancer Research UK Cambridge Research Institute (CRI).

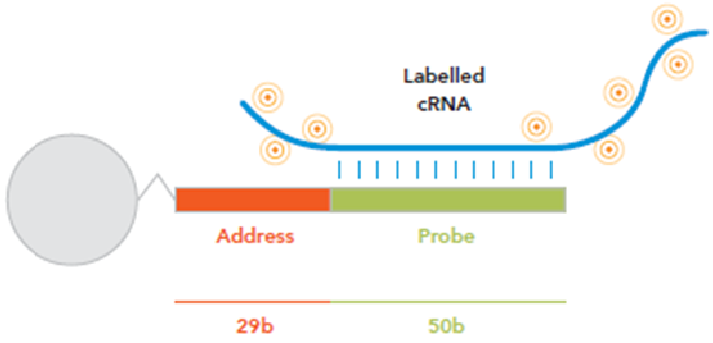


Figure 6. Gene expression probe. Gene expression probes are attached to beads, which are then assembled into arrays. For simplicity, this figure shows only one oligomer attached to the bead; actual beads have hundreds of thousands of copies of the same sequence attached. Figure and description from http://www.illumina.com/.

2.3.1.2 Selection of variable probes

To ensure variability in the gene expression phenotype, the intersection of the top 18,000 most variable probes in each of the four populations was selected from the 47,294 probes, resulting in a set of 13,797 probes. An additional set of probes with large differences in rank variability between populations was also selected by ranking all transcripts by variability within each population and making all pairwise comparisons between populations to quantify difference in rank between population pairs. The top 1% of transcripts with largest absolute value rank difference from each population pair comparison were selected. The union of these lists provided an additional 2,021 probes. Probes mapping to chromosomes X and Y, as well as those mapping to the mitochondrion genome were discarded. Probes with no match in the human genome

Build 35 were also removed, as were 469 probes that contained SNPs in their sequence. This resulted in a subset of 14,456 probes (mapping to 13,643 unique autosomal genes) that were highly variable within and between populations and were used for association analysis (Chapter 3). Variable probe selection was carried out by Barbara Stranger and Manolis Dermitzakis at the WTSI.

2.3.2 HapMap Phase 3

2.3.2.1 RNA preparation, gene expression quantification and normalization

Total RNA was extracted from LCLs of the 792 unrelated individuals of the HapMap Phase 3 (Coriell). Gene expression (mRNA levels) was quantified using illumina's commercial whole genome expression array, Sentrix Human-6 Expression BeadChip version 2 (~48,000 transcripts interrogated; illumina) as described previously (in this case only two IVTs were performed). This work was carried out by Catherine Ingle, James Nisbett, and Magdalena Sekowska at the WTSI.

Hybridization intensity values were normalized on a $\log_2$ scale using a quantile normalization method (Bolstad, Irizarry et al. 2003) across all replicates of a single individual followed by a median normalization method across individuals of a single population. GIH, LWK, MEX and MKK populations were normalized for admixture using a customized version of *Eigenstrat* which outputs principal component adjustments for expression data (Price, Patterson et al. 2006). Expression values were adjusted using ten primary axes of variation from intra-population PCA and these normalized expression values were used as input for the association analysis. Normalization and PCA correction were performed by Stephen Montgomery at the WTSI.

2.3.2.2 Probe selection

illumina's Sentrix Human-6 Expression BeadChip version 2 array covers over 24,000 unique, curated genes from the Refseq collection, as well as genes for which annotation is less well-established. In this case, probes were not filtered for expression variability, but were restricted to those corresponding to Refseq genes. SSAHA (Sequence Search and Alignment by Hashing Algorithm) (Ning, Cox et al. 2001), an algorithm for very fast matching and alignment of DNA sequences, was used to map probes on the Ensembl genes using the Ensembl Application Programme Interface (API) (http://www.ensembl.org/info/data/api.html Ensembl 49 NCBI Build 36). It was found that 22,512 probes mapped to 19,862 Ensembl genes and depending on the number of transcripts, some genes were covered by multiple probes. Conversely, a subset of probes mapped to more than one Ensembl gene and were discarded, as were probes mapping on chromosomes X and Y. Following filtering, a non-redundant total of 21,800 probes (corresponding to 18,226 Ensembl genes) was used for association analysis (Chapter 4). Mapping of probes on Ensembl genes using the Ensembl API was carried out with the help of Nathan Johnson at the European Bioinformatics Institute (EBI).

2.3.3 GenCord

2.3.3.1 GenCord sample collection

Umbilical cords were collected from 85 newborns of Western European origin born at the maternity ward of the University of Geneva Hospital, for which pregnancies were full term or near full term (38-41 weeks). For each sample, informed consent was obtained after an interview of the mother with a trained nurse and the project was approved by the University of Geneva Hospital Ethics Committee. From each umbilical cord three cell types were derived: 1) primary fibroblasts, 2) LCLs and 3) phytohemagglutinin (PHA) stimulated primary T-cells. In addition, total buffy coat was

frozen in RPMI medium (Invitrogen, Carlsbad, California, USA) 10% DMSO (Sigma, St. Louis, Missouri, USA), 20% FCS (Invitrogen) for future studies. This work was carried out by Samuel Deutsch at the UGMS.

2.3.3.2 GenCord cell line preparation

Cord blood was collected in 50 ml falcon tubes containing 10 ml of anti-coagulants (Sodium citrate and EDTA, Sigma) and kept at 4°C for less than 24 hours prior to treatment. For separation, cord blood was diluted two-fold in PBS (Invitrogen), layered on Ficoll-Paque (GE Healthcare Lifesciences, Chalfont St. Giles, UK) and centrifuged for 30 minutes at 800g. The mononuclear cell layer was removed, washed twice in 40 ml of PBS and re-suspended in 1 ml of RPMI 20% FCS, 1% antibiotics (Amimed, Basel, Switzerland).

For fibroblast preparation, cord tissue was finely cut under sterile conditions in 1 ml DMEM 10% FCS, 1% antibiotics (Amimed), transferred to a T25 flask and cultured upside-down for 12 hours to allow cells to attach to the surface of the flask. Flasks were turned around and left for approximately one week until fibroblast clusters appeared. Fibroblasts were then expanded with standard procedures. For preparation of LCLs, 300 µl of re-suspended cells and 100 µl of EBV were transferred to a 24-well plate well and cultured in an incubator at 37°C, 5 % CO_2. Fresh medium was added and replaced every 2-3 days. Cells were kept in culture for no less than 21 days prior to freezing. For PHA stimulated T-cell preparation, re-suspended mononuclear cells were diluted to a concentration of 1 x 106 cells/ml in RPMI (Invitrogen) with 5 µg/ml of PHA (Sigma), and cultured for five days with 2/3 medium replacement after 2.5 days. A subset of samples was characterized by flow cytometric analysis for expression of CD3, CD25 and CD69 (Becton Dickinson, Franklin Lakes, New Jersey, USA) revealing a homogenous activated T-cell population.

RNA from each cell type was prepared with RNeasy columns with on-column DNAse treatment (Qiagen, Venlo, The Netherlands), quantified with NanoDrop (Thermo Scientific, Waltham, Massachusetts, USA) and analyzed with a 2100 Bioanalyzer (Agilent, Santa Clara, California, USA). This work was carried out by Samuel Deutsch at the UGMS.

2.3.3.3 RNA preparation, gene expression quantification and normalization

Total RNA was extracted from fibroblasts, LCLs, and T-cells of the 85 unrelated individuals of the GenCord as described above. Two one-quarter scale Message Amp II reactions (Ambion) were performed for each RNA extraction with 200 ng of total RNA. 1.5 μg of cRNA was hybridized to illumina's WG-6 v3 Expression BeadChip array to quantify transcript abundance as described previously. In total there were two technical replicates (labelling and hybridization) for each RNA sample. This work was carried out by Catherine Ingle, James Nisbett, and Magdalena Sekowska at the WTSI. Intensity values were $\log_2$ transformed and normalized independently for each cell type using quantile normalization for sample replicates, and median normalization across all individuals. Each cell type was renormalized using the mean of the medians of each cell type expression values. Normalization was carried out by Stephen Montgomery at the WTSI.

2.3.3.4 Probe selection

The illumina WG-6 v3 Expression BeadChip array covers over 27,000 unique coding transcripts belonging to the Refseq collection. For the majority of these transcripts annotation is well-established, with approximately 7,000 transcripts having provisional annotation. This array also covers non-coding transcripts, as well as experimentally confirmed mRNA sequences aligning to EST clusters. Only probes corresponding to transcripts with good or provisional annotation (Refseq genes) were selected for

association testing. A total of 36,156 probes with Refseq IDs were queried for their corresponding Ensembl gene IDs in Biomart (Ensembl 50, NCBI Build 36). Of these, 23,805 probes had a corresponding Ensembl gene ID and after discarding probes mapping to chromosomes X and Y, as well as those that mapped to more than one Ensembl genes, 22,651 probes (corresponding to 17,945 RefSeq genes and 15,596 Ensembl genes) were used for subsequent analysis (Chapter 5).

2.4 ASSOCIATION TESTS

Additive linear regression (LR) and Spearman rank correlation (SRC) were used to test for association in cis between SNP genotypes and expression levels of genes. For each gene, variants mapping in a 2 Mb window centred on the TSS were tested for association (cis eQTLs used in the eQTL-nsSNP study described in Chapter 3 were identified in a previous study (see 2.2.1.1) that defined cis eQTLs as variants mapping in a 2 Mb window centred on the probe midpoint). This 2 Mb window defines the genomic region tested for cis association with gene expression. Particular tests and analyses conducted for each of the three studies are discussed in the relevant sections of this chapter.

2.4.1 Additive linear regression

A main effects additive LR model was used to test for association between SNP genotype and probe expression levels. The additive effect of a SNP genotype was tested by coding the genotypes at each locus as 0, 1 and 2 corresponding to counts of alphabetically sorted alleles in each genotype (e.g. counting the number of G alleles for a A/G SNP: AA = 0, AG = 1, GG = 2). Normalized $\log_2$ expression was regressed on SNP genotypes for each gene, and the following additive model was fitted: the genotype X_i

of individual i at the given SNP may be classified as one of three states $X_i = 0$, 1 or 2. The linear regression fitted was:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

Where Y_i is the normalized $\log_2$ expression levels of the probe for individual i, i=1…n and ε_i are independent normally distributed random variables with mean 0 and constant variance (Stranger, Forrest et al. 2005). The nominal parametric p-value of the test of no association (i.e. $b_1 = 0$), the slope, and r^2 for each SNP-probe pair were reported (Figure 7). LR however is sensitive to outlier effects and for this reason association tests were also carried out using SRC. SRC performs at a level equivalent to LR, detecting 77% - 86% of the associations uncovered by LR (Stranger, Nica et al. 2007).

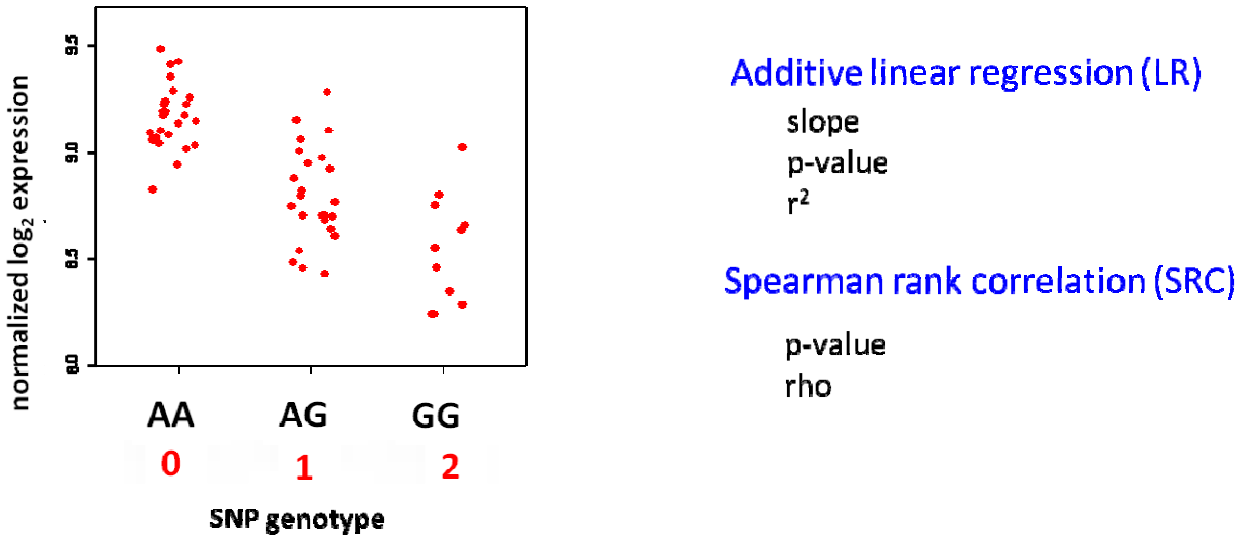


Figure 7. Statistical tests employed to associate SNP genotype with normalized $\log_2$ mRNA intensity levels. SNP genotypes were coded as 0, 1 or 2 (in this case corresponding to AA, AG and GG respectively). Linear regression (LR) was used to test the additive effects of SNP genotype on mRNA intensity (expression levels). The slope, the p-value and r^2 of the test were reported. To avoid outlier effects that affect LR output, Spearman rank correlation (SRC), a non-parametric test, was also used. The p-value and the correlation coefficient (rho) were reported.

2.4.2 Spearman rank correlation

SRC was also used to test for association between SNP genotypes and probe expression levels. SRC is a non-parametric measure of correlation that assesses how well an arbitrary monotonic function describes the relationship between two rank-ordered variables, without making any other assumptions about the particular nature of the relationship between these variables. Variables are initially converted into ranks (in this case the lower expression values are assigned lower ranks) and a correlation analysis is performed. When two observations are equal (tied) the average rank is used. SRC yields a statement of the degree of interdependence of the scores of the two variables, the Spearman correlation coefficient or rho. Rho describes the strength and direction of the correlation. The nominal p-value for the test of no association and rho were reported.

2.5 MULTIPLE TEST CORRECTION

To assess significance of association between SNP genotype and probe expression levels, 10,000 permutations of each expression phenotype relative to the genotypes were performed for each gene (Churchill and Doerge 1994; Doerge and Churchill 1996; Stranger, Forrest et al. 2005; Stranger, Nica et al. 2007) (Figure 8). For each round of permutations, the minimal permuted p-value was reported and a distribution of 10,000 minimum permuted p-values was generated. An association to gene expression was considered significant if the nominal p-value from the association test (observed p-value) was lower than the 0.5, 0.01, 0.001 and 0.0001 tail of the distribution of the minimal permuted p-values, defining four permutation significance thresholds. For each gene, the most stringent p-value was retained.

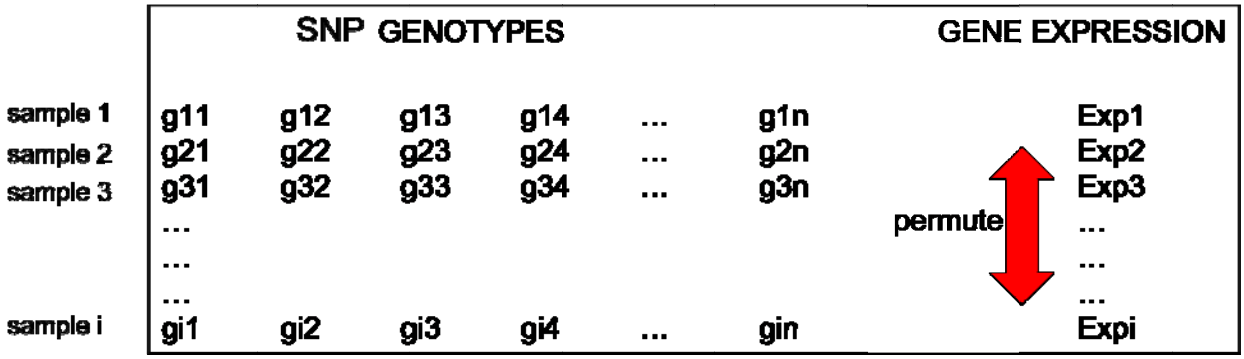


Figure 8. Significance levels for each gene were determined through permutations. 10,000 permutations of expression levels (Epx1, Exp2, Exp3…Expi) relative to genotypes (g11, g21, g31…gi1) were performed for each gene and for all individuals (samples) in the single population analysis. The minimal permuted p-value obtained for each round of permutations was used to generate a distribution of minimal permuted p-values for each gene. Four significance thresholds were defined at the 0.5, 0.01, 0.001 and 0.0001 tails of the distribution. Significant associations of SNP genotype to expression levels were those for which the association test p-value (observed p-value) was lower than the selected permutation significance threshold.

2.6 eQTL-nsSNP interaction study (Chapter 3)

2.6.1 The interaction model

HapMap Phase 2 data were used to explore interactions between regulatory and protein-coding variants and their impact on gene expression in cis and trans. The regulatory variants tested for interaction were cis eQTLs (MAF ≥ 0.05) identified in a previous study (Stranger, Nica et al. 2007) located within a 2 Mb window centred on the probe midpoint. The protein-coding variants tested were nsSNPs mapping in Refseq genes.

The model of interaction brings together quantitative and qualitative variation as follows: a gene for which a cis eQTL has been detected will be expressed at different quantities among individuals in the population (Pastinen and Hudson 2004; Stranger, Nica et al. 2007) (Figure 9 a). On the other hand, genes containing nsSNPs give rise to

protein products that differ in quality by a single amino acid (Figure 9 b). In the case where a gene with an identified cis eQTL also contains an nsSNP, the resulting protein products will differ not only in quantity, but also in quality (amino acid sequence) among individuals (Figure 9 c, also see Figure 14). The co-existence of these two variant types may have cis and trans effects on gene expression. In cis, the eQTL (or rather the regulatory element tagged by the eQTL) can modify (magnify or mask) the functional effect of the nsSNP. This is a cis modification effect and nsSNPs harboured in genes with varying expression levels are hereon termed DE (differentially expressed). In trans, the different protein ratios arising from modification in cis may affect their downstream targets, leaving an imprint on genome-wide expression levels. This impact in trans is a true epistatic effect and can be explored using the specific and testable biological model presented.

The proposed model is centred on the concept of DE nsSNPs, and two strategies were employed to detect these variants. The first strategy involved scanning all genes with cis eQTLs for nsSNPs. The second strategy involved direct association testing of nsSNP genotype with expression levels of the gene it is harboured in. In this second case, the nsSNP can act as an eQTL for its own gene's expression levels. To summarize, an nsSNP is DE if: 1) it maps in a gene for which at least one cis eQTL has been identified or 2) it shows a significant association with its own gene's expression levels. The nsSNP shown in Figure 9 c is DE.

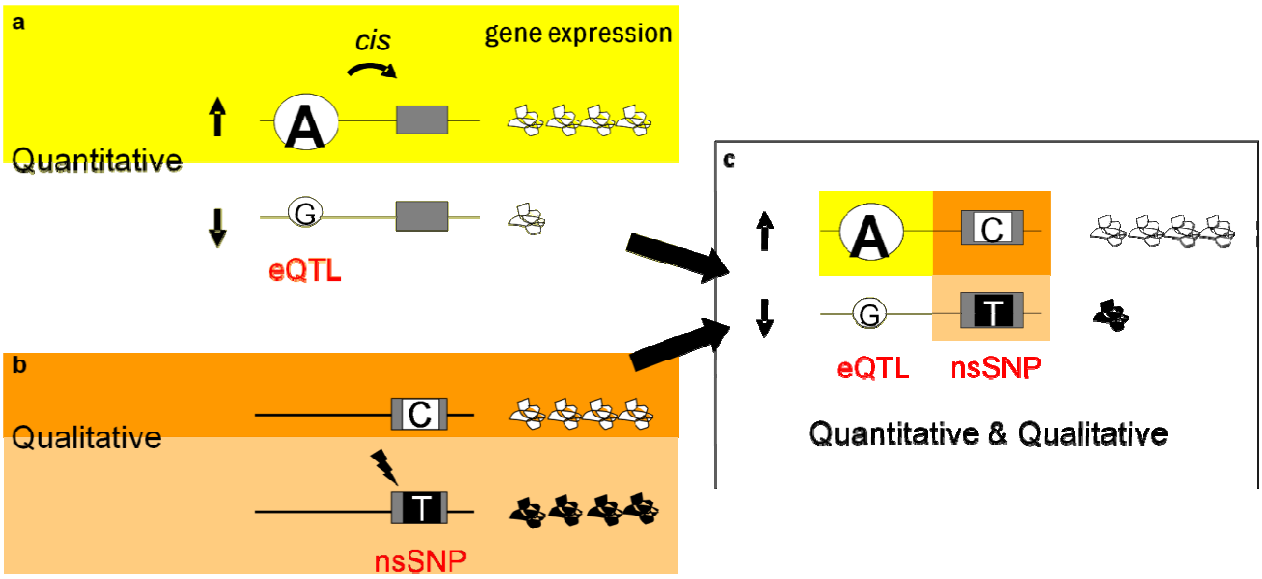


Figure 9. eQTL-nsSNP interaction. a) Genes with identified cis eQTLs have quantitative differences in their expression (high vs. low expression levels). **b)** Genes that possesses an nsSNP give rise to protein products that differ qualitatively by a single amino acid (white vs. black protein product). **c)** If a gene possesses both a cis eQTL and an nsSNP, the resulting protein products will differ in quantity and quality. This is an example of an interaction in cis, where the functional effect of the nsSNP is modified by the cis eQTL. Furthermore, if this gene has downstream targets, their expression may be influenced through a trans effect on gene expression. See also Figure 14 in Chapter 3.

2.6.2 Single population nsSNP association test

One way to determine whether an nsSNP is DE is to perform a direct association test of nsSNP genotype and expression levels of the gene harbouring it. LR was used to test for association in cis between: 1) nsSNP genotypes (for nsSNPs with MAF ≥ 0.05) for the unrelated individuals of each HapMap Phase 2 population (60 CEU, 45 CHB, 45 JPT, and 60 YRI) and 2) normalized $\log_2$ quantitative gene expression measurements for the same individuals. Association testing was performed for each population separately and significance thresholds for each gene were assigned through permutations of expression values relative to genotypes. An association with gene expression was

considered significant if the nominal p-value from LR was lower than the 0.01 tail of the distribution of minimal permuted p-values. Correction for false positives was carried out by calculating the ratio of expected false positives at a given threshold over the number of significant associations at the same threshold (this is an approximation of the false discovery rate (FDR)).

2.6.3 Multiple population nsSNP association test

To increase power of association detection I combined data (SNP genotypes and normalized expression values) for unrelated individuals of multiple populations and repeated association testing. Three different multiple population comparison panels were compiled: 1) CEU-CHB-JPT-YRI, 2) CEU-CHB-JPT and 3) CHB-JPT. Association tests were carried out for each population panel separately using LR. In this case, correction for significance was through conditional permutations (Figure 10) whereby the correlated structure of gene expression values within each population was retained by randomizing data within each population (Stranger, Nica et al. 2007). This approach accounts for population differentiation and prevents detection of spurious associations. For each of the 14,456 probes in each multiple population panel, expression values were permuted among individuals of a single population followed by regression analysis of the grouped multi-population expression data against the grouped multi-population permuted nsSNP genotypes. Four significance thresholds were selected (0.05, 0.01, 0.001, 0.0001) and an association to gene expression was considered significant if the nominal p-value from the linear regression test was lower than the 0.01 tail of the distribution of minimal permuted p-values. Correction for false positives was carried out as described in section 2.6.2.

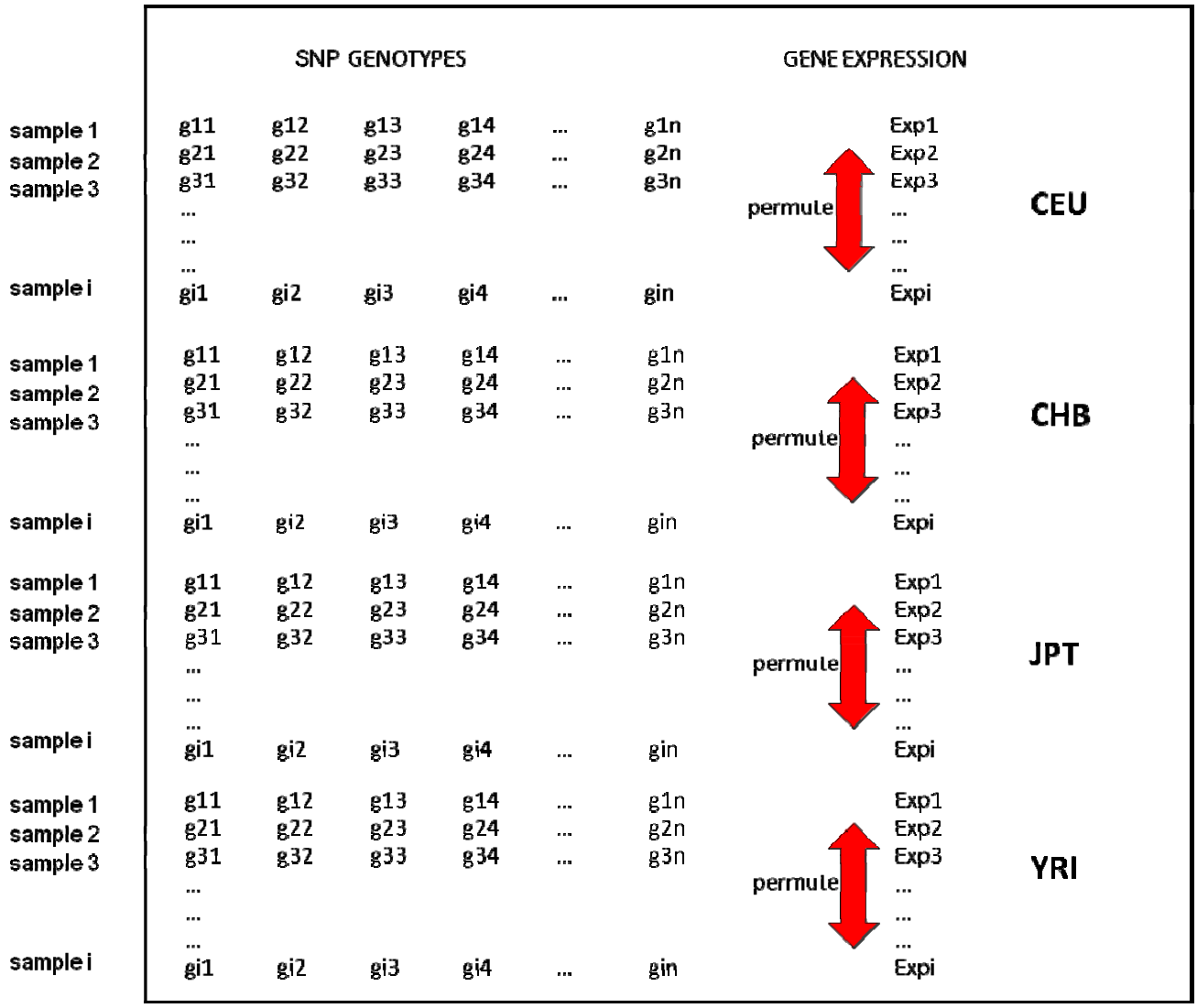


Figure 10. Conditional permutations were used to determine significance levels for each gene in the multiple population analysis. Ten thousand permutations of expression levels (Exp1, Exp2, Exp3…Expi) relative to genotypes (g11, g21, g31…gi1) were performed for each gene, in an approach where the correlated structure of expression values within each population was retained. This was achieved by randomizing expression data within each population. This approach accounts for population differentiation and prevents detection of spurious associations.

2.6.4 eQTL-nsSNP linkage disequilibrium analysis

Differential expression of nsSNPs is most likely driven by regulatory variants tagged by eQTLs that are in LD with the nsSNP. To address this, I explored the distribution of r^2 (a

measure of statistical correlation between alleles at two loci) (Hartl and Clark 2007) for eQTL-nsSNP pairs in which the nsSNP: a) showed a significant association with its own gene's expression levels and b) showed no such association. LD values were calculated by a pairwise estimation, for eQTLs and nsSNPs genotyped in the same individuals, and that mapped within a 100 kb window of each other (Ensembl 46). LD values were calculated by Daniel Rios at the EBI. The distributions of r^2 estimates for eQTLs-nsSNP pairs with and without an associated nsSNP were compared using a Mann-Whitney (M-W) test. Significant results were those for which M-W p-value ≤ 0.05.

2.6.5 Allele-specific expression assay

2.6.5.1 DNA and RNA preparation for allele-specific expression assays

The association tests employed make predictions about DE nsSNPs. ASE assays were used for the biological verification of a subset of these predictions (Figure 11). Genomic DNA (gDNA) and total RNA were extracted from LCLs of the unrelated CEU and YRI HapMap individuals (Coriell) using Qiagen's AllPrep kit. RNA was treated with Turbo DNA-free (Ambion) to minimize gDNA contamination. The RNA was concentrated and further cleaned with RNeasy MinElute columns (Qiagen). Total RNA and gDNA were quantified using a Nanodrop Spectrophotometer (Thermo Scientific) and either Quant-iT RNA or DNA reagents (Invitrogen). Double stranded (ds) cDNA was synthesised from 250 ng of cleaned RNA. The first strand was synthesised with Superscript III (Invitrogen) and random hexamers. The second strand was synthesised with DNA polymerase I (Invitrogen), ribonuclease H (Invitrogen) and dNTPs. The 96-well plate containing the ds cDNA samples was cleaned using Multiscreen PCR plate (Millipore). This work was carried out by Matthew Forrest at the WTSI.

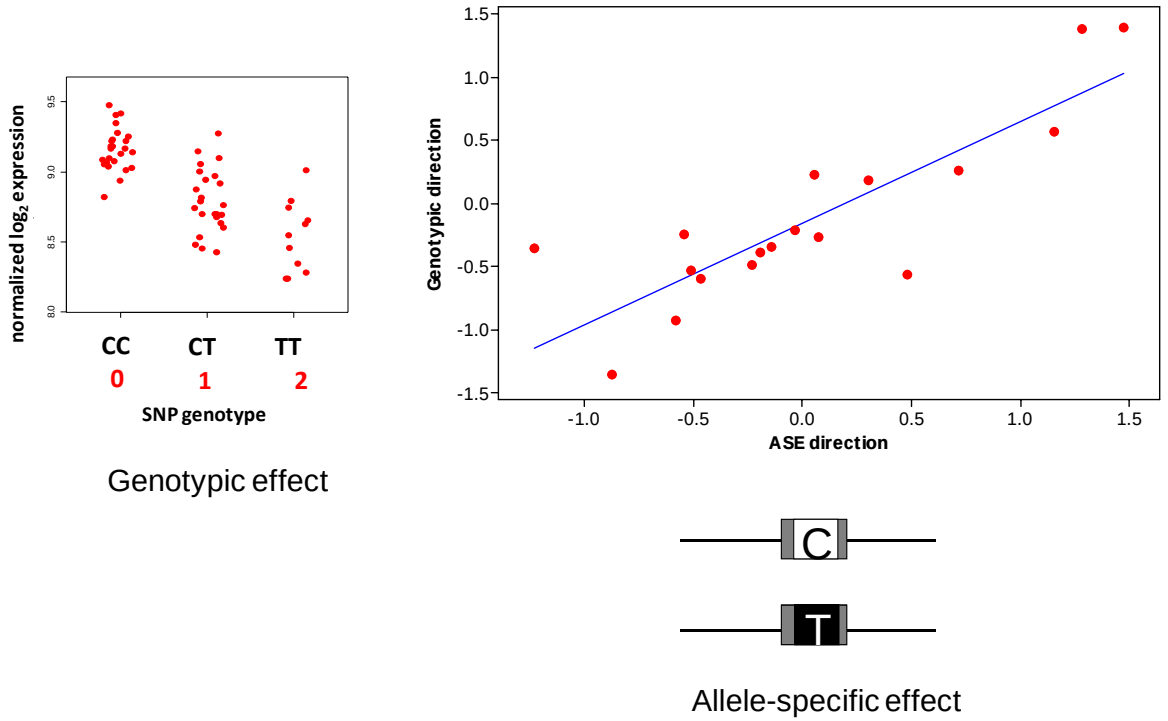


Figure 11. Statistically predicted vs. biologically verified differentially expressed (DE) nsSNPs. Allele-specific expression (ASE) assays were used to verify differential expression of nsSNPs (cis eQTLs) predicted from statistical analyses. In this scatterplot the experimentally determined allelic effect (x axis) of a C/T nsSNP is compared to the genotypic effect predicted from the association test (y axis).

2.6.5.2 illumina allele-specific expression array

A custom made Oligo Pool All (OPA) array (illumina) based on the Golden Gate assay was used to assay ASE. Only exonic SNPs ≥ 45bp from both exon edges were chosen for submission to illumina for assay design, to ensure that the assay would work equally well for genomic and cDNA. SNPs that failed according to illumina's design scores were discarded. Paired ds cDNA and gDNA were dried down in 96-well plates and re-suspended in 5µl of HPLC purified water. Golden Gate assays were then run for all samples using the manufacturer's standard protocol for gDNA (i.e. ds cDNA was treated exactly the same way as gDNA). Reactions were hybridised to 8×12 Sentrix

Array Matrix (SAM) Universal Probe Sets so that 96 arrays could be run in parallel. Each bead type (probe) is present on a single array on average 30 times. All reactions were run in duplicate, so that each LCL had two ds cDNA replicate and two gDNA replicate hybridizations. SAMs were scanned with a Bead Station (illumina). A total of 1,536 assays were interrogated on the array, but only 141 were nsSNPs from this study and only 28 were selected based on data quality for further analysis. This work was carried out by Matthew Forrest at the WTSI.

2.6.5.3 Allele-specific expression assay data pre-processing

Data from each array were summarised by calculating the per bead type average of 4 quantities after outlier removal: the $\log_2$(Cy3) and $\log_2$(Cy5) intensities, average log-intensities ($^1/_2\log_2$(Cy5.Cy3)) and log-ratios ($\log_2$(Cy5/Cy3)). Outliers were beads with values more than three absolute deviations from the median. Arrays with low dynamic range (determined using an inter-quartile range cut-off of < 1 for either the $\log_2$(Cy3) or $\log_2$(Cy5) summary intensities) were discarded. The summarised data were normalized by median centring of log-ratios. Normalisation was carried out in R using the *Beadarray* package (Dunning, Smith et al. 2007) by Matthew Ritchie at the CRI. Direction of expression (high/low) was assigned to alleles for nsSNPs fulfilling the threshold criteria from the association study (adjusted $r^2 \geq 0.27$; i.e. the nsSNP explained at least 27% of the variance in gene expression so the effect is expected to be large) and the ASE assay (average cDNA log-intensity ≥ 12 within a population).

2.6.6 Amino acid substitution effect

Given that nsSNPs are likely to be functional I explored three aspects of the resulting amino acid substitution: a) relative position of substitution on the peptide, as a percent of peptide total length. b) hydrophobicity change in peptide resulting from the amino acid substitution. For each pair of variant sequences the hydrophobicity at the position

of the variant amino acid was calculated using the Kyte-Doolittle algorithm (Kyte and Doolittle 1982) and a window size of seven amino acids (centred on the variant amino acid). The difference between hydrophobicity scores was then taken for each of the variant pairs in the dataset. c) Pfam score change in peptide sequence resulting from the amino acid substitution (Finn, Tate et al. 2008). All sequences were searched against the profile-HMM library provided by the Pfam database (release 22.0) using hmmpfam from the HMMer software package (version 2.3.2, http://hmmer.janelia.org/) and a default cut off E-value of 10. Only the HMM_ls library was used so that domain assignments to a pair of variant sequences were comparable. The set of Pfam domain assignments were then filtered such that only the domains that overlapped with the SNP position and that at least one of the domain assignments from a pair of variant sequences scored above the Pfam defined gathering threshold, were considered in the subsequent analysis. The difference between the two E-values was taken for each of the variant pairs in the dataset. Pfam scores were provided by Robert Finn from the Pfam team at the WTSI.

2.6.7 Impact of eQTL-nsSNP interaction in trans

The impact of interactions between eQTLs and nsSNPs on gene expression in trans was tested for the CEU population. In a previous study (Stranger, Forrest et al. 2005), trans effects were found to be weak in the YRI population and the number of individuals in the CHB and JPT Phase 2 populations limit the power for detection of trans effects. To test the trans effect of eQTL-nsSNP interactions I pooled the minor allele homozygote and the heterozygote into a single genotypic category and then coded genotypes as 0 (major allele homozygote) or 1 (heterozygote and minor allele homozygote) for both eQTL and nsSNP. As a result, four possible eQTL-nsSNP genotypic combinations are possible: 0-0, 1-0, 0-1, 1-1 (Figure 12). Analysis of variance (ANOVA) was performed

using the *R* software package (R Development Core Team 2008) to test the effects of: the eQTL, the nsSNP, and the eQTL x nsSNP interaction term against gene expression phenotypes in trans. In each case the gene from which the eQTL-nsSNP pair originated was excluded from the association test. To ensure all genotypic combinations were present and to avoid outlier effects, tests were carried out for 22 SNP pairs with low LD ($D' \leq 0.5$) between eQTL and nsSNP and a MAF ≥ 0.1 for both variants.

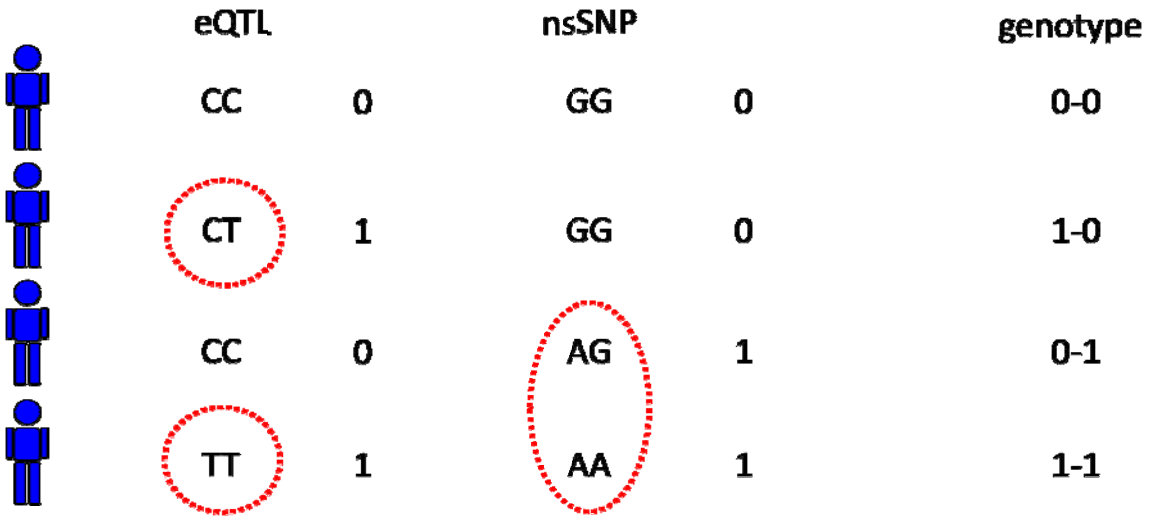


Figure 12. Genotypes for eQTL and nsSNP were combined and used to test for impact of the eQTL-nsSNP interaction on gene expression in trans. The minor allele homozygote and the heterozygote were collapsed into a single genotypic category (red circles represent genotypes that were pooled for each variant). This resulted in two genotypic categories coded as 0 (major allele homozygote) or 1 (heterozygote and minor allele homozygote) for both eQTL and nsSNP. As a result, when combining eQTL-nsSNP genotypes four combinations (shown in the right column) are possible: 0-0, 1-0, 0-1, 1-1.

To assess significance of interaction p-values a single permuted dataset of expression values relative to combined genotypes was generated and the p-value distributions of interaction terms for observed and permuted data were compared. To further evaluate the robustness of observed interactions, I permuted eQTL genotypes

relative to nsSNP genotypes and gene expression phenotypes, and re-ran the ANOVA association test for the top ten most significant interactions.

2.7 eQTL FINE-SCALE ARCHITECTURE STUDY (CHAPTER 4)

2.7.1 Recombination hotspot interval mapping and LD filtering

LD is a useful property of the genome as it enables genome-wide mapping of variation associated with a phenotype. At a smaller scale however LD impedes fine-mapping as multiple correlated variants can show a significant association with a trait. The aim of this study was to identify those cis eQTLs that tag the effects of independent regulatory elements and in this way detect independent cis regulatory signals for a gene. To do this I mapped eQTLs in recombination hotspots and recombination hotspot intervals using data on the recombination patterns in the genome (McVean, Myers et al. 2004; Myers, Bottolo et al. 2005; Winckler, Myers et al. 2005).

A recombination hotspot interval was defined as the space between two recombination hotspots and represents a segment of DNA with an independent recombination history (McVean, Myers et al. 2004). Recombination hotspot intervals were constructed using hotspot coordinates estimated from HapMap Phase 2 data and coordinates were lifted over to Build 36 using the UCSC liftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). eQTLs were mapped in intervals and only the most significant eQTL per interval was considered for further analysis. Correlation between this subset of eQTLs is still possible if LD extends across intervals and to ensure that independent signals were identified, the least significant variant from eQTL pairs with $D' \geq 0.5$ for a given gene was removed. Independent eQTLs (or regulatory intervals) therefore define genomic units likely to carry independent functional regulatory elements. The strategy described is shown in Figure 13.

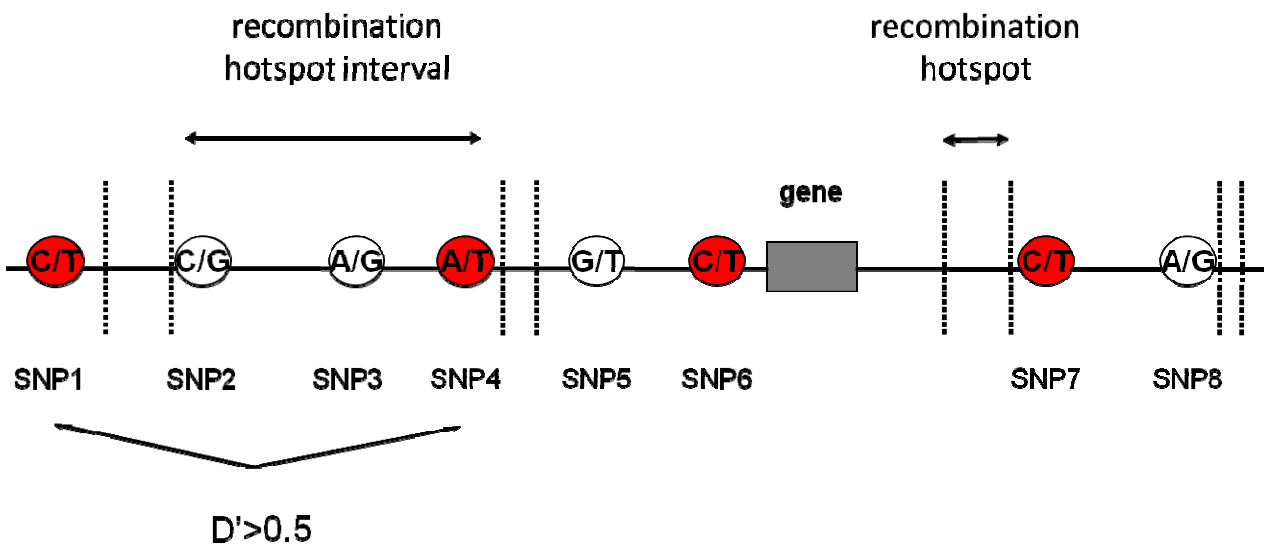


Figure 13. Detecting independent eQTLs (intervals). Recombination hotspot intervals were defined as the space between two consecutive recombination hotspots (McVean, Myers et al. 2004) and represent an approximation of genomic units with an independent history. To identify intervals with an independent effect on gene expression in cis, eQTLs were mapped in recombination hotspot intervals and the most significant eQTL for a given interval (shown in red) was retained. Further control for correlation was performed by excluding the least significant eQTL from eQTL pairs with a D' > 0.5 (e.g. SNP1-SNP4, with SNP4 being the most significant of the two). In this example SNP4, SNP6 and SNP7 are independent cis eQTLs defining independent intervals. A modified version of this strategy was used to test the effects of interactions between SNPs that have an impact on gene expression in cis: SNPs with a nominal (uncorrected) p-value < 0.001 were mapped in intervals and SNP pairs with a D' > 0.5 were excluded from association testing. In this example an interaction effect would be tested for the following pairs: SNP4-SNP6, SNP6-SNP7, and SNP4-SNP7 (see section 2.7.3).

D' was chosen over r^2 to filter for LD as it is a metric of the degree of historical recombination that has occurred between two variants (Hartl and Clark 2007). r^2 on the other hand is an indicator of statistical correlation and not of historical relationships. As a result, if two SNPs have different MAFs, but there has been no historical recombination between them in the samples studied, r^2 can take low values, but D' = 1. Under such a scenario, if one SNP displays a strong association with expression levels of a gene, it can be that the other SNP is also associated with expression levels of the same gene, even if r^2 is low. This is possible as the two SNPs may be tagging the same

functional variant since there has been no historical recombination between them. Using a D' threshold (which translates to an even lower r^2) ensures that the two signals are historically independent. Upon recombination hotspot interval mapping and LD filtering, I determined the number of independent intervals detected for each gene at the 0.01 and 0.001 permutation thresholds. This analysis was carried out for HapMap Phase 3 (Chapter 4) and GenCord data (Chapter 5). For GenCord an overlap analysis was carried out to determine the extent to which independent eQTLs (intervals) are shared across the three cell types studied.

2.7.2 Independent eQTL distance to transcription start site

To describe the cis regulatory landscape around genes, p-values and effect sizes (rho) of the most significant eQTL per gene were plotted relative to the TSS. This was done for both HapMap Phase 3 and GenCord data. HapMap Phase 3 data were analysed by Barbara Stranger at the WTSI.

2.7.3 eQTL-eQTL cis interaction

To further characterise cis regulatory architecture, HapMap Phase 3 SNP pairs were tested for an interaction with an impact on gene expression using the CEU and YRI populations. The interaction model employed in this analysis was identical to that described in section 2.6.7, but instead of testing interactions between regulatory (eQTLs) and protein-coding (nsSNPs) variation, I explored interactions between SNPs likely to tag regulatory variants. Variants tested were not filtered for permutation threshold significance (and are not termed eQTLs), but were SNPs with an observed nominal p-value < 0.001 from the SRC cis association test. These variants were chosen so that SNPs that do not necessarily have large marginal effects are included in the interaction test. (Ideally all SNPs for a given gene should be tested for an interaction to uncover variants

that influence gene expression through interactions. At the time of writing, this was being explored in collaboration with Doug Speed and Simon Tavaré at the CRI). SNPs were mapped in recombination hotspot intervals and the most significant SNP per interval was kept. SNP pairs were constructed for each gene in cis, and pairs with D' ≥ 0.5 were excluded. ANOVA was used to test the main effects of each SNP, as well as the SNP x SNP interaction term, on gene expression in cis. A single permutation of expression values relative to genotypes was performed to assess significance of the interaction p-values. The p-value distributions of the interaction term for observed and permuted data were compared.

2.8 eQTL CELL TYPE SPECIFICITY STUDY (CHAPTER 5)

2.8.1 Association analysis

GenCord data were used to investigate the cell type specificity of cis eQTLs. A total of 22,651 probes covering 17,945 autosomal RefSeq genes (15,596 Ensembl genes) were tested for cis association with SNP genotypes using SRC. Cis association tests encompassed SNPs mapping in a 2 Mb window centred on the TSS. Following quality control and filtering for MAF ≥ 5%, a total of 394,651 SNPs were included in the analysis. Significance thresholds for each gene were assigned after 10,000 permutations of expression values relative to genotypes. To explore sharing and cell type specificity of significant associations, I compared eQTLs and genes across cell types and determined those that passed significance thresholds in all three, in at least two and in only one cell type (overlap analysis).

2.8.2 Repeated-measures ANOVA to investigate eQTL cell type specificity

I used repeated-measures ANOVA (RMA), programmed in *R* (R Development Core Team 2008), to investigate the robustness of sharing and cell type specificity of associations. I tested shared and cell type-specific SNP-probe pairs identified from the overlap analysis (at the 0.001 permutation threshold), in tests where the repeated measure was the cell type. The analysis was run for pairs of cell types and the significance of the SNP x cell type interaction term was assessed. The expectation is that the interaction term will be significant for those eQTLs that were identified as cell type-specific.

2.8.3 Allele-specific expression assay

ASE assays were used to validate a subset of cell type-specific eQTLs. Thirty five transcript SNPs (seven in fibroblasts, 14 in LCLs, and 14 in T-cells) in genes with identified cell type-specific eQTLs were tested for ASE in each cell type. The expectation is that allelic imbalance will be observed for the cell type in which the eQTL was detected. 800 ng of total RNA, in a total volume of 20 µl from fibroblasts, LCLs, and T-cells was converted to cDNA using hexaprimers (Superscript II, Invitrogen). This work was carried out by Christelle Borel at the UGMS. gDNA (~40 ng) from LCLs and cDNA (~30 ng) from each cell type, as well as an RNA control (~30 ng) from 293 T-cells were genotyped using Sequenom's MassArray allele specific assay without competitor (iPLEX Gold assay, Sequenom, San Diego, California, USA). Assays for all SNPs were designed using the eXTEND suite and MassARRAY Assay Design software version 3.1 (Sequenom). Amplification was performed in a total volume of 5µL containing the DNA, 100 nM of each PCR primer, 500 nM of each dNTP, 1.25 x PCR buffer (Qiagen), 1.625 mM MgCl2 and 0.2 U HotStar Taq (Qiagen). Reactions were heated to 95°C for 15 minutes followed by 45 cycles at 94°C for 20 s, 56°C for 30 s, 72°C for 60 s and a final

extension at 72°C for 3 minutes. Unincorporated dNTPs were SAP digested prior to iPLEX Gold allele specific extension with mass modified ddNTPs using an iPLEX Gold reagent kit (Sequenom). SAP digestion and extension were performed according to the manufacturer's instructions with reaction extension primer concentrations adjusted to between 0.731-2.193 μM, dependent upon primer mass. Extension products were desalted and dispensed onto a SpectroCHIP using a MassARRAY Nanodispenser (Sequenom) prior to analysis with a MassARRAY Analyzer Compact mass spectrometer (Sequenom). Allele-specific peak heights from the mass spectra of gDNA and cDNA were analysed to detect transcript SNPs showing allelic imbalance. This work was carried out by Naomi Hammond at the WTSI. The ratio of the two alleles of transcript SNPs was analysed in RNA samples of individuals who were double heterozygotes for both the eQTL and the transcript SNP.

2.8.4 Biological properties of cell type-specific associations

Gene Ontology (GO) terms (Ashburner, Ball et al. 2000) were used to investigate the biological properties of cell type-specific gene associations (at the 0.001 permutation threshold). GO terms were assigned to Ensembl Genes (Ensembl 50) and were then mapped on to their GO Slim ontologies. GO Slim represents a cut-down version of GO and gives a broader overview of the ontology (Ashburner, Ball et al. 2000). Fisher's exact tests were used to compare GO Slim terms corresponding to gene associations that were cell type-specific vs. associations that were shared in all three cell types. Significant associations were those for which Fisher's exact p-value ≤ 0.05.

2.8.5 Tissue entropy

Gene expression entropy was used as a proxy to gene expression specificity (Jongeneel, Delorenzi et al. 2005; Schug, Schuller et al. 2005; Martinez and Reyes-Valdes 2008). The

expression of genes possessing an eQTL in a single cell type (0.001 permutation threshold) was investigated using the GNF/Novartis expression atlas, which contains expression data for 10,424 Ensembl genes from 38 tissues (Su, Wiltshire et al. 2004). The GNF/Novartis data were used to calculate gene expression entropy as described in Schug et al (2005). Briefly, for expression levels measured in N tissues, the relative expression of a gene *g* in a tissue *t* is defined as:

$$p_{t|g} = W_{g,t} / \Sigma_{1 \le t \le N} w_{g,t}$$

where $w_{g,t}$ is the expression level of the gene in that tissue. The entropy of a gene's expression across N tissues is defined as follows:

$$H_g = \Sigma_{1 \le t \le N} - p_{t|g} log_2(p_{t|g})$$

To assess cell type specificity of associations, I compared the entropy distributions for genes with associations that were cell type-specific vs. associations that were: a) three cell type-shared, b) at least two cell type-shared, c) two cell type union. Significant differences in entropy distributions were those for which M-W p-value ≤ 0.05.

3 MODIFIER EFFECTS BETWEEN REGULATORY AND PROTEIN-CODING VARIANTS

In this chapter I will:

- Outline how interactions between genetic variants have an impact on phenotypes.
- Explain why detecting interactions is challenging.
- Put forth a biological framework that can be used to test for interactions between regulatory (eQTLs) and protein-coding variants (nsSNPs) with an impact on gene expression.
- Demonstrate a modification effect in cis, arising from the eQTL-nsSNP interaction, that also has a trans effect on gene expression.
- Discuss the biological implications of this interaction.

3.1 CONTEXT-DEPENDENT EFFECTS ON PHENOTYPES: INTERACTIONS

To date, most association studies attempt to link single genetic variants to a specific phenotype (Brem, Yvert et al. 2002; Morley, Molony et al. 2004; Stranger, Forrest et al. 2005; Goring, Curran et al. 2007). Most of the systems that underlie cellular, developmental and physiological function however are composed of many elements that interact with one another, often in complex ways (Phillips 2008). As a result the extent to which a phenotype is shaped by genetic factors may not be a simple reflection of their independent effects, but is likely to arise in part from context-dependent effects, such as interactions between genetic factors, as well as interactions between genetic factors and the environment (Gibson 2008; Phillips 2008; Flint and Mackay 2009). The interaction between genetic variants that results in a phenotypic effect conditional on the combined presence of two or more variants is called epistasis (Brem, Storey et al.

2005; Nagel 2005). Epistasis may arise from a variety of underlying mechanisms. Over the years geneticists have used this term to describe subtly different genetic phenomena including the functional relationship between genes, the genetic ordering of regulatory pathways and the quantitative differences of allele-specific effects (Phillips 2008; Cordell 2009). Philips (2008) defines three forms of epistasis: functional epistasis (the molecular interactions that proteins and other genetic elements have with one another), compositional epistasis (the blocking of one allelic effect by an allele at another locus), and statistical epistasis (the average effect of substitution of alleles at combinations of loci, with respect to the average genetic background of the population). Hartl and Clark (2007) define epistasis as any situation in which the genetic effects of different loci that contribute to a phenotypic trait are not additive. In this thesis I refer to epistasis as a property of specific alleles at two loci whose interaction has an impact on gene expression, and will use the term interchangeably with the term interaction.

3.2 Prevalence and biological significance of interactions

The prevalence and biological significance of epistasis has always been an area of interest in the field of genetics, but its contribution to phenotypic variation has remained obscure, largely because genetic interactions have proven difficult to test (Musani, Shriner et al. 2007; Cordell 2009). This difficulty arises primarily because it is unclear which variant combinations should be tested and under which model of epistasis. To date, such an approach has been most feasible for specific genes or biological pathways that have been well-characterised, mostly in model organisms.

One of the best studied examples of epistasis is coat colour in mammals. In mice, an adaptive transition from dark to light coat colour accompanied the movement of dark-coloured forest mice from the forest to the beach (Steiner, Weber et al. 2007; Phillips 2008). The genetic basis for this transition stems from an interaction between

structural changes to the *agouti* locus and regulatory changes to the *Mc1r* locus. Obesity is another phenotype in mice that is affected by epistatic interactions and an extended network of epistatic QTLs has been discovered on chromosomes 4, 17, and 19 that controls regulation of fat pad depots and body weight (Stylianou, Korstanje et al. 2006).

A classic example of an interaction between regulatory and protein-coding variation is the *Adh* locus in *Drosophila* (Laurie, Bridgham et al. 1991; Stam and Laurie 1996). A series of regulatory SNPs in complex LD and with an impact on protein concentration, modify the effects of a protein-coding variant affecting the catalytic efficiency of this enzyme. Catalytic efficiency and protein levels determine overall enzyme activity. This example illustrates that large effects attributed to a single locus may arise as a consequence of multiple associated interacting variants and is a case of a modification effect in cis where the protein-coding effect is magnified or masked through the action of regulatory variants. More recent studies in *Drosophila* reveal epistatic effects between genes affecting traits such as ovariole number (Orgogozo, Broman et al. 2006) and olfactory avoidance (Sambandan, Yamamoto et al. 2006).

In cases where little is known about the genes sculpting a phenotype, addressing the possibility of epistasis becomes more challenging. A recent study interrogating cardiac dysfunction in *Drosophila* (Ocorr, Crawley et al. 2007) identified a major susceptibility locus for this trait, but highlighted the importance of examining the phenotype in different genetic backgrounds to detect variants whose effects are manifest through interactions with the prime susceptibility locus. The extent of epistasis in a more global way has been demonstrated in yeast where experiments on gene expression revealed that interacting locus pairs are involved in the inheritance of over half of all transcripts (Brem, Storey et al. 2005; Boone, Bussey et al. 2007). Furthermore, a large proportion of the eQTLs attributable to interaction effects were not detected by single locus tests. This suggests that analysis of interaction effects in other systems is likely to uncover additional associations.

In humans, most documented cases of epistasis have been detected in instances where there are biological clues as to which genes should be tested. Epistasis between two multiple sclerosis (MS) associated human leukocyte antigen (HLA) alleles was demonstrated by Gregerson et al. (2006) who showed that one allele modifies the T-cell response that is activated by a second allele, through activation-induced apoptosis contributing to a milder form of MS-like disease. Similarly, Oprea et al. (2008) demonstrated that a specific modifier effect is protective against spinal muscular atrophy (SMA). SMA arises from a homozygous deletion of the *SMN1* gene, but some deletion homozygotes escape the disease phenotype due to the modulating effects of expression of PLS2.

Risk for nicotine dependence and lung cancer was shown to be sculpted by interactions between functional variants in genes belonging to the neuronal nicotinic acetyl choline receptor (nAChR) family (Wang, Cruchaga et al. 2009). nAChR genes encode pentameric ligand-gated ion channels that mediate fast signal transmission at synapses and modulate the release of neurotransmitters. Nicotine is an exogenous agonist of these receptors, and variations in nAChR genes are strong candidate risk factors for nicotine dependence and lung cancer. The authors of this study showed that interactions between a coding variant, that changes amino acid sequence in the α5 nicotine receptor subunit gene *CHRNA5* (D398N), and non-coding variants that regulate the gene's expression levels confer risk for nicotine dependence and lung cancer. They conclude by stating that by establishing this cis modification effect they have identified a potential drug target.

With the explosion of successful GWAS over the past three years, the natural next step is genome-wide interaction testing (Cordell 2009). Detecting epistasis is crucial as it is likely to uncover new variants affecting phenotypes. Additionally, epistasis may mask the genetic impact of variants and impede replication of primary associations. Differential fixation of variants that modulate the primary disease variant can therefore

affect the degree of penetrance of disease alleles and the need to address this property of genes in a systematic, genome-wide approach is becoming increasingly pressing. The case of MS clearly illustrates this: as with most complex disorders, MS has a polygenic heritable component characterised by underlying complex genetic architecture (Oksenberg, Baranzini et al. 2008). Association studies to date have met with modest success in identifying MS-causing genes, and a large proportion of phenotypic variation remains unexplained. The expectation is that this residual variation arises at least in part, as a consequence of gene-gene interactions.

In this study I explored the extent to which regulatory variants modify protein-coding effects in cis and tested whether this modification effect has an impact on gene expression of other genes in the genome in a trans effect. This work has been described in (Dimas, Stranger et al. 2008).

3.3 BIOLOGICAL FRAMEWORK TO DETECT INTERACTIONS

Most strategies that address the effects of epistasis in humans involve millions of agnostic pairwise tests falling into one of two broad categories: exhaustive testing of interactions between all pairs of variants across the genome (Marchini, Donnelly et al. 2005), or testing of interactions between all pairs of variants with an independent main effect on the phenotype (Marchini, Donnelly et al. 2005; Evans, Marchini et al. 2006; Dixon, Liang et al. 2007). It is not entirely clear whether improvements in statistical methods will be sufficient to address the problem of epistasis. Therefore the development of realistic biological models of epistatic interactions may reduce the statistical cost of dealing with many comparisons and facilitate the development of such methodologies.

In this study I present a biological framework for global survey of interaction effects in humans, which avoids exhaustive testing of agnostic pairs and involves

prioritisation of variants to be tested. Two types of functional variants are common throughout the human genome and are present at appreciable frequencies in populations: regulatory variants with an impact on the expression patterns and levels of genes (Pastinen and Hudson 2004; Birney, Stamatoyannopoulos et al. 2007; Forton, Udalova et al. 2007; Spielman, Bastone et al. 2007; Stranger, Nica et al. 2007) and protein-coding variants affecting protein sequence (Rodriguez-Trelles, Tarrio et al. 2003; Birney, Stamatoyannopoulos et al. 2007). To date, the effects of these variants have been considered independently of each other. In this study I evaluated the joint effects of regulatory and protein-coding variants on genome-wide expression phenotypes in humans to highlight an underappreciated angle of functional variation.

As outlined in section 2.6.1 the proposed model brings together quantitative and qualitative variation, by testing the cis and trans impact on gene expression observed when a gene with an identified regulatory variant (eQTL) also contains protein-coding variation (nsSNP). Under such a scenario, and assuming that mRNA levels are indicative of mature protein levels, the resulting protein products will differ in quantity (expression level) and quality (amino acid sequence) among individuals (Figure 9). Depending on the historical rate of recombination between eQTLs and nsSNPs, different allelic combinations (haplotypes) can arise on the two homologous chromosomes in a population (Figure 14). As a consequence, phasing (the arrangement of alleles at each variant position with respect to one another) can differ between individuals in the population. Such an interaction results in a modification (magnification or masking) of the functional impact of the protein-coding variant. If the modified gene product has downstream targets, then expression of these target genes may also be affected in a trans manner.

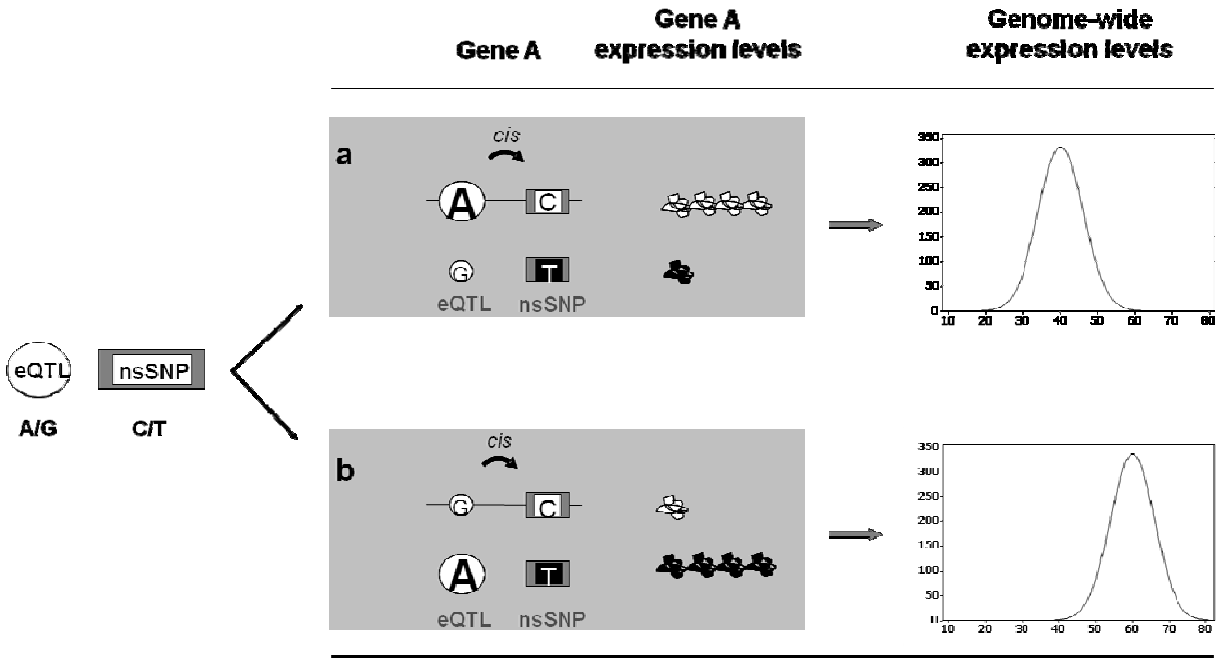


Figure 14. Illustration of a hypothetical epistatic interaction between a regulatory (eQTL) and a protein-coding variant (nsSNP). Two double heterozygote individuals may be genotypically identical, but the phasing of alleles can be different and may result in very distinct phenotypes between individuals. In **a)** the A allele of the eQTL drives high expression levels of the protein arising from the C allele of the nsSNP. In **b)** the G allele of the eQTL drives low expression levels of the protein arising from the C allele of the nsSNP. If the protein-coding variant is functionally important then this interaction in cis can give rise to different means in the distribution of a complex trait phenotype (e.g. genome-wide expression levels) as shown on the right (trans effect).

3.4 MODIFICATION EFFECT IN CIS: DIFFERENTIALLY EXPRESSED NSSNPS

Using this model as a main principle, I explored the degree to which nsSNPs can be modulated by cis eQTLs. eQTLs were identified in a previous study (Stranger, Nica et al. 2007) in LCLs of the unrelated individuals of the Phase 2 HapMap populations (60 CEU, 45 CHB, 45 JPT and 60 YRI) (Table 4). LCLs represent one particular cell type and even though there may be some effect arising from EBV transformation, it has been demonstrated that genetic effects on gene expression, such as the ones I describe below,

are readily identifiable, mappable, and replicate in independent population samples generated decades apart (Dimas, Deutsch et al. 2009).

	0.01 permutation threshold			
Population	LR: significant genes	CEU-CHB-JPT-YRI multi-population	CEU-CHB-JPT multi-population	CHB-JPT multi-population
CEU	606	1,186	1,149	1,071
CHB	**634**	**1,186**	**1,149**	**1,071**
JPT	679	1,186	1,149	1,071
YRI	**742**	**1,186**	**1,149**	**1,071**
Non-redundant union	**1,746**			
4 populations	114			
≥2 populations	**533**			

Table 4. eQTLs detected in the HapMap Phase 2 populations (0.01 permutation threshold). Adapted from (Stranger, Nica et al. 2007).

Two strategies were applied to detect DE nsSNPs. The first strategy involved scanning genes with known cis eQTLs (Stranger, Nica et al. 2007), for nsSNPs. The aim was to identify nsSNPs that are predicted to be DE as a consequence of a nearby regulatory variant tagged by the eQTL. I identified 606, 634, 679 and 742 genes with at least one eQTL at the 0.01 permutation threshold (estimated FDR of 20%) (Table 4). Of these genes 159, 168, 180 and 202 (union of 484) were found to contain 286, 304, 311 and 393 nsSNPs respectively (union of 909) (Table 5). I infer that these nsSNPs are DE as they reside in genes with experimentally-derived varying expression levels. This means that there are allelic effects on gene expression such that, depending on the genotypes of the eQTL and nsSNP and on the phasing of their alleles, one can make predictions about the relative abundance of the two alleles of a transcript in the cell.

a

	nsSNPs					
Population	Total nsSNPs interrogated	With identified eQTL	Single population association	Single population DE	Multi-population association	Total DE
CEU	5,686	286	242	452		
CHB	5,335	304	276	478		
JPT	5,328	311	267	487		
YRI	6,093	393	255	574		
Non-redundant union	8,233	909	703	1,355	587	1,502

b

	genes					
Population	Total genes interrogated	With identified eQTL	Single population association	Single population DE	Multi-population association	Total DE
CEU	3,579	159	196	307		
CHB	3,412	168	226	322		
JPT	3,410	180	210	325		
YRI	3,692	202	211	364		
Non-redundant union	4,518	484	560	863	461	973

Table 5. a) nsSNPs and b) genes interrogated for differential expression. (DE: differentially expressed)

The second strategy for DE nsSNP discovery involved direct association testing (using LR) between nsSNP genotype and expression levels of the gene in which the nsSNP resides. This strategy aimed to identify DE nsSNPs that are in LD with a regulatory variant that drives expression levels. Depending on the strength of the regulatory effect, such variants may or may not have been detected in the initial scan for eQTLs (Stranger, Nica et al. 2007). Relative distances between eQTLs and nsSNPs can vary, but in the special case where this distance is short in genetic terms, the two variants may be in LD (McVean, Spencer et al. 2005). Under these circumstances it is expected that the nsSNP itself will demonstrate some degree of association with expression levels of the gene in which it resides. I tested for genotype-expression associations in each population separately and in three multiple population sample panels (see section 2.6.3).

For the single-populations analysis, with significance evaluated at the 0.01 permutation threshold, 56 nsSNPs and 34 genes are expected to have at least one significant association by chance. I detected 242, 276, 267 and 255 nsSNPs (union of 703; estimated FDR of 21%) with significant for the CEU, CHB, JPT and YRI populations respectively (Table 6 a). These associated nsSNPs correspond to 196, 226, 210 and 211 genes (union of 560; estimated FDR of 16%) (Table 6 b). For the multiple-population analysis I detected 345, 362 and 417 nsSNPs (estimated FDR of 15%) for the four, three and two population groups respectively (Table 6 a), corresponding to 284, 296 and 320 significant genes (estimated FDR of 11%) (Table 6 b). Overall, the multiple-population analysis yielded a total of 587 nsSNPs with significant associations, corresponding to 461 genes. Taken together, the association analyses indicate that 884 nsSNPs (688 genes) across the four populations are associated with expression levels of the genes they are in, suggesting that they are in LD with regulatory variants driving their expression. In this specific case of association, the nsSNP itself serves as a proxy for the regulatory variant and knowledge of associated nsSNP genotype for an individual provides a prediction of relative abundance of the two transcript alleles.

a

	0.01 permutation threshold						
	1	2	3	4			
Population	significant nsSNPs	CEU-CHB-JPT-YRI multipop	CEU-CHB-JPT multipop	CHB-JPT multipop	Overlap 1&2	Overlap 1&3	Overlap 1&4
CEU	242	345	362	417	111	139	104
CHB	276	345	362	417	126	162	224
JPT	267	345	362	417	136	161	203
YRI	255	345	362	417	102	86	90
Nonredundant	703						
4 populations	34						
≥2 populations	233						

b

	0.01 permutation threshold						
	1	2	3	4			
Population	significant genes	CEU-CHB-JPT-YRI multipop	CEU-CHB-JPT multipop	CHB-JPT multipop	Overlap 1&2	Overlap 1&3	Overlap 1&4
CEU	196	284	296	320	99	117	87
CHB	226	284	296	320	109	129	183
JPT	210	284	296	320	114	125	156
YRI	211	284	296	320	87	77	82
Nonredundant	560						
4 populations	31						
≥2 populations	196						

Table 6. a) nsSNP and b) gene cis associations detected in single and multiple populations.

To summarize, two classes of DE nsSNPs were discovered: a) 909 nsSNPs mapping in genes with a previously identified eQTL (considering nsSNPs of all frequencies) and b) 884 nsSNPs showing a significant association with expression levels of the gene they are in (considering nsSNPs with MAF ≥ 0.05) (Figure 15). From a non-redundant total of 8,233 nsSNPs tested in four populations, 1,502 of these (~18.2%) are predicted to be DE. It is a plausible biological hypothesis that mature protein levels mirror transcript levels on average and as a consequence, this high fraction of DE nsSNPs may have important implications for levels of protein diversity in the cell.

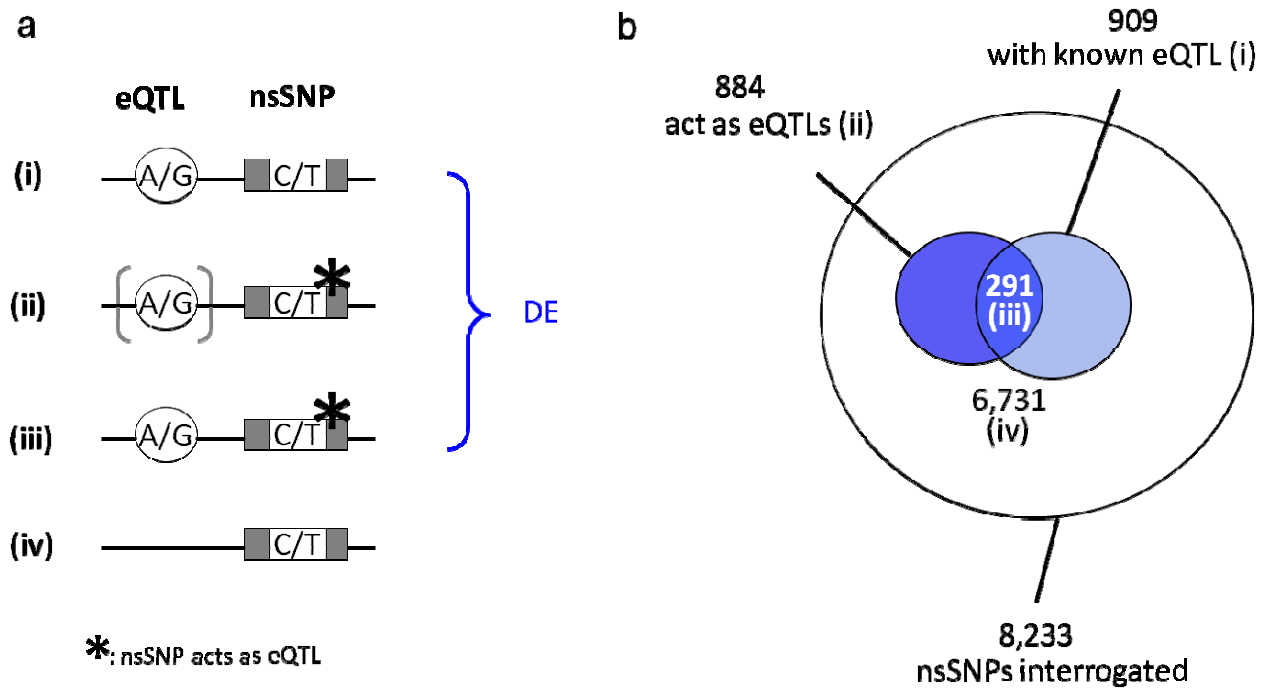


Figure 15. Strategies applied to discover differentially expressed (DE) nsSNPs. a) Two approaches were employed to discover DE nsSNPs: nsSNPs mapping in genes with a known eQTL (i) and nsSNPs that were associated with expression levels of the gene they map in (ii). In (ii) the presence of a cis-acting regulatory variant is implied. For some nsSNPs with a significant association, an identified cis eQTLs also exists (iii). In all other cases the nsSNPs interrogated were not inferred to be to be DE (iv). **b)** Of the 8,233 nsSNPs studied, 909 mapped in a gene with an identified eQTL (i), 884 were found to be associated with levels of expression of the gene they reside in (ii), 291 nsSNPs with an identified eQTL also showed a significant association with expression levels (iii) and 6,731 nsSNPs showed no evidence of differential expression (iv). Taken together over 18% of nsSNPs were found to be DE.

3.4.1 Linkage disequilibrium between eQTLs and nsSNPs

Of the 884 DE nsSNPs detected through association testing, only 291 also possess a previously identified eQTL. This suggests that eQTL detection in our previous study was conservative and that nsSNPs can act as tags of undiscovered regulatory variants. With this in mind, it is expected that LD between eQTL-nsSNP pairs in which the nsSNP had a significant association with gene expression, will be greater than LD between eQTL-nsSNP pairs in which the nsSNP was not associated. To explore this, I

used data from the single population analysis, and compared the distribution of r^2 values between the two eQTL-nsSNP pair types. As expected, much higher LD was found for eQTL-nsSNP pairs where the nsSNP showed a significant association (M-W p-value < 0.0001) (Figure 16). This confirms that in most cases, association of the nsSNP with its gene's expression is due to a regulatory variant tagged by the eQTL.

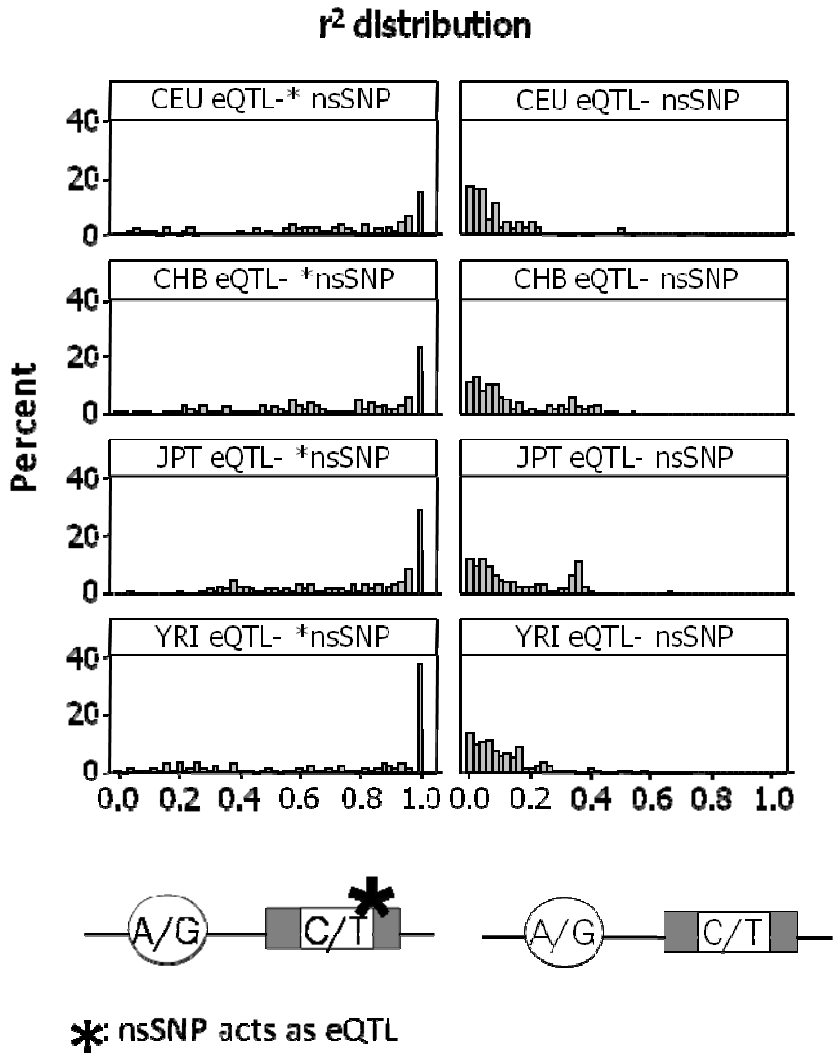


Figure 16. Linkage disequilibrium (LD) properties of eQTL-nsSNP pairs. The distribution of r^2 (a measure of LD) was compared between eQTL-nsSNP pairs in which the nsSNP acts as an eQTL (i.e. showed a significant association with its gene's expression levels) and SNP pairs in which the nsSNP was not associated. As expected, r^2 values are much higher in the first case, where the nsSNP is thought to act as a tag of the functional regulatory variant nearby.

3.4.2 Experimental verification of differentially expressed nsSNPs

Thus far I have described relative abundance estimates for transcripts of genes containing nsSNPs using genotypic associations. To verify the statistical predictions of nsSNP association tests, it was necessary to perform direct allele-specific quantification. A subset of nsSNPs were tested for ASE (Pastinen, Ge et al. 2006; Forton, Udalova et al. 2007) in heterozygote CEU and YRI individuals. The initial experiment included a total of 141 nsSNPs predicted to be DE, but the assay performed was new and proved noisy. As a result it was possible to confirm and analyse signals for 28 nsSNPs, after filtering for association $r^2 > 0.27$ and ASE mean RNA intensity > 12. For heterozygous individuals at each nsSNP, I assigned relative expression of the two alleles and subsequently compared the experimentally derived relative abundance (ASE results) with the predictions of relative abundance from the genotypic association test. Predicted and experimentally-quantified relative expression of nsSNP alleles were in agreement for 89% (16 out of 18) and 90% (9 out of 10) of nsSNPs tested in the CEU (Figure 17 a) and the YRI populations (Figure 17 b) respectively. This is in agreement with the estimated FDR and suggests strongly that the relative abundance of alternative coding transcripts can be inferred reliably by genotypic associations.

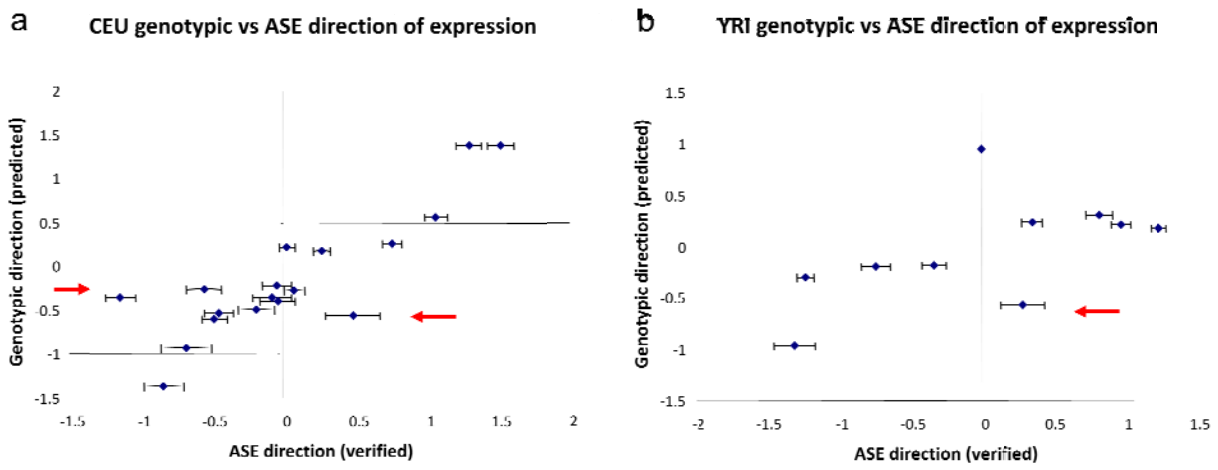


Figure 17. Comparison of statistically predicted and experimentally verified direction of nsSNP allelic effects. The predictions of the nsSNP association test were in agreement with the experimentally verified direction of expression in **a)** 89% and **b)** 90% of the cases studied in the CEU and YRI populations respectively. Red arrows point to the cases where association predictions did not agree with allele-specific expression (ASE) results.

3.4.3 Properties of differentially expressed nsSNPs

To assess the potential biological impact of DE nsSNPs I compared three functional attributes of amino acid substitutions arising from DE nsSNPs and non-DE nsSNPs (testing nsSNPs with MAF ≥ 0.05, to assess common nsSNP consequences). I investigated: 1) the relative position of substitution on the peptide, as different effects may arise depending on whether the nsSNP is at the beginning or the end of the peptide (Figure 18 a), 2) the resulting change in peptide hydrophobicity which may alter the interactions of a protein (Kyte and Doolittle 1982) (Figure 18 b) and 3) the resulting change in Pfam score (a measure of amino acid profile in each position of a protein domain) (Finn, Tate et al. 2008), which assesses the integrity of protein domains that are evolutionary conserved and likely to harbour important functions (Figure 18 c). In all cases the properties of DE nsSNPs were not different from those of non-DE nsSNPs (M-W p-value ≥ 0.05). Though indirect and not comprehensive, this finding

suggests that DE nsSNPs may be a random subset of nsSNPs. If these variants have a functional impact, this will be modified (magnified or masked) by the regulatory variant tagged by the eQTL.

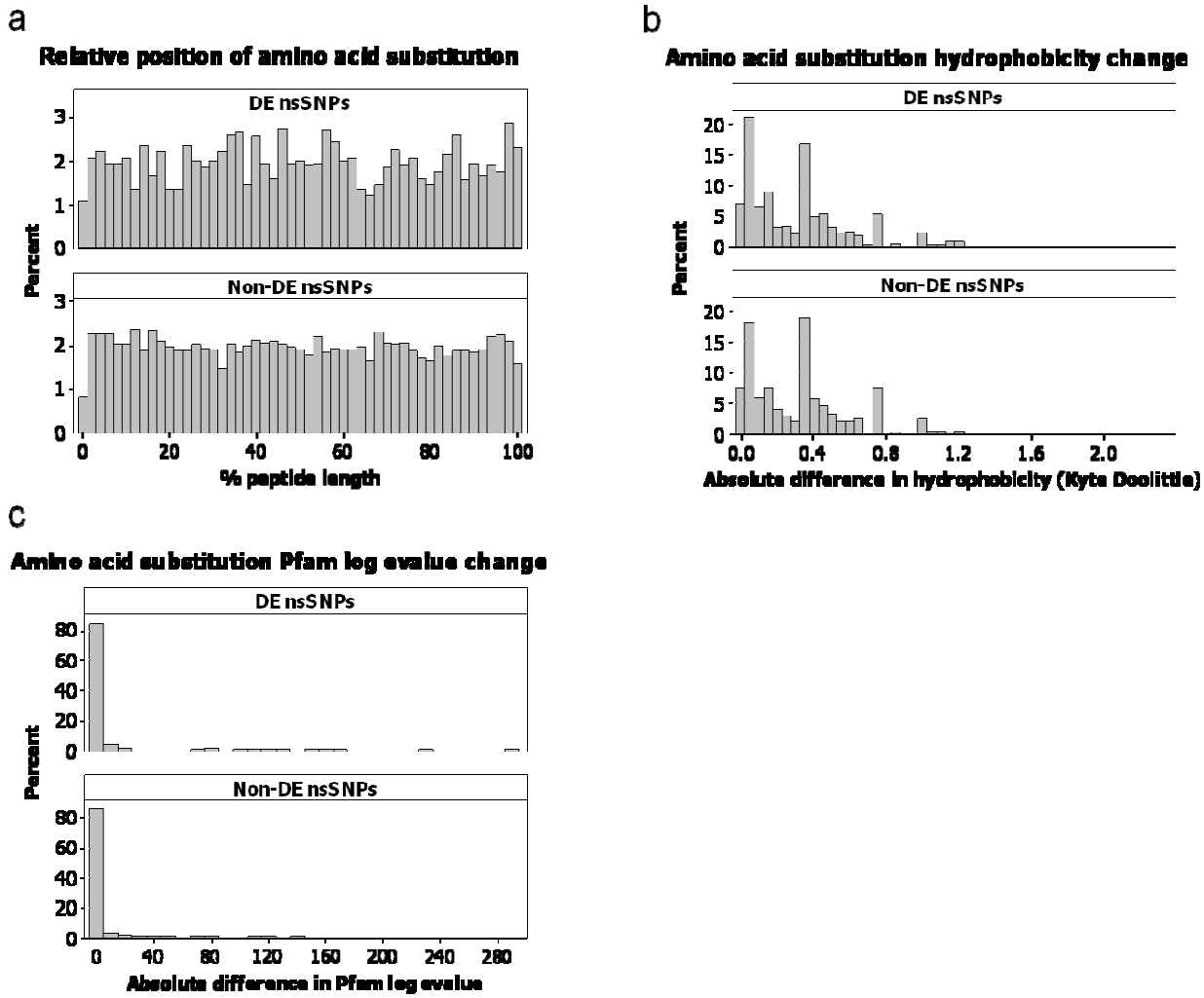


Figure 18. Comparison of biological properties of differentially expressed (DE) vs. non-DE nsSNPs. Three functional attributes of the amino acid substitutions resulting from DE nsSNPs vs. non-DE nsSNPs were compared: **a)** relative position of substitution on the peptide, **b)** resulting change in peptide hydrophobicity and **c)** resulting change in Pfam score when searched against the Pfam profile Hidden Markov Model library. In all cases Mann-Whitney (M-W) tests did not reveal a significant difference between DE and non-DE nsSNPs (M-W p-value ≥ 0.05) and DE nsSNPs appear to be a random subset of nsSNPs. Therefore, if a random nsSNP has a phenotypic effect, this is likely to be magnified or masked through differential expression driven by cis-acting regulatory variants.

To assess how many DE nsSNPs have a known function, I explored the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/omim/) and found that 71 (out of 1,502) DE nsSNPs had an OMIM entry (OMIM nsSNPs, the genes they map in and the predicted health impact are shown in the Appendix). DE nsSNPs were found to map in genes with a role in cancer susceptibility (*BRAC1* (+113705), *BARD1* (+113705)), asthma and obesity (*ADRB2* (+109690)), Crohn disease (CD) (*DLG5* (*604090)), myokymia (*KCNA1* (*176260)), diabetes (*OAS1* (*164350)), chronic lymphatic leukaemia (*P2RX7* (*602566)) emphysema and liver disease (*PI* (+107400)), severe keratoderma (*DSP* (+125647)), and familial hypercholesterolemia (*ABCA1* (+600046)). In some cases the functional role of the nsSNP is unclear and the noise in reported functional effects in OMIM is well-known and difficult to assess in a study such as the present. However there are examples where specific effects have been attributed to nsSNPs. For example, rs28931610 in *DSP* is predicted to change disulphide bonding patterns and alter the peptide tertiary structure, rs28933383 in *KCNA1* causes a substitution in a highly conserved position of the potassium channel and is predicted to impair neuronal repolarization, rs28937574 in *P2RX7* is a loss of function mutation associated with chronic lymphatic leukaemia, rs28931572 in *PI* entails a replacement of a polar for a non-polar amino acid and is predicted to disrupt tertiary structure of the protein, and rs2230806 in *ABCA1* is associated with protection against coronary heart disease in familial hypercholesterolemia. The modulation of such strong effects by cis regulatory variation may increase the complexity and severity of the biological impact.

3.5 eQTL-nsSNP epistatic effect in trans

Thus far I have presented evidence for a modification effect in cis. In cases where the gene containing the DE nsSNP has downstream targets, then it is likely that the expression of target genes is also affected. The aim of this analysis was to test for the

genome-wide effects of this interaction directly, in a statistical framework. To do this I carried out ANOVA to test the main effects of eQTLs and nsSNPs as well as their interaction term (eQTL x nsSNP) on genome-wide gene expression. The rationale behind this approach is that if an eQTL-nsSNP interaction is biologically relevant, its effect may influence gene expression in trans. The power to detect an interaction is maximized when all combinations of genotypes are present, each at appreciable frequencies in the population. To increase power of interaction detection, rare homozygotes were pooled with heterozygotes into a single genotypic category, creating a 2x2 table of genotypes (section 2.6.7). This does not introduce bias in the test statistic as shown by permutations below. Analyses were performed for the CEU population as CHB and JPT population samples were small (45 individuals) and YRI have shown low levels of trans effects in previous studies (Stranger, Nica et al. 2007). I tested 22 eQTL-nsSNP pairs with low LD ($D' \leq 0.5$) and a MAF ≥ 0.1 for both SNPs, against genome-wide expression. At the 0.001 nominal p-value threshold, roughly 331 significant associations are expected (assuming a uniform distribution of p-values) for the interaction term. I detected 412, which corresponds to an estimated FDR of 80%. This is an overall weak signal, but the signals at the tail of the distribution appear to be real given the limited power of this analysis (Figure 19 a).

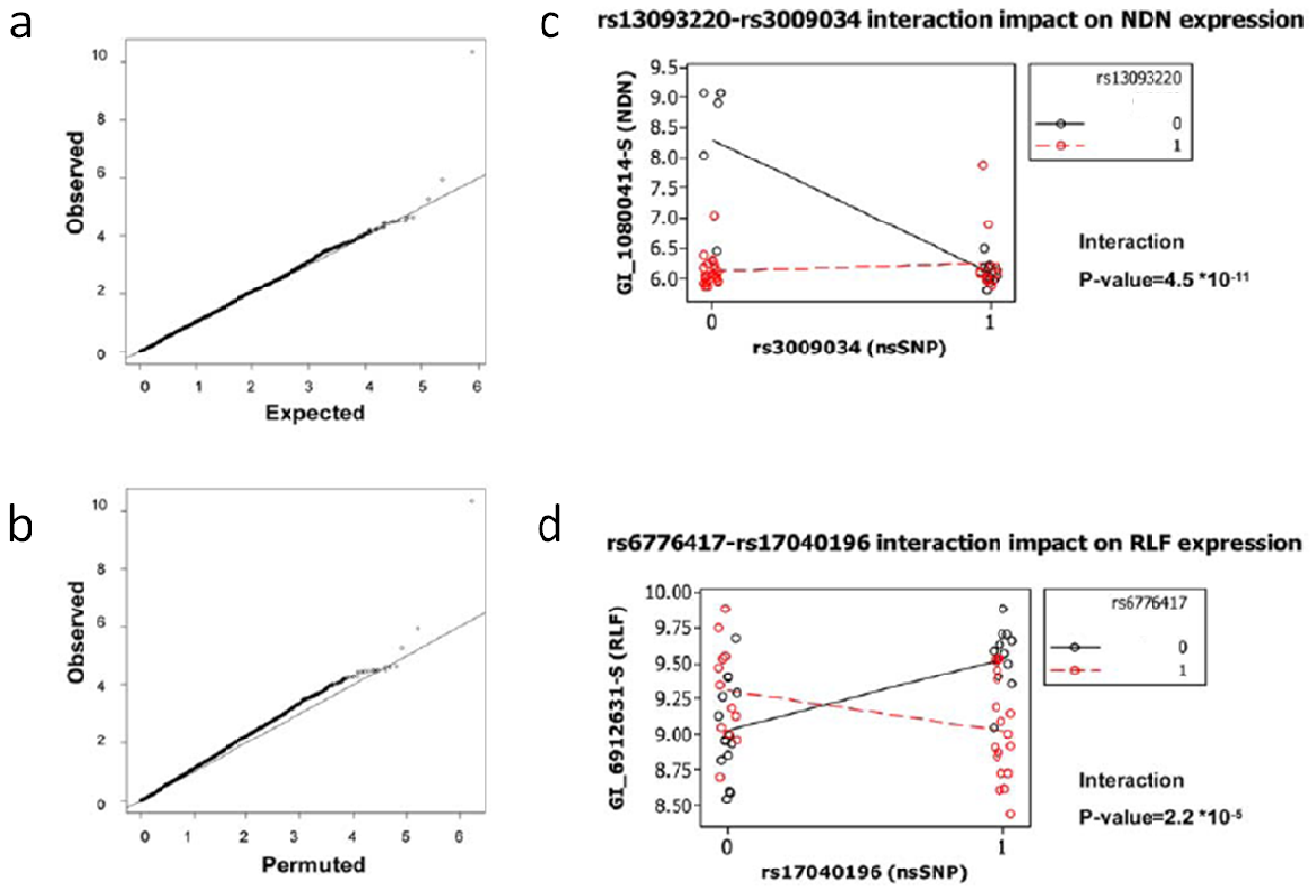


Figure 19. Impact of eQTL-nsSNP genetic interaction on trans gene expression. a) QQ plot of observed vs. expected $-\log_{10}$ p-values of the interaction term from analysis of variance (ANOVA) under the assumption of a uniform distribution of expected p-values. **b)** QQ plot of observed vs. permuted $-\log_{10}$ p-values of the interaction term from ANOVA. **c)** The interaction between rs13093220 (eQTL) and rs3009034 (nsSNP) on chromosome 3 is associated with changes in expression of gene *NDN* (probe ID GI_10800414-S) on chromosome 15 (interaction p-value = $4.5*10^{-11}$). **d)** The interaction between rs6776417 (eQTL) rs17040196 (nsSNP) on chromosome 3 is associated with changes in expression of gene *RLF* (probe ID GI_6912631-S) on chromosome 1 (interaction p-value = $2.2*10^{-5}$).

To test for potential biases in the statistic used, I carried out the same tests using permuted gene expression values (a single permutation was performed by maintaining the correlated structure of gene expression data, see section 2.6.7) relative to the eQTL-nsSNP genotypes. I explored the p-value distribution of the eQTL-nsSNP interaction for observed and permuted data (Figure 19 b) and found an abundance of low p-values in

the observed data. There appears to be some degree of p-value inflation in the observed data relative to the permuted data which is most likely due to correlations in gene expression values. However this does not affect the enrichment of p-values seen at the tails of the observed distribution relative to distributions from expected and permuted values. The observed results therefore show enrichment relative to a uniform distribution of p-values (permutation was not performed to assess significance thresholds, but to assess enrichment of tests with low p-values in the observed data). To further evaluate the robustness of the interactions, I repeated the analysis for the top ten eQTL-nsSNP significant pairs against their corresponding trans-associated gene expression phenotype, after permuting eQTL genotypes relative to nsSNP genotypes and gene expression values. As expected, the significance of the interaction term vanishes in the permuted data. The conditional effects of alleles at the eQTL and nsSNP loci can therefore have a very different impact on the expression of other genes in the cell. This conditional effect on gene expression is illustrated in Figure 19 c and Figure 19 d which show two examples of eQTL-nsSNP interactions (interaction term p-values = 4.5×10^{-11} and 2.2×10^{-5} respectively). In Figure 19 c rs3009034 has an effect on gene expression of gene *NDN* only if the genotype of rs13093220 is homozygous for the common allele. The phenotypic effect of such interactions is even more prominent in Figure 19 d where opposite directions of the effect of rs1704196 are observed. Table 7 shows summary statistics and specific information of SNPs and genes for the ten most significant interactions with a trans effect.

eQTL	nsSNP	source chr	eQTL loc	eQTL MAF	nsSNP loc	nsSNP MAF	trans gene Hugo	trans chr	eQTL p-value	nsSNP p-value	Interaction p-value
rs13093220	rs3009034	3	75853445	0.453	75864268	0.242	NDN	15	1.04E-04	3.41E-05	4.56E-11
rs2280902	rs2272720	8	2064479	0.433	2008828	0.467	TFF2	21	1.88E-02	3.35E-03	1.20E-06
rs2280902	rs2272720	8	2064479	0.433	2008828	0.467	Hs.525661	15	1.03E-01	1.16E-02	5.41E-06
rs6776417	rs17040196	3	14674018	0.267	14720861	0.35	RLF	1	1.46E-01	2.89E-01	2.22E-05
rs6776417	rs6790129	3	14674018	0.267	14730621	0.342	RLF	1	1.46E-01	2.89E-01	2.22E-05
rs6776417	rs17040196	3	14674018	0.267	14720861	0.35	C6orf57	6	3.40E-01	8.91E-02	2.36E-05
rs6776417	rs6790129	3	14674018	0.267	14730621	0.342	C6orf57	6	3.40E-01	8.91E-02	2.36E-05
rs6776417	rs17040196	3	14674018	0.267	14720861	0.35	Hs.121413	2	8.84E-01	8.96E-01	3.10E-05
rs6776417	rs6790129	3	14674018	0.267	14730621	0.342	Hs.121413	2	8.84E-01	8.96E-01	3.10E-05
rs2280902	rs2272720	8	2064479	0.433	2008828	0.467	hmm11130	2	3.51E-01	6.33E-02	3.17E-05

Table 7. eQTL-nsSNP pairs with the most significant interaction effects in trans. Summary statistics and information about mapping location as well as source and trans-affected genes are shown. (chr: chromosome, loc: location, MAF: minor allele frequency)

3.6 Conclusions

I have presented a biological framework to interrogate functional genetic variation by focusing on a specific case of epistasis between regulatory and protein-coding variants. I demonstrated that regulatory variants may have an impact on the protein diversity of cells by differentially modulating the expression of protein-coding variants. In cis, regulatory variants can amplify or mask the functional effects of protein-coding variants. If the coding variant has a role in disease, such an interaction is likely to result in a milder or more severe phenotype to the one expected if only the protein-coding variant were present. Cis interactions were also shown to affect the expression of other genes in the cell in a trans effect, revealed only if an interaction between variants is specifically tested for.

The conditional and context-dependent effects of alleles of variants are likely to have important consequences for complex and quantitative phenotypic traits (Flint and Mackay 2009). In this study I put forth a biological framework for considering and conditioning existing disease associations on known regulatory and protein-coding

variants, in an approach that also provides a potential explanation for the differential penetrance of known disease variants. The abundance of cis regulatory and protein-coding variants in human populations and the generic nature of this type of epistatic interaction (no assumptions made about specific biological pathways) makes it likely that such interactions are common genetic factors underlying complex traits and their consideration is likely to reveal important associations that have not been detected to date. Furthermore, this consideration is particularly important for studies that fail to replicate primary disease associations in newly tested populations, since some of the failures may be due to differential frequency of modifier alleles between the first and second population. Consideration of such interactions may assist in better interpretation of non-replicated signals.

4 FINE-SCALE ARCHITECTURE OF THE CIS REGULATORY LANDSCAPE

In this chapter I will:

- Discuss that LD is a useful property of the genome for association studies at the large scale, but that it can impede the fine-mapping of functional variants.
- Outline a number of approaches employed to enable localization and identification of functional variants.
- Present a strategy used to scan all cis eQTLs detected for a given gene and to identify those that tag independent effects on gene expression.
- Describe the genetic architecture of the cis regulatory landscape and show that multiple regulatory elements can interact to regulate expression in cis.

4.1 FROM GENOME-WIDE ASSOCIATION HITS TO FUNCTIONAL VARIANTS

The power of a SNP to show association with a phenotype is related to its correlation coefficient with the causal variant (Ioannidis, Thomas et al. 2009). This correlated structure of variants in the genome has made it possible to carry out GWAS and identify a plethora of associations between genetic variants and complex traits. However, the variants discovered are not necessarily the ones that give rise to phenotypes, but are more likely tags of functional drivers. Furthermore, when a locus is identified by SNP association, the causal mutation itself need not be a SNP (Altshuler, Daly et al. 2008). For example variants in the *IRGM* gene were found to be associated with CD, but subsequent analysis indicated that the causal mutation is most likely a

deletion upstream of the promoter affecting tissue-specific expression (McCarroll, Huett et al. 2008).

Using GWAS-detected regions as a starting point, the field is currently focusing on strategies for the localization and identification of true functional variants. It is only when these variants are discovered that it will be possible to piece together the biological pathways and processes sculpting complex traits and disease risk. Fine-mapping and identification of functional variants is not an easy task as the correlated structure between variants can impede fine-mapping, with patterns of LD determining the number of markers required to detect and fine-map an association (Mackay, Stone et al. 2009). If a group of markers is in high LD, it is only necessary to genotype one of them as a proxy for all others in the LD block. In pure breeds of dogs for example, where LD blocks are large, only a few markers are required to detect candidate regions. However it is not possible to localize functional variants precisely using this approach (Sutter and Ostrander 2004). In species such as *Drosophila*, LD declines rapidly over short physical distances and knowledge of all sequence variants is necessary for association mapping (Carbone, Jordan et al. 2006), but localization of variants with an impact on the phenotype is precise. Given the extent of LD in humans, genetic variants are likely to have a number of close proxies (Slatkin 2008). A detailed survey of 5 Mb of the human genome (Encyclopedia of DNA Elements or ENCODE regions) genotyped and sequenced in HapMap individuals, revealed that over half of all common SNPs have at least 10 other SNPs in their proximity with an $r^2 > 0.8$ (International HapMap Consortium 2005).

Fine-mapping established associations involves selecting a set of non-redundant SNPs that are in perfect, or near perfect correlation (Ioannidis, Thomas et al. 2009). The rationale behind this approach is that one of the variants selected is the functional driver of the phenotype. Consequently, fine-mapping requires detailed knowledge of variation. Currently the most complete catalogue of human genetic variation is the

HapMap Phase 2, (four million SNPs genotyped for four geographically distinct populations), which covers roughly 30% of common variants. A much more detailed assay of variation will be provided by the ongoing 1000 Genomes Project which involves sequencing the genomes of 1,000 individuals (http://www.1000genomes.org). Deeper sequencing will subsequently reveal rarer variants (International HapMap Consortium 2005).

GWAS interrogating regulatory variation are also faced with the same issues when it comes to localization of functional variants. Association studies of SNP genotypes with transcript levels reveal that for most genes multiple cis eQTLs exist (Stranger, Nica et al. 2007; Dimas, Stranger et al. 2008; Dimas, Deutsch et al. 2009). In such cases, it is likely that most variants mapping to the same genetic locus and are in high LD do not tag independent regulatory effects. On the contrary, SNPs with promising association signals are those that are not in LD and are expected to contribute independent effects to the phenotype of interest (Ioannidis, Thomas et al. 2009). Single loci however may harbour multiple independent functional variants, as is the case of chromosome 8q24 which contains seven independent risk alleles for prostate cancer (Haiman, Patterson et al. 2007).

4.2 NARROWING DOWN THE REGION OF INTEREST

Mapping eQTLs has two components: detection and localization (Mackay, Stone et al. 2009). eQTL detection depends on effect sizes and allele frequencies and delimits a broad genomic region harbouring regulatory elements. Localization or fine-mapping of eQTLs depends on the recombination frequency between regulatory elements and markers. Many approaches have been employed to fine-map eQTLs mostly by narrowing down the region likely to harbour the regulatory variant. In general, the smaller the space outlined by significant associations, the narrower the region that has

to be surveyed for variation, although this can be complicated by local patterns of LD, population history and non-genetic factors. Despite all this, one of the ways forward is to make use the properties of the genome (e.g. information about recombination hotspot intervals) and integrate data from various fields to limit the size of the genomic space to be scanned. Some approaches employed thus far are discussed below.

One approach, employed by Veyrieras et al (2008) who studied HapMap Phase 2 LCLs, involved taking the position of the most significant SNP as an estimate of the location of the functional site. The authors point out that this is only a rough proxy and that these SNPs are unlikely to be true functional variants since: a) HapMap Phase 2 contains only about a third of common SNPs, b) some significant SNP associations may arise if the SNP is in LD with CNVs and c) non-functional SNPs in strong LD with the causal SNP may have lower p-values just by chance. A Bayesian hierarchical model incorporating information about the physical location of SNPs, as well as SNP functional annotation was used to create a high-resolution map of cis regulatory variation. Thirty three percent of most significant eQTLs were found to map within 10 kb of the TSS, and immediately upstream of transcription end site (TES). The former are likely to be polymorphisms that affect the strength of TF binding sites and influence the rate of transcription. The latter may have an impact on microRNA binding and subsequent transcript degradation. eQTLs were also found to be more frequent in exons compared to introns, suggesting that these polymorphisms may affect transcript stability or rate of degradation.

Another study interrogating cis regulatory variation employed allelic expression to measure the relative expression of alleles within a sample, assaying both primary (unspliced) transcripts and mRNA (Pastinen and Hudson 2004). This approach yields direct (vs. statistically inferred) relationships between SNPs and cis regulatory differences (Verlaan, Ge et al. 2009), but does not detect differences in transcript levels driven by variants unlinked to the primary transcript. Allelic expression screening in

LCLs and primary osteoblasts revealed that even for genes that were expressed in both tissue types, identical haplotypes exerted different effects in ~ 50% of the cases. Therefore the same haplotype can display different regulatory effects depending on the tissue it is acting in. (Note that in this study each tissue type originated from one of two populations. Both populations however were of Northern European origin).

A third study investigated the relationship between expression levels of 4,200 genes and proportion of European ancestry in LCLs from African American individuals (Price, Patterson et al. 2008) who inherit variable proportions of African and European ancestries. It was shown that expression differences in individuals of different ancestry proportions reflect expression differences between African and European populations. Using information on an individual's ancestry at the location of a gene whose expression was being analysed, ancestry effects were employed to quantify the relative contributions of cis and trans regulation of human gene expression. The authors estimated that 12 ± 3% of all heritable variation in human gene expression is due to cis variants. However, as they point out, distinction between cis and trans was somewhat imprecise due to the extended length (> 10Mb) of segments of continental ancestry in African Americans.

The examples above illustrate that association analyses testing marker panels cannot differentiate causal SNPs from proxies. Identifying causal variants will be aided by obtaining a more complete catalogue of genetic variation (e.g. 1000 Genomes), but also by cataloguing variants with a functional role on a genome-wide scale. This is the aim of the ENCODE Project (Birney, Stamatoyannopoulos et al. 2007), whose ultimate goal is to find all functional elements in the genome across different cell types. In its pilot phase, a number of techniques were employed to analyse 1% (30Mb) of the human genome.

With the same aim in view, two recent studies focused on identifying regulatory elements across the genome. In the first study Heintzman et al (2009) used a chromatin

immunoprecipitation (ChIP)-based microarray method to identify promoters, enhancers and insulators in multiple cell types and investigate their role in cell type-specific gene expression. Over 55,000 potential transcriptional enhancers were identified, marked with highly cell type-specific histone modification patterns. The patterns detected correlated strongly to cell type-specific gene expression programmes on a global scale and were functionally active in a cell type-specific manner. In contrast, the chromatin state at promoters, as well as binding of CTCF (a major protein involved in insulator activity), were largely invariant across diverse cell types. The second study used *in vivo* mapping of p300 binding to identify regulatory sequences that control the spatial and temporal expression of genes (Visel, Blow et al. 2009). p300 is a near-ubiquitously expressed transcriptional co-activator and a component of enhancer-associated protein assemblies. ChIP of p300, followed by massively parallel sequencing led to mapping of several thousand p300 binding sites in mouse embryonic forebrain, midbrain and limb tissue. Eighty six of the identified sequences were tested in a transgenic mouse assay and enhancer activity was detected in nearly all cases.

In this study I dissected the fine-scale architecture of the cis regulatory landscape using eight of the eleven HapMap Phase 3 populations. I designed and applied a strategy to filter all cis eQTLs detected for a given gene and identify those that tag independent regulatory elements. I also explored the extent to which pairs of interacting variants shape expression levels in cis to highlight the complexity and multidimensionality of gene regulation. At the time of writing this work was in preparation for publication.

4.3 HAPMAP PHASE 3 CIS EQTLS

With the availability of additional populations, as well as additional individuals per population, HapMap Phase 3 provides greater power for eQTL detection within and

across populations. SRC was used to test for association in cis between SNP genotypes (of approximately 1.2 million SNPs per population) and transcript levels of 18,226 Ensembl genes, independently in each population and considering only unrelated individuals (Table 8). All SNPs mapping in a 2 Mb window, centred on the TSS of genes were tested and correction for significance was through permutations. Gene expression for GIH, LWK, MEX and MKK was PCA-corrected and analysed against non-PCA-corrected genotypes (see section 2.3.2.1). This work was carried out in collaboration with Barbara Stranger and Stephen Montgomery at the WTSI.

	SNPs	Probes	Total tests in cis
CEU	1,223,718	21,800	26,690,513,298
CHB	1,115,926	21,800	24,339,461,986
GIH	1,174,223	21,800	25,598,061,400
JPT	1,096,051	21,800	23,905,968,361
LWK	1,249,643	21,800	27,242,217,400
MEX	1,163,286	21,800	25,359,634,800
MKK	1,284,097	21,800	27,993,314,600
YRI	1,306,038	21,800	28,485,994,818

Table 8. HapMap Phase 3 SNPs, probes and total association tests performed in cis.

At the 0.01 permutation threshold of significance, roughly 180 genes are expected to have one significant association by chance. We detected 657, 774, 698, 795, 773, 472, 947 and 799 genes in CEU, CHB, GIH, JPT, LWK, MEX, MKK and YRI populations respectively (estimated FDR of 20-40%) (Table 9). From a non-redundant union of 3,130 gene associations (18% of all genes tested) 1,074 (34%) were shared in at least two populations and 63 (2%) had a significant association in all eight populations.

	0.01 permutation threshold	
	genes	**FDR**
CEU	657	0.28
CHB	**774**	**0.24**
GIH	698	0.26
JPT	**795**	**0.23**
LWK	773	0.24
MEX	**472**	**0.39**
MKK	947	0.19
YRI	**799**	**0.23**
Non-redundant	**3,130**	
> 2 populations	1,074	
8 populations	**63**	

Table 9. HapMap Phase 3 cis significant gene associations.

To explore the location and strength of cis eQTLs for each of the eight populations, the distance of the most significant cis eQTL per gene was mapped relative to the TSS. In agreement with previous studies (Stranger, Nica et al. 2007; Veyrieras, Kudaravalli et al. 2008) a strong signal was found close to the TSS, with no discernable trend in a 5′ or 3′ direction (Figure 20). This symmetrical trend has also been documented in the analysis of the ENCODE Consortium (Birney, Stamatoyannopoulos et al. 2007) and is likely to reflect variation in core regulatory sequences such as promoter elements.

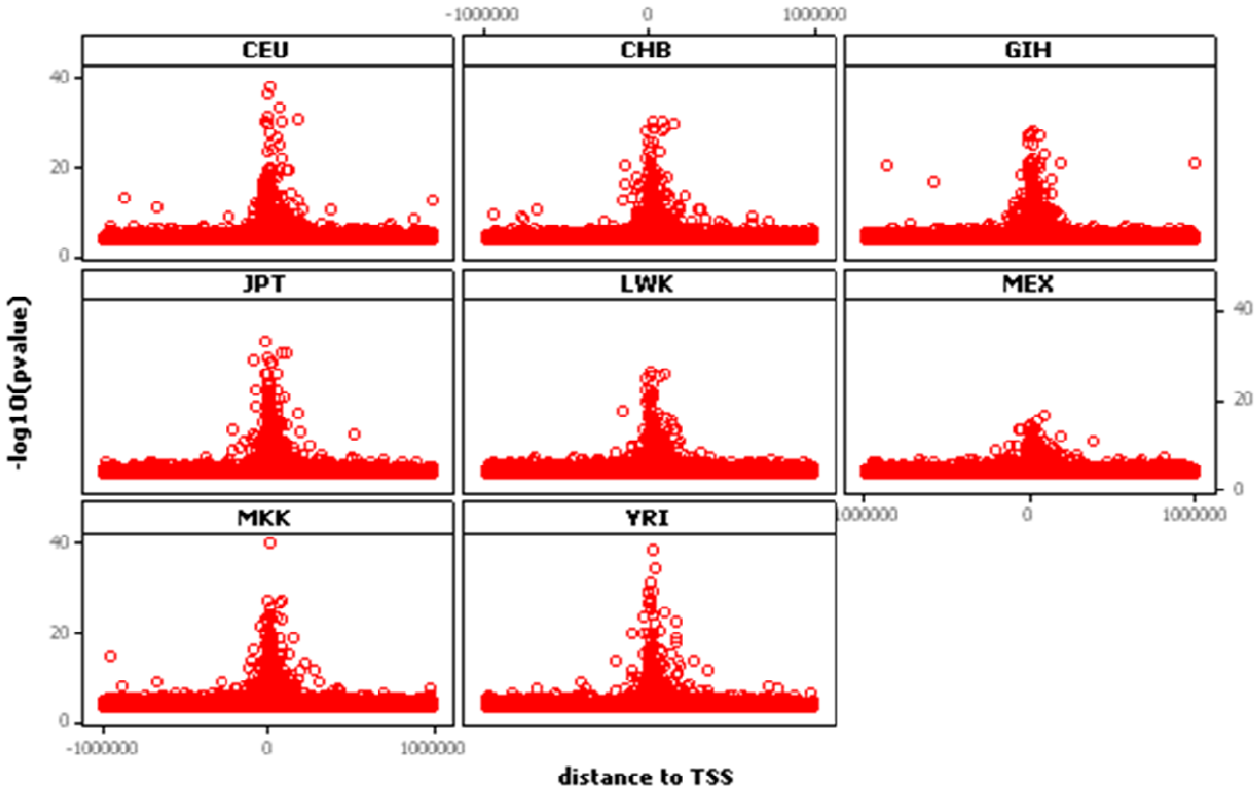


Figure 20. Distance (in bases) of the most significant cis eQTL per gene to the transcription start site (TSS). For HapMap Phase 3 populations (0.01 permutation threshold) the strength and abundance of cis eQTLs decrease with increasing distance from the TSS.

4.4 INDEPENDENT REGULATORY INTERVALS

Over half of the genes with a significant association at the 0.01 permutation threshold possess more than one SNP with a significant association in each of the eight populations (Figure 21 a and Figure 21 b). Multiple eQTLs identified for a given gene most probably tag the effects of the same regulatory element. Gene regulation however is dependent on the joint action of multiple regulatory elements (Figure 22) and the aim of this study was to identify cis eQTLs that tag independent regulatory effects (independent eQTLs or regulatory intervals).

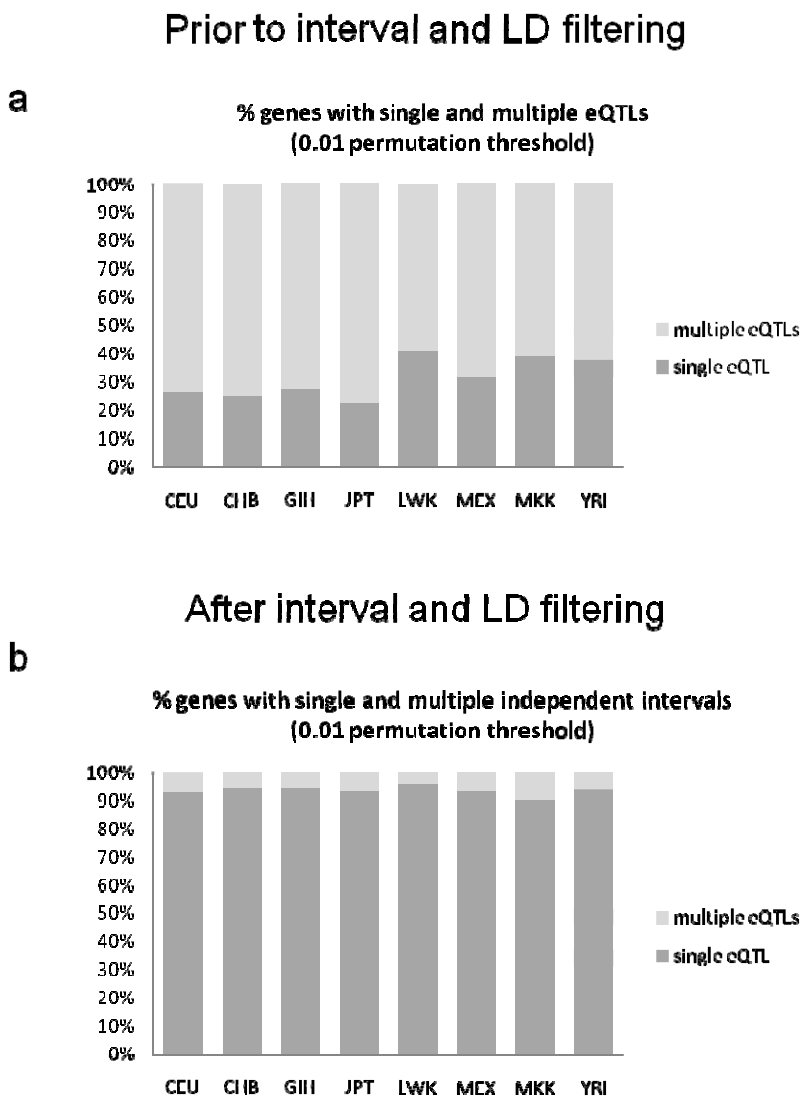


Figure 21. Percent of genes with multiple cis eQTLs and independent intervals. a) shows the % of genes possessing multiple cis eQTLs prior to recombination hotspot interval mapping and LD filtering and **b)** shows the % of genes possessing multiple independent cis eQTLs (intervals) (0.01 permutation threshold).

A detailed description of the strategy employed to do this has been given in section 2.7.1. Briefly, for a given gene eQTLs were mapped in recombination hotspot intervals, the most significant eQTL per interval was retained and remaining eQTLs were filtered further to exclude the least significant variant from variant pairs with a D' > 0.5. This rigorous filtering strategy ensures that surviving eQTLs tag the effects of

independent regulatory elements. Furthermore, since filtering is strict the count of independent cis eQTLs most likely represents the lower bound of the true number of regulatory elements controlling the expression of genes. As expected, the number of cis eQTLs detected for each gene after filtering is much lower (Figure 21 c and Figure 21 d).

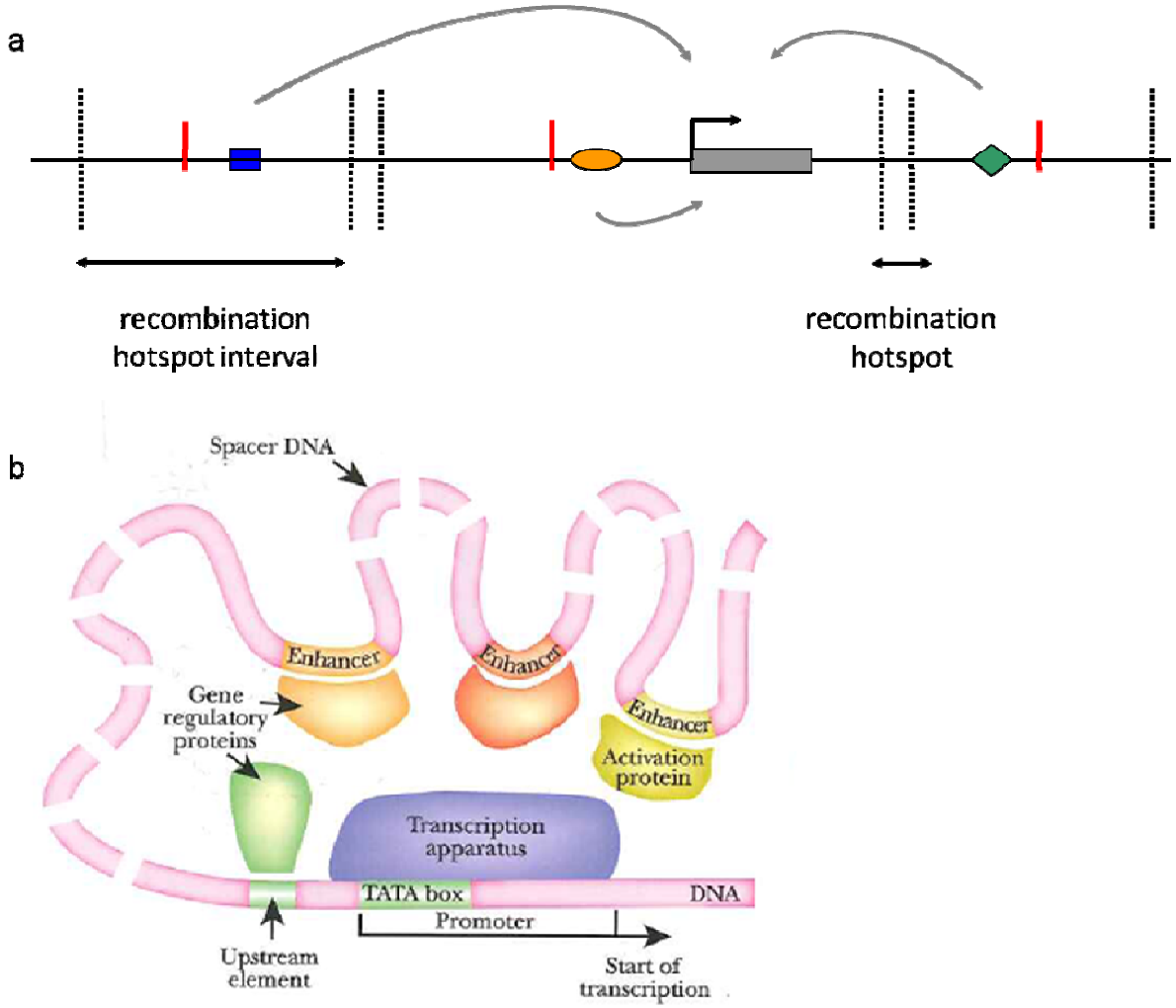


Figure 22. Multiple independent regulatory elements control gene expression. a) Regulatory elements interact with each other to control levels of transcription. In this example independent regulatory elements (with variation in the population) are shown in blue, orange and green and map in different recombination hotspot intervals. The red bars represent SNPs tagging the effects of these elements. **b)** The action of multiple elements controls transcription initiation. Folding of DNA allows numerous activators bound to enhancer sequences to make contact with the basal transcription complex. From (Clark 2005).

At the 0.01 permutation threshold the number of genes possessing multiple independent intervals ranged from 5-10% across the eight populations (Table 10). Specifically, 50 genes with multiple eQTLs (8% of all genes tested), 46 (6%), 44 (6%), 55 (7%), 36 (5%), 34 (7%), 97 (10%) and 52 (7%) were detected for the CEU, CHB, GIH, JPT, LWK, MEX, MKK and YRI populations respectively. Taken together, multiple independent regulatory intervals were detected for approximately seven percent of genes. This observation is in agreement with a mechanism for gene regulation involving the coordinated action of multiple elements (Figure 22).

	0.01 permutation threshold							
No. of intervals	**CEU**	**CHB**	**GIH**	**JPT**	**LWK**	**MEX**	**MKK**	**YRI**
1	607	728	654	740	737	438	850	747
2	**40**	**35**	**43**	**48**	**34**	**30**	**77**	**46**
3	6	9	1	4	2	4	16	3
4	**1**			**2**			**4**	**1**
5	1	2						2
6	**1**							
7								
8								
9								
10								
11	1							
12								
13				1				
Total	**657**	**774**	**698**	**795**	**773**	**472**	**947**	**799**
genes with ≥ 2 intervals	50	**46**	**44**	**55**	**36**	**34**	**97**	**52**
% genes with ≥ 2 intervals	7.61	5.94	6.30	6.92	4.66	7.20	10.24	6.51

Table 10. HapMap Phase 3 independent eQTLs (intervals) at the 0.01 permutation threshold.

To address the extent to which gene activity is controlled by common regulatory sequences across populations, I explored sharing of independent eQTLs (intervals). This was done for all regulatory intervals detected in each population and comparison was not restricted to intervals detected for a given gene (the latter analysis was ongoing at the time of writing). At the 0.01 permutation threshold and from a non-redundant

union 3,288 independent intervals, 2,281 (70%) were found in a single population, 404 (12%) were shared in exactly two populations, 201 (6%), 145 (4%), 84 (3%), 65 (2%), 52 (2%) and 56 (2%) were shared in exactly three, four, five, six, seven and all eight populations respectively. Taken together, roughly 31% of intervals were found in at least two populations (Table 11). The high proportion of intervals detected in only one population suggests that even for the same cell type, genes are regulated to some extent by different regulatory elements across populations. Conversely, sharing of intervals implies sharing of regulatory elements. Relative sharing in ≥ five populations increased with higher significance stringency. The lower degree of sharing at the 0.01 permutation threshold may arise as a consequence of winner's curse (Goring, Terwilliger et al. 2001; Lohmueller, Pearce et al. 2003; Ioannidis 2008) which states that the effect sizes discovered when applying specific statistical significance thresholds are inflated compared to true effect size. Consequently the discovery sample usually achieves higher significance than replication samples. In this analysis the degree of sharing across populations may be underestimated if a gene with a significant association in one population barely fails significance correction in a second population. Sharing is likely to be further underestimated due to the fact that eQTL detection is affected by allele frequency differences across populations. Therefore a regulatory element may be active in multiple populations, but detected via an eQTL only in a fraction of these groups.

	0.01 permutation threshold	
No. populations	No. intervals	%
1	2,281	69.37
2	404	12.29
3	201	6.11
4	145	4.41
5	84	2.55
6	65	1.98
7	52	1.58
8	56	1.70
Total intervals	3,288	
>2 populations	1,007	30.63

Table 11. Sharing of intervals for HapMap Phase 3 cis significant genes.

4.5 eQTL-eQTL INTERACTION IN CIS

As outlined above, the genetic architecture of cis regulatory landscapes is complex with multiple regulatory intervals controlling gene expression. To dissect cis regulatory architecture further, I explored the degree to which interactions between variants in cis affect expression levels. DNA sequences containing enhancer elements for example are known to loop over great distances (> 1 Mb) and make physical contact with regulatory elements close to the TSS, in an interaction that affects initiation and rate of transcription (Figure 22 b). To detect such interactions, I applied a similar strategy to that used in Chapter 3 to test for interactions between regulatory and protein-coding variants (also see section 2.6.7). This analysis was carried out for the CEU and YRI populations.

SNPs with a nominal (uncorrected) p-value < 0.001 from the SRC association test were mapped in recombination hotspot intervals and the most significant SNP per interval was retained. SNP pairs with a D' > 0.5 across intervals were excluded from the analysis. Filtering for permutation significance was not performed to include variants that do not necessarily have large marginal effects on the phenotype, but whose impact

on gene expression may be revealed through an interaction. I carried out ANOVA to test the independent effects of each SNP as well as the SNP x SNP interaction term on gene expression in cis. Assuming a uniform distribution of nominal p-values, at the 0.01 nominal p-value threshold approximately 47 and 79 significant associations are expected by chance for the interaction term in the CEU and the YRI populations respectively. I detected 87 and 131 associations corresponding to an estimated FDR of 54% and 60% respectively (Table 12). At the stricter 0.001 nominal p-value threshold, approximately 5 and 8 significant associations are expected by chance for the interaction term in the CEU and the YRI populations respectively. I detected ten and 22 corresponding to an estimated FDR of 47% and 36% respectively (Table 12). Although this is not a very strong signal, given the strict filtering and relatively low power of this analysis, an enrichment of significant interaction terms is observed.

		0.01 nominal p-value threshold			0.001 nominal p-value threshold		
	Total tests	Expected	Observed	FDR	Expected	Observed	FDR
CEU	4,692	46.92	87	0.54	4.69	10	0.47
YRI	7,866	78.66	131	0.60	7.87	22	0.36

Table 12. Expected and observed significant interaction terms for CEU and YRI.

To explore this signal further, I conducted a single permutation of expression levels relative to genotypes. The p-value distributions of observed and permuted interaction terms were compared and an abundance of low p-values was found in the observed data for both populations (Figure 23). This suggests that gene expression in cis is sculpted to a certain extent by interacting regulatory elements.

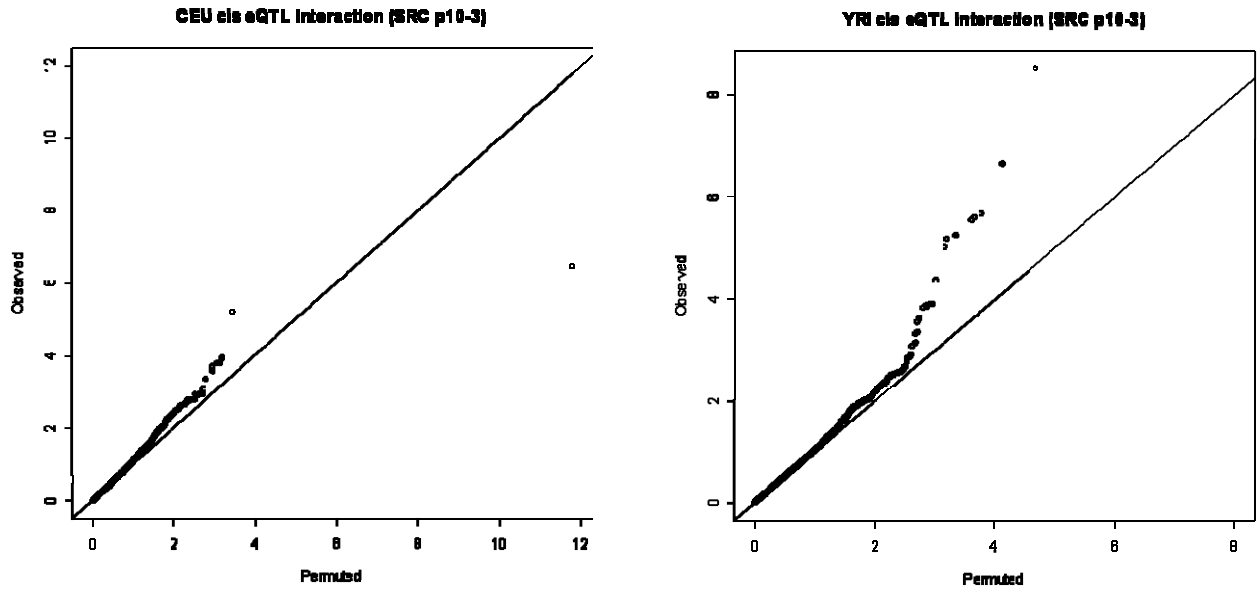


Figure 23. QQ plots of observed vs. permuted cis interaction p-values for the CEU and YRI HapMap Phase 3 populations. The signal at the tail of the observed distributions suggests that interactions between variants in cis influence expression levels of genes.

4.6 CONCLUSIONS

The signals detected in GWAS stem from markers that are not likely to be the causal variants. Furthermore, these markers typically delineate large genomic spaces that harbour causal variants. Replication of signals in independent studies provides corroborating evidence of causality, but the problem of delimiting the space carrying the functional variants remains. In this chapter I have presented a strategy that makes use of the properties of the genome and can be employed to restrict the space likely to contain regulatory elements controlling gene expression in cis. Using eight of the HapMap Phase 3 populations I demonstrated that seven percent of genes (0.01 permutation threshold) across all populations possess multiple independent regulatory intervals. The strategy applied involved strict filtering to remove highly correlated markers likely to tag the same regulatory element. As outlined in section 1.4.1, regulatory element length ranges from a few to a few hundred bp. Recombination

hotspot intervals on the other hand have a median length of 9,000 bp, ranging from a minimum of 998 bp to a maximum 31,495,264 bp. As a result, a single interval may contain multiple regulatory elements. The strategy employed in this study involved selection of the most significant eQTL per interval. Therefore, the number of truly independent eQTLs acting on genes in cis is likely to be higher and the method employed is most probably conservative.

The complexity of the regulatory landscape is further demonstrated through evidence of interactions between genetic variants with a small marginal impact on gene expression in cis. Using the CEU and YRI populations I explored the extent to which SNPs mapping in different intervals jointly affect cis expression levels. Although relatively underpowered, also because the ability to detect an interaction decays substantially when proxies of the functional variants are used, this study presents evidence for a cis interaction between regulatory variants. This approach does not test markers without marginal effects and cannot reveal variants that manifest themselves only in the context of an interaction. Consequently, the extent to which expression is influenced by interactions between variants in cis is likely to be an underestimate. A potentially more informative approach is to test all SNP pairs in the vicinity of a gene. This is currently being explored in collaboration with Doug Speed and Simon Tavaré at the CRI.

This study has highlighted the complex architecture of the cis regulatory landscape. GWAS of phenotypes in which expression levels are likely to play a crucial role should take this observation into consideration. Furthermore, integrating this information with studies on trans gene regulation will help piece together a more complete picture of gene expression control.

5 CELL TYPE SPECIFICITY OF CIS REGULATORY VARIATION

In this chapter I will:

- Underline that most studies investigating regulatory variation to date explore expression in a single cell type.
- Stress the value of documenting cell type-specific regulatory variation.
- Describe a resource and experimental strategy that enable detection of eQTLs across cell types.
- Outline that the majority of eQTLs identified using this resource are cell type-specific.
- Emphasize the value of large collections of LCLs.

5.1 THE VALUE OF STUDYING DIFFERENT CELL TYPES

Variation influencing gene expression can manifest itself as gene expression differences among populations, among individuals in a population, among tissues, and in response to environmental factors. As discussed in the previous chapters, the genetic basis of the first two types of gene expression variation has been investigated in a number of studies with the quantification of mRNA in one tissue and the identification of eQTLs in a single or multiple populations (Adams, Kerlavage et al. 1995; Reymond, Marigo et al. 2002; Su, Cooke et al. 2002). The complex developmental program in higher eukaryotes however results in a vast set of highly specialized cell types, whose fate is determined to a large extent by the combination of expressed genes and their level of expression. During development, but also in differentiated cells, some genes exhibit ubiquitous patterns of expression while others display tissue-specific activity (Myers, Gibbs et al. 2007; Emilsson, Thorleifsson et al. 2008; Schadt, Molony et al. 2008). The extent to which

genetic variation manifests itself as tissue-specific gene expression patterns remains unknown and eQTL cell type specificity remains underexplored. A handful of studies have identified eQTLs in certain human (Myers, Gibbs et al. 2007; Emilsson, Thorleifsson et al. 2008; Schadt, Molony et al. 2008) and mammalian (Cotsapas, Williams et al. 2006; Campbell, Kirby et al. 2008) tissues but a systematic study comparing eQTLs across a wide range of cell types, while controlling for confounding associations, such as population samples and differences in technology or statistical methodology, is lacking in humans. Studies in model organisms however are highlighting the value of interrogating regulatory variation systematically and in a tissue-specific context (Petretto, Mangion et al. 2006; Huang, Shifman et al. 2009).

The importance of documenting cell type-specific regulatory variation is high given the role of gene expression patterns in determining cell type during development, in shaping higher level phenotypes and in determining disease risk. In cases such as asthma (Moffatt, Kabesch et al. 2007) and colorectal cancer (Valle, Serena-Acedo et al. 2008) documenting genetic control of gene expression variation is likely to shed light on mechanisms of disease pathogenesis. Furthermore there is growing evidence that causative variants identified in GWAS are likely to behave in a cell type-specific manner (Wellcome Trust Case Control Consortium 2007). Cataloguing cell type-specific regulatory variation can therefore serve to connect biological pathways controlling cellular activities in health and disease (Emilsson, Thorleifsson et al. 2008; Schadt, Molony et al. 2008; Wu, Delano et al. 2008).

The case of CD, an autoimmune inflammatory disease of the gastrointestinal tract, illustrates the critical role of eQTLs in elucidating disease pathogenesis. GWAS revealed a strong signal in a 1.25 Mb gene desert of chromosome 5p13.1 (Libioulle, Louis et al. 2007; Wellcome Trust Case Control Consortium 2007). Expression association studies quantifying transcript levels in LCLs (Libioulle, Louis et al. 2007), revealed that the same region showed a strong association with transcript levels of

PTGER4. Knockout mice for *PTGER4* have increased susceptibility to colitis, rendering this gene a strong susceptibility candidate for CD (Servitja, Pignatelli et al. 2009).

In cases such as the above, where disease and gene expression signals map to the same chromosomal location (Figure 24 a), integrating information from both sources may provide important clues about the genes and functional pathways involved in disease pathogenesis. CD is an immune system disease and studying expression in immune system-derived LCLs has proven informative in terms of pointing to candidate genes. In this respect LCLs are a relevant cell type to study for CD. In the case of other phenotypes however (e.g. diabetes) expression association signals in LCLs may not yield signals that track disease association (Figure 24 b). Interrogating expression in pancreatic-islet β-cells might provide more clues for the pathogenesis of diabetes (Nica and Dermitzakis 2008).

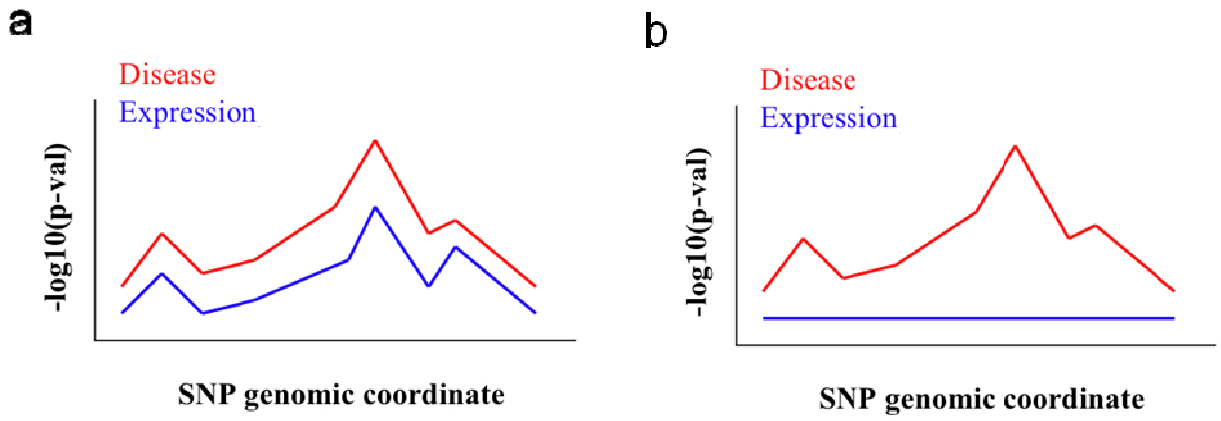


Figure 24. Disease and expression signals from genome-wide association studies (GWAS). The x axis represents chromosomal location, the y axis shows the significance of association for SNPs along the chromosome. In **a)** expression and disease association signals track one another, implying that expression of the particular gene in the cell type studied may be involved in disease pathogenesis. This is the case for Crohn disease (CD) where a SNP on chromosome 5 was associated with expression levels of *PTGER4* in LCLs and also showed a significant association to disease. In **b)** expression and disease association signals do not track one another, implying that expression of the particular gene in the cell type studied is probably not relevant for the disease. Given the important role of gene expression in disease pathogenesis, it is necessary to investigate multiple cell types to determine whether there are cases in which

expression signals mirror those of disease. In this way it will be possible to identify loci with a functional role in disease pathogenesis. Figure adapted from (Nica and Dermitzakis 2008).

It is not clear how straightforward it will be to determine which cell type or tissue is relevant for a particular disease or complex trait. As with candidate gene studies it may turn out that in some cases the relevant cell type is not the one that was identified as a candidate based on existing biological knowledge. Interrogating expression in blood-derived LCLs for example has proven useful for identifying genes implicated in the pathophysiology of autism (Nishimura, Martin et al. 2007). Gene expression profiles from males with autism and non-autistic controls clearly distinguished cases from controls. It is yet not clear how many cell types and tissues will be adequate to provide a catalogue of regulatory variation, but this approach contributes to efforts using functional genomic information to interpret the biological effects of disease or complex trait variants. To date, such efforts are hindered by the limited availability of the relevant cell type to perform the functional assays. Understanding the degree of tissue-specificity of regulatory variation will enable us to assess how much we are missing by interrogating only a limited number of tissues and will provide clues as to how many tissues will be required to capture the spectrum of functional consequences of disease-causing variants (McCarthy and Hirschhorn 2008; Nica and Dermitzakis 2008).

In this study, I assessed cell type specificity of variants impacting gene expression by quantifying mRNA levels in three cell types from each of 85 individuals, and by identifying shared and cell type-specific eQTLs. I also explored the fine-scale architecture of cis regulatory landscapes conditioning on cell type, to determine the extent to which genes are regulated by common or cell type-specific regulatory elements. This work has been described in (Dimas, Deutsch et al. 2009).

5.2 DETECTING CIS EQTLS IN THREE CELL TYPES

Eighty five individuals from the GenCord resource were studied to explore the cell type-specific distribution of cis regulatory variation. GenCord is a collection of cell lines derived from umbilical cords of individuals of Western European origin (see section 2.3.3). Sample collection was performed systematically on full term or near full term pregnancies, to ensure homogeneity for sample age. mRNA levels were quantified in primary fibroblasts, LCLs, and primary T-cells for 48,804 probes using the illumina WG-6 v3 Expression BeadChip array. Data from 22,651 probes, mapping to 17,945 autosomal RefSeq genes (15,596 Ensembl genes) were analysed. The same samples were genotyped on the illumina 550K SNP array. Following quality control (SNPs with missing data were removed) and minor allele frequency filtering (MAF ≥ 5%), 394,651 SNPs were used for association testing. PCA detected ten potential outlier individuals from the genotype data (Figure 25) who were subsequently removed from the analysis. eQTL discovery and all other properties of the results for 75 vs. 85 individuals were almost identical (Figure 26).

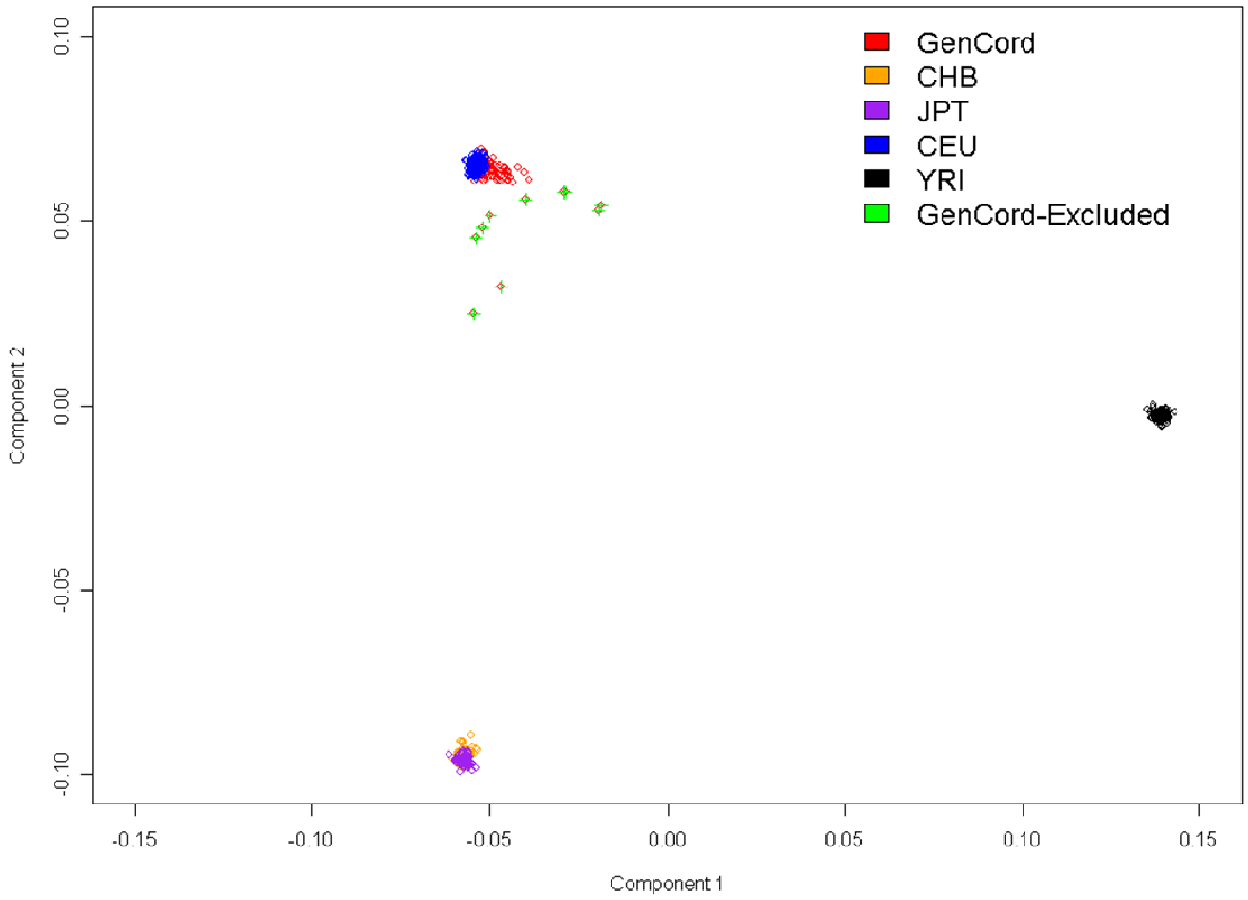


Figure 25. Principal components analysis (PCA) of the GenCord and HapMap Phase 2 populations. GenCord individuals were clustered with the HapMap populations (CEU, CHB, JPT and YRI) to assess relative population stratification in the samples. Given the observed clustering along the first two principal components, ten outliers were removed from the analysis (GenCord-Excluded).

I explored associations in cis, by testing SNPs mapping within a 2 Mb window centred on the TSS of genes. SRC was used to test for association between SNP genotype and mRNA levels, after intensity normalization and $\log_2$ transformation, performed separately for each cell type. A total of 6,083,130 tests were performed and significance thresholds for each gene were assigned through permutations. For 75 individuals at the 0.01 permutation threshold I discovered 2,146, 2,155 and 2,046 genes with significant cis eQTLs in fibroblasts, LCLs and T-cells respectively, with an

estimated FDR of 7%. At the stricter 0.001 permutation threshold, I discovered 427, 442 and 430 genes with significant cis associations in fibroblasts, LCLs and T-cells respectively, with an estimated FDR of 4% (Figure 26). The genomic distribution of detected associations at the 0.001 threshold in each cell type is shown in Figure 27.

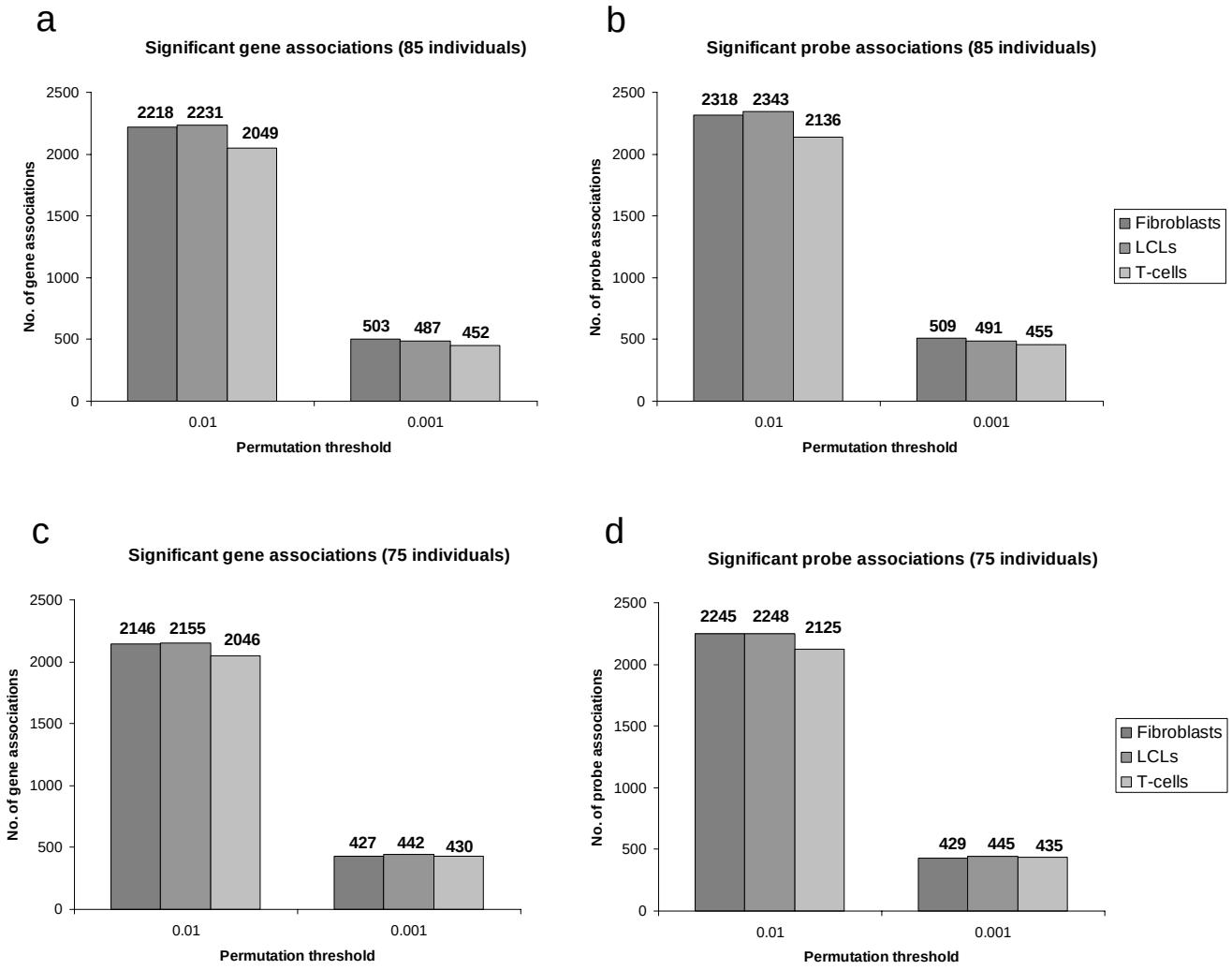


Figure 26. Significant gene and probe associations in GenCord cell types at the 0.01 and 0.001 permutation thresholds. Numbers on top of the histogram bars represent counts of associations. **a)** and **b)** show gene and probe associations detected using 85 individuals and **c)** and **d)** show gene and probe associations detected using 75 individuals, after removal of 10 outliers. Association detection was highly similar in both analyses, with comparable estimated false discovery rates (FDR = 7% for genes and 10% for probes for the 0.01 permutation threshold and 3% for genes and 3-4% for probes at the 0.001 permutation threshold for both the 85 and 75 individuals analyses).

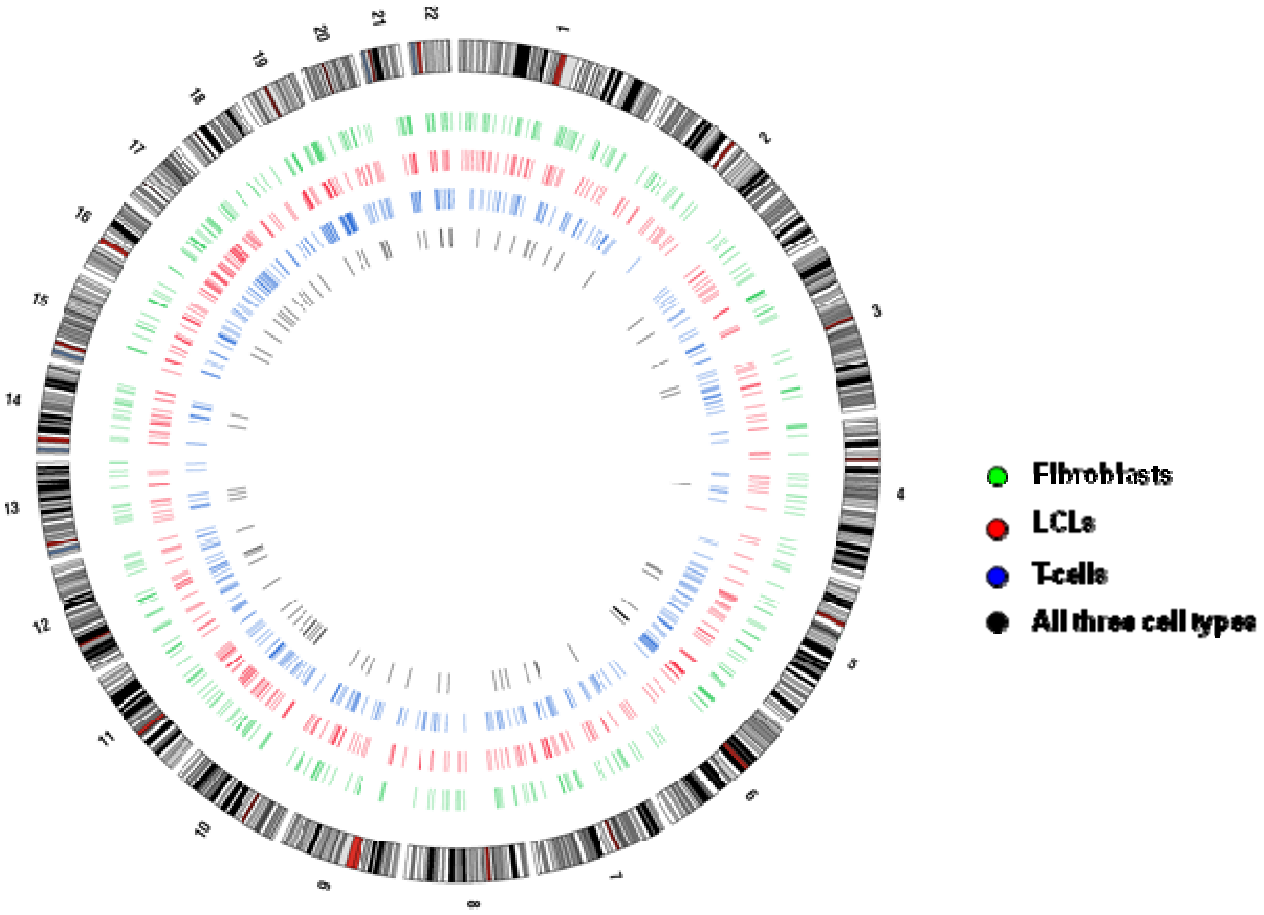


Figure 27. Genome-wide map of cis eQTLs in GenCord three cell types. cis eQTLs at the 0.001 permutation threshold are shown as colour-coded lines on their corresponding chromosomal location. Internal black lines represent genes with eQTLs in all cell types.

5.3 Replication of cis eQTLs detected in LCLs

There has been long debate about the stability of eQTLs detected in LCLs from different samples, experiments and technologies, as well as the use of large collections of these cell lines. In the present study I assessed how well previously described eQTLs from the CEU HapMap Phase 2 (International HapMap Consortium 2007; Stranger, Nica et al. 2007) are replicated in GenCord LCLs. The expectation is that a large proportion of

eQTLs will be shared, as both populations are of European descent and share similar allele frequency spectra. Due to differences in probe sequence content between the illumina v1 array (used for HapMap Phase 2 CEU) and the illumina v3 array (used for GenCord) it was possible to compare a small subset of SNP-probe associations. Comparisons were made for cases where the SNP was present in both HapMap and GenCord and the probe had identical sequence between illumina v1 and v3 expression arrays. Strict filtering was performed to avoid confounding effects arising from: a) differences in probe efficiency, b) the possibility that probes covered alternative splicing products from the same gene and c) the occurrence of probes in the v1 array containing SNPs. Of the 5,898 SNP-probe pairs that survived the 0.001 permutation threshold in HapMap Phase 2 CEU, 137 SNP-probe pairs (44 probes, some associated with multiple SNPs) were also tested in GenCord LCLs. The distribution of nominal (uncorrected) p-values from the association test for these SNP-probe pairs is greatly enriched for very low p-values, with 114 nominal p-values < 0.001 (83%) (Figure 28). Therefore, previously detected eQTLs were well-replicated, despite the long separation time between tests, demonstrating the stability of LCLs. These data highlight the value of large collections of LCLs from different cohorts for studies of gene expression and disease interpretation.

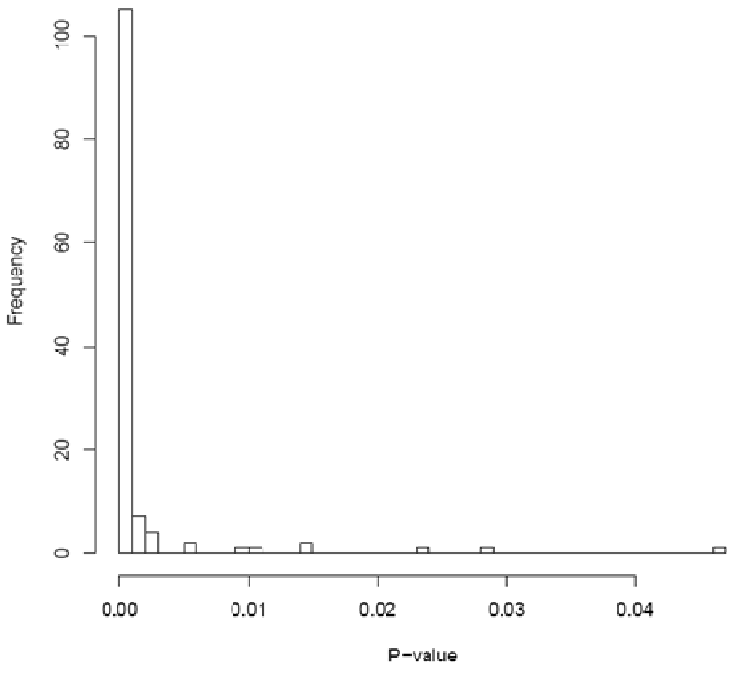


Figure 28. Replication of nominal (uncorrected) p-values in GenCord of SNP-probe associations initially identified as significant in HapMap Phase 2. I tested 137 identical SNP-probe pairs (44 probes) in GenCord LCLs. Of these, 114 SNP-probe pairs (83%) have a nominal p-value < 0.001 in GenCord LCLs, suggesting good replication of eQTLs between experiments.

5.4 SHARING AND CELL TYPE SPECIFICITY OF CIS EQTLS

Having established the robustness of eQTLs through replication, I interrogated the cell type specificity of regulatory effects by exploring genes with cis eQTLs that were: a) shared in all three cell types, b) shared in two cell types and c) cell type-specific. At the 0.001 permutation threshold, I identified a non-redundant set of 1,007 genes with cis eQTLs of which 86 (8.5%) were shared in all three cell types, 120 (12%) were shared in exactly two of the cell types and 801 (79.5%) were cell type- specific (Table 13 for genes, Table 14 for probes; results for the 0.01 permutation threshold are also shown). The proportion of cell type-specific eQTLs was similar to previous estimates of eQTL tissue specificity and alternative splicing reported in a study interrogating two tissue types, sampled however from different groups of individuals (Heinzen, Ge et al. 2008).

		0.01 permutation threshold	% shared	0.001 permutation threshold	% shared
All 3 cell types	Fibroblasts - LCLs - T cells	227	4.5	86	8.5
Exactly 2 cell types	Fibroblasts - LCLs	296	5.9	38	3.8
	Fibroblasts - T cells	**270**	**5.4**	**35**	**3.5**
	LCLs - T cells	288	5.7	47	4.7
Cell type specific	Fibroblasts	1,353	26.9	268	26.6
	LCLs	**1,344**	**26.7**	**271**	**26.9**
	T cells	1,261	25.0	262	26.0
Total significant in each cell type	Fibroblasts	2,146		427	
	LCLs	**2,155**		**442**	
	T cells	2,046		430	
3 cell type union		5,039	100.0	1,007	100.0
Total tested		15,670		15,670	

Table 13. Cell type-shared and specific gene associations. This table shows gene associations that were: i) shared in all three cell types, ii) shared in two cell types and iii) cell type-specific.

		0.01 permutation threshold	% shared	0.001 permutation threshold	% shared
All 3 cell types	Fibroblasts - LCLs - T cells	170	3.0	79	7.7
Exactly 2 cell types	Fibroblasts - LCLs	231	4.1	35	3.4
	Fibroblasts - T cells	**212**	**3.8**	**35**	**3.4**
	LCLs - T cells	225	4.0	49	4.7
Cell type-specific	Fibroblasts	1,632	29.1	280	27.1
	LCLs	**1,622**	**28.9**	**282**	**27.3**
	T cells	1,518	27.1	272	26.4
Total significant in each cell type	Fibroblasts	2,245		429	
	LCLs	**2,248**		**445**	
	T cells	2,125		435	
3 cell type union		5,610	100.0	1,032	100.0
Total tested		22,651		22,651	

Table 14. Cell type-shared and specific probe associations. This table shows probe associations that were: i) shared in all three cell types, ii) shared in two cell types and iii) cell type-specific.

The relative sharing of gene associations across cell types is shown in Figure 29 a and Figure 30 (probe associations shown in Figure 29 b). The degree of gene (and probe) sharing is an overestimate of overlapping genetic effects as expression of genes for which eQTLs were identified in all three or at least two cell types is not necessarily driven by the identical regulatory elements.

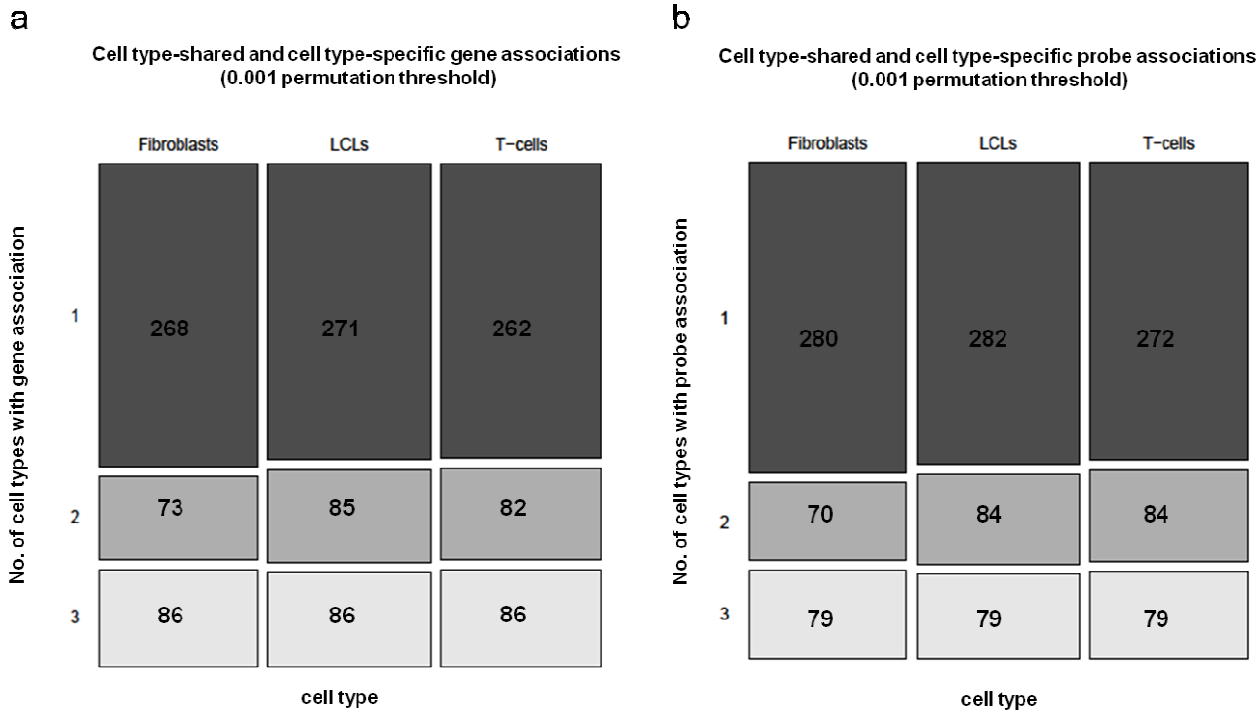


Figure 29. Relative sharing of significant genes and probes in three cell types. Cell type-shared and cell type-specific associations for **a)** genes and **b)** probes (0.001 permutation threshold). Each bar indicates the full fraction of genes or probes for which eQTLs were detected in each cell type. Light grey indicates the fraction of genes/probes with eQTLs overlapping in all three cell types, dark grey indicates the fraction of genes/probes with an overlap in at least one other cell type, and black indicates the fraction of genes/probes with cell type-specific eQTLs.

As expected, a proportion of variation controls expression levels in a similar way across cell types and this most probably reflects regulation of processes common to all cells. At the 0.001 permutation threshold, of the genes with cis eQTLs common to two or more cell types, 124 (12.3%) were shared between fibroblasts and LCLs, 121 (12.0%)

were shared between fibroblasts and T-cells, and 133 (13.2%) were shared between LCLs and T- cells (Table 14). Increased eQTL sharing between LCLs and T-cells is most likely due to the related function and common developmental origin of these cells.

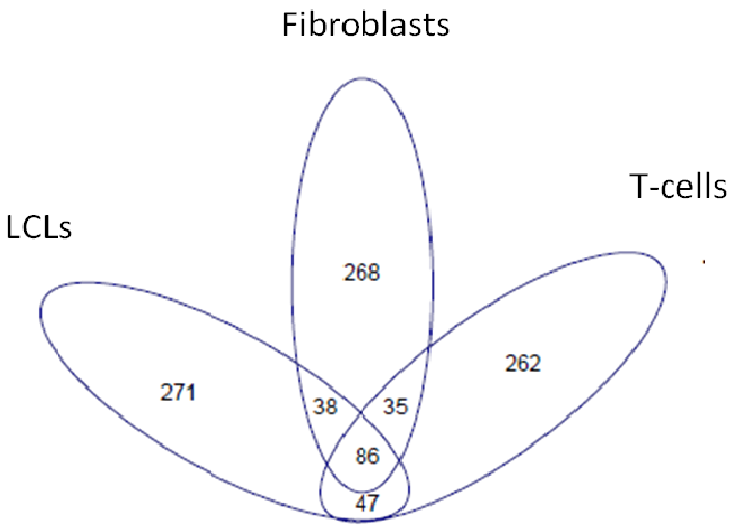


Figure 30. Venn diagram of genes with cis eQTLs in three cell types. Cell type-specific counts, two-way and three-way sharing is shown. Figure by Manolis Dermitzakis.

The most striking result from this analysis is the prominence of cell type specificity. 268 (26.6% of total), 271 (26.9%) and 262 (26.0%) of gene associations were found only in fibroblasts, LCLs and T-cells respectively (Table 13). It is plausible that cell type-specific eQTLs can arise if a gene is expressed in one cell type, but not in another. To test this I explored the medians and variances of gene expression in each cell type, and found that genes with cell type-specific signals had significantly higher expression variance in the cell type where the eQTL was detected (M-W p-value < 0.0001 for all comparisons). Medians of gene expression values for the same genes were either marginally significantly or not significantly different, meaning that all genes included in this analysis were largely expressed in all cell types. Furthermore, it is estimated that all genes with cell type-specific cis eQTLs are expressed to some level in

all three cell types. This suggests that the majority of cell type specificity is not a result of the presence or absence of gene expression between cell types, but is due to differential expression resulting from cell type-specific use of regulatory elements.

5.5 Dissecting eQTL cell type specificity

To dissect the nature of the overlap of cis eQTLs across cell types, I compared the direction of the allelic effect (i.e. assignment of high/low expression to eQTL alleles) between pairs of cell types in cases where SNP-gene associations were significant for both cell types. The direction (sign of Spearman rho) was in complete agreement for all pairwise cell type comparisons at the 0.001 permutation threshold (Figure 31) (99% agreement for 0.01 permutation threshold). This observation implies that regulatory variants are active across cell types in the same manner.

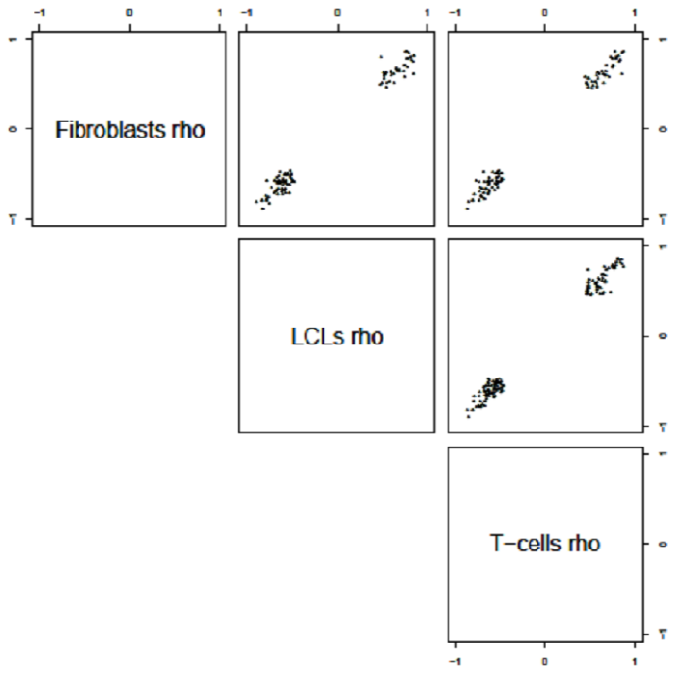


Figure 31. Comparison of the direction of the allelic effect of overlapping SNP-probe associations between pairs of cell types. The plots indicate the value of Spearman rho (effect size) for the same SNP-probe associations between cell types at the 0.001 permutation threshold. In all cases the direction of the allelic effect (indicated by the sign of rho) is the same.

To assess the strength of the cell type specificity observed, I performed RMA on cell types. Cell type specificity is expected to be reflected in the SNP x cell type interaction term, where any cell type-specific association is expected to have a significant interaction term. For cell type-specific eQTLs I found 61 % enrichment of low p-values in RMA (quantified by estimation of FDR (Storey and Tibshirani 2003) (Figure 32). No such enrichment was observed for cell type-shared eQTLs. RMA however is relatively limited in this type of analysis, as the power to detect an interaction term is never maximized. This is because reversal of allelic effect between cell types is not observed.

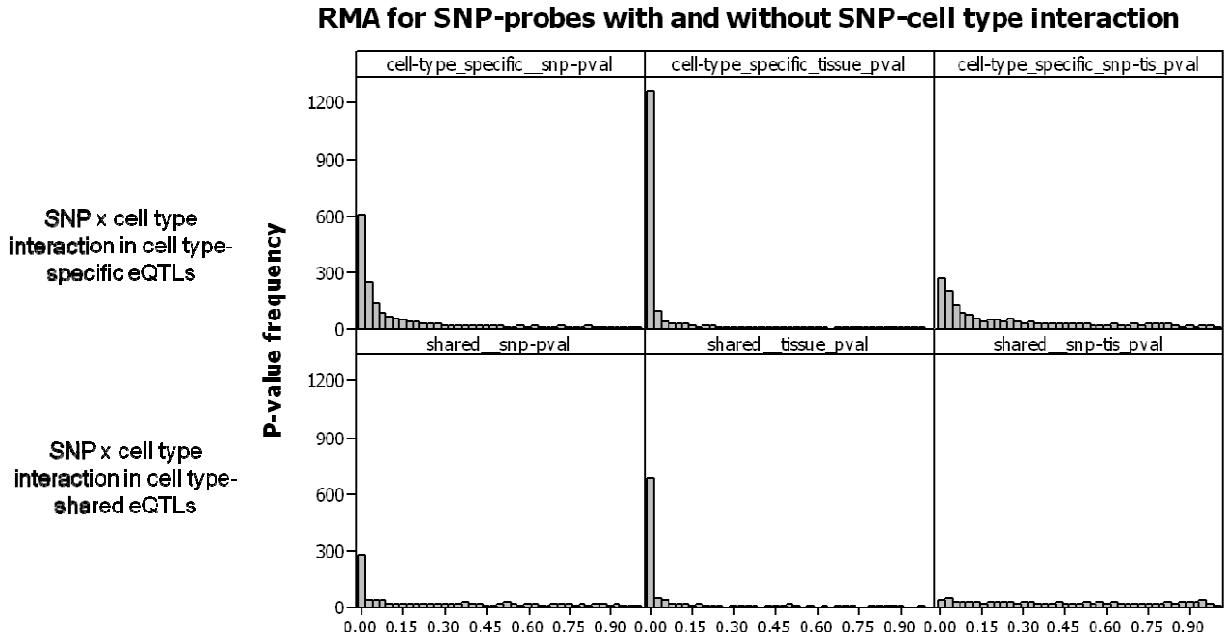


Figure 32. Repeated-measures ANOVA (RMA) to confirm eQTL cell type specificity. RMA association testing (using cell type as the repeated measure) of SNP-probe pairs significant in all three, exactly two and in only one cell type confirmed cell type specificity. Enriched low p-values were observed for SNP-cell type interactions corresponding to those associations that were defined as cell type-specific from the association overlap analysis.

ASE assays were used to validate a subset of cell type-specific eQTLs discovered for genes that also possess transcript SNPs. The ratio of the two transcript alleles was measured in individuals who were double heterozygotes for both the eQTL and the transcript SNP. For 35 transcript SNPs (seven in genes with fibroblast eQTLs, 14 in LCL eQTL genes and 14 in T-cell eQTL genes) extensive allelic imbalance was observed for the cell type in which the eQTL was detected (Figure 33). This imbalance was not observed for ratios of the same eQTL-transcript SNP pairs in the two cell types where the eQTL was not detected (paired t-test p-value = 5.6×10^{-7}). Taken together, these results confirm the signal of cell type specificity statistically and experimentally.

Limited sharing of associations between cell types may arise as a consequence of winner's curse (Goring, Terwilliger et al. 2001; Lohmueller, Pearce et al. 2003; Ioannidis 2008). A cross-threshold assay of sharing revealed that overlapping associations among cell type pairs increased slightly at relaxed significance thresholds for one cell type (Figure 34). Even with relaxed thresholds however over half of associations detected remain cell type-specific.

To further quantify the extent of winner's curse I selected significant SNP-probe pairs from one cell type, and explored their nominal (uncorrected) p-value distribution in the other two cell types. As expected, these distributions were enriched for low p-values, reflecting associations that are shared between cell types (Figure 35). When SNP-probe associations with significant associations in the secondary cell type were removed (i.e. shared associations at the same and at the lower significance threshold), the resulting nominal p-value distributions demonstrated only small enrichment for low p-values.

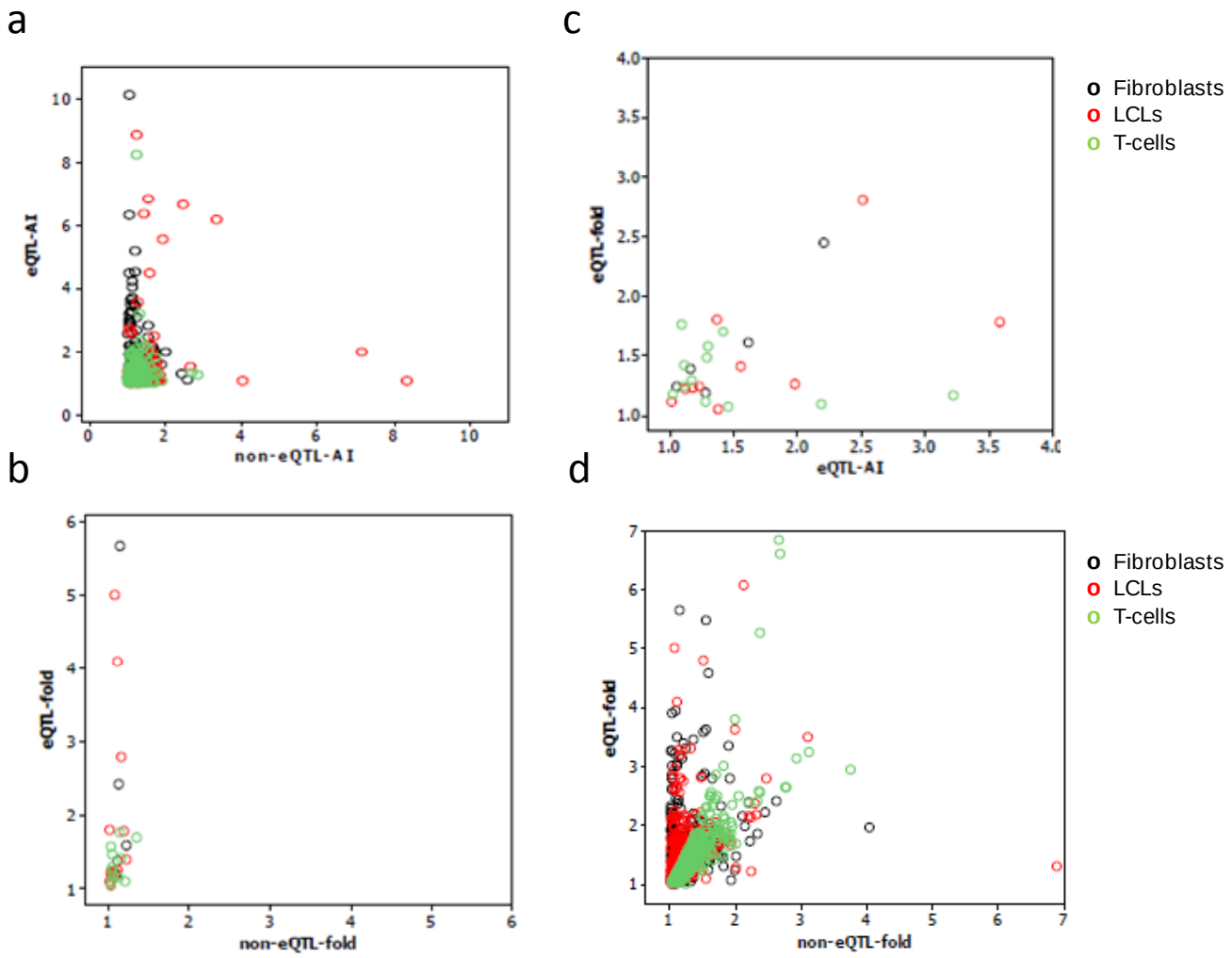


Figure 33. Allele-specific verification of eQTLs. a) Degree of allelic imbalance in double heterozygote individuals (for eQTL and transcript SNP) for 35 assayed transcript SNPs. The y axis shows the ratio of the two alleles in the cell type where the eQTL was initially discovered for each individual, and the x axis shows the mean of the ratio for the other two cell types for each individual. Data points are colour-coded to indicate cell type. The degree of allelic imbalance is more pronounced in the eQTL cell type vs. the non-eQTL cell types. **b)** Fold change difference in expression between the medians of the two homozygote classes of the population for the subset of 35 eQTLs that were confirmed by allele-specific expression (ASE). The plot shows fold change in the eQTLs cell type (y axis) and the non-eQTL cell types (x axis). As expected, the pattern is very similar to the one observed in a). **c)** The fold change estimated from the ratio of homozygotes (y axis) and allelic imbalance (x axis). The correlation is very strong and highly significant (Pearson's correlation coefficient r = 0.685, p-value < 0.0001). **d)** Fold change between the medians of the two homozygote classes of the population for the eQTL cell type (y axis) and the non-eQTL cell types (x axis). As expected the fold change is substantially higher for the eQTL cell type with a mean fold change of 1.55 and a range of 1.07 to 2.65.

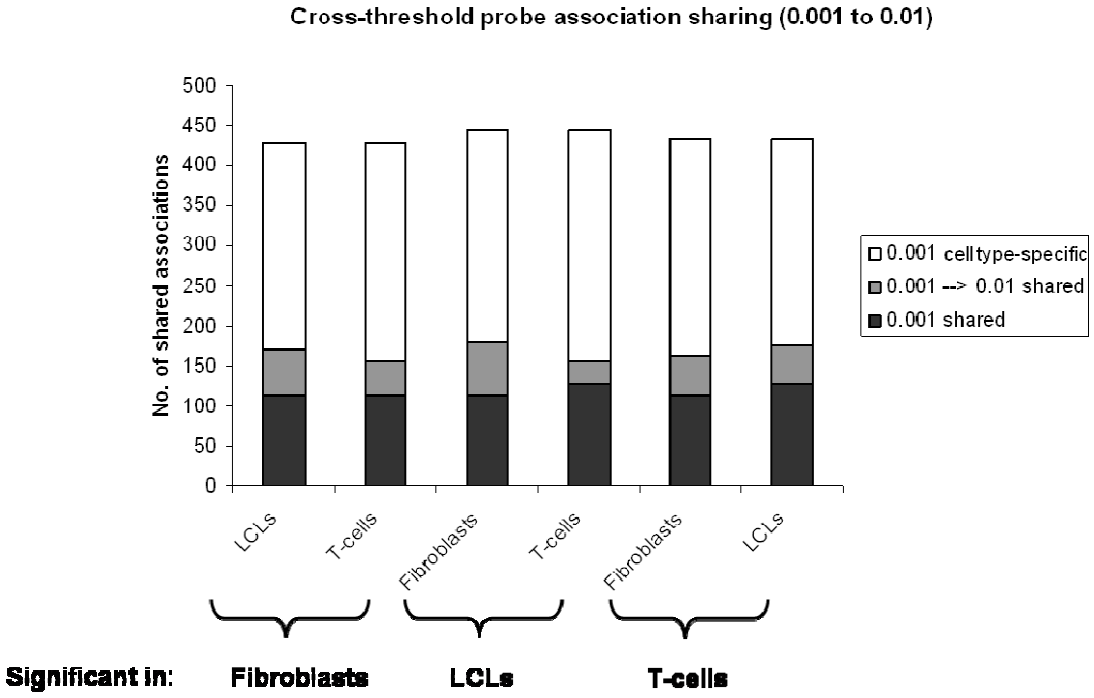


Figure 34. Cross-threshold probe association sharing (exploring the extent of winner's curse). I explored whether association sharing in cell type pairs increases when the significance threshold is relaxed for one cell type. Probe association sharing was found to increase from 28-35% to 40-50% when considering significant associations at the 0.001 permutation threshold in one cell type and the 0.01 permutation threshold in the replication cell type.

I thus quantified the fraction of significant cis eQTLs from one cell type that is not nominally significant (p-value prior to correction > 0.05) in either of the other two cell types. Using this principle of replication, it is estimated that 54%, 50% and 54% of cis eQTLs in fibroblasts, LCLs and T-cells respectively are cell type-specific, amounting to 69% of all cis eQTLs at the 0.001 permutation threshold. Consequently the limited overlap of cis eQTLs between cell types is unlikely to result from winner's curse and a substantial fraction of eQTLs is truly unique to each cell type.

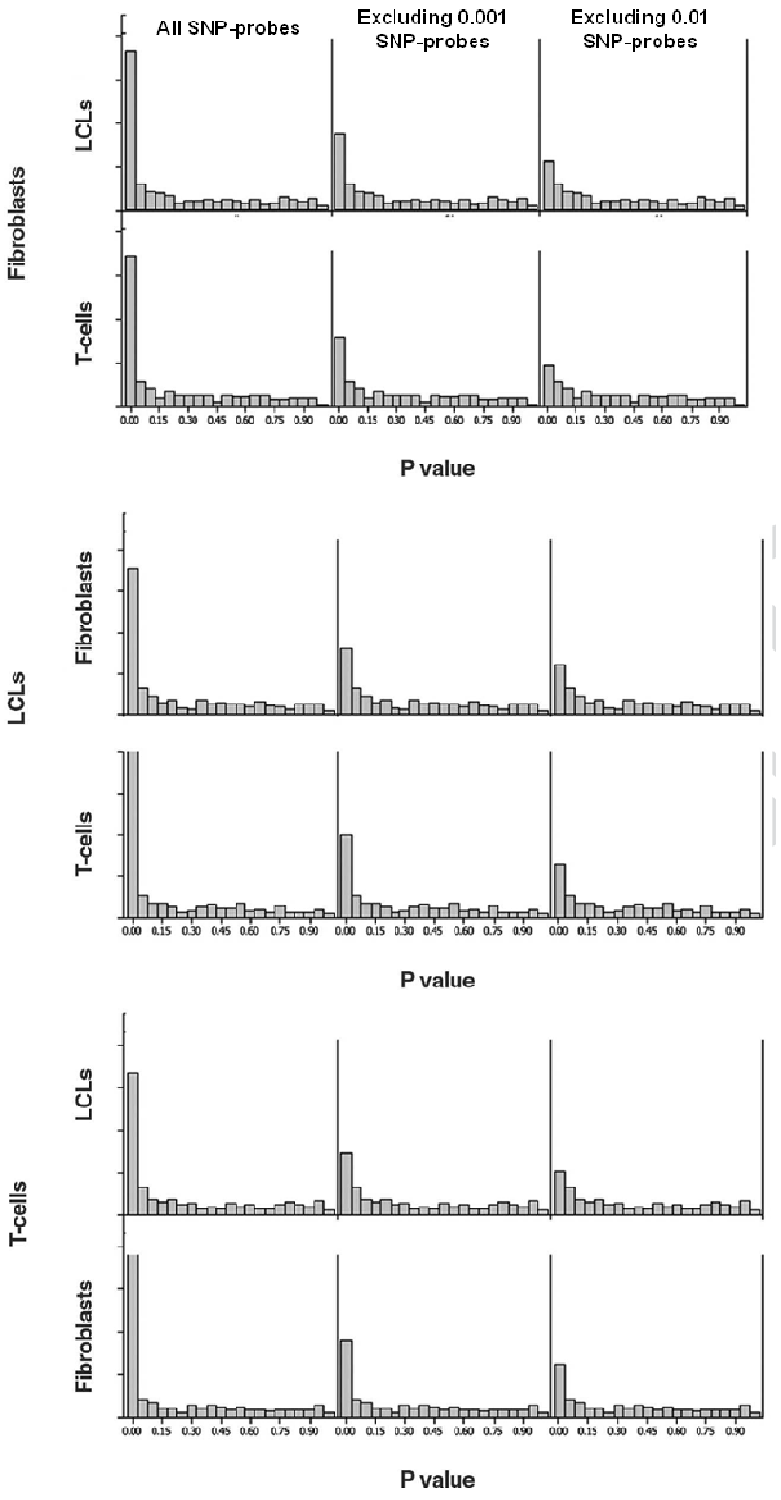


Figure 35. SNP-probe pair nominal (uncorrected) p-value distributions for the two secondary cell types conditional on the reference cell type eQTL (0.001 permutation threshold). The panels on the horizontal axis correspond to secondary cell type p-values for: i) all SNP-probes, ii) excluding SNP-probes significant at the 0.001 permutation threshold and iii) excluding SNP-probes significant at the 0.01 permutation threshold.

5.6 INDEPENDENT EQTLS

Experimental data are accumulating in an effort to annotate the regulatory landscape around genes (Birney, Stamatoyannopoulos et al. 2007). In agreement with previous studies (Stranger, Nica et al. 2007; Veyrieras, Kudaravalli et al. 2008) I found that, on average, the strength and density of cis associations detected for a given gene decay symmetrically with increasing distance from the gene's TSS (Figure 36). As discussed in Chapter 4, the correlated structure of variants within a genomic region due to LD enables association studies as it reduces the number of markers required for testing association with a phenotype, but can impede fine-mapping. The strategy described in 2.7.1 was used to identify independent eQTLs. eQTLs were mapped in recombination hotspot intervals, the most significant eQTL per interval was retained and the least significant eQTL from eQTL pairs with D' < 0.5 between intervals was excluded to derive independently-acting cis eQTLs. At the 0.001 permutation threshold and averaged across three cell types, 5.1% of genes with identifiable eQTLs possess more than one independent interval carrying a significant eQTL (Table 15). In LCLs this number is 4.5% which is comparable to 7.6% of genes with multiple independent eQTLs detected for the HapMap Phase 3 CEU population (Table 10).

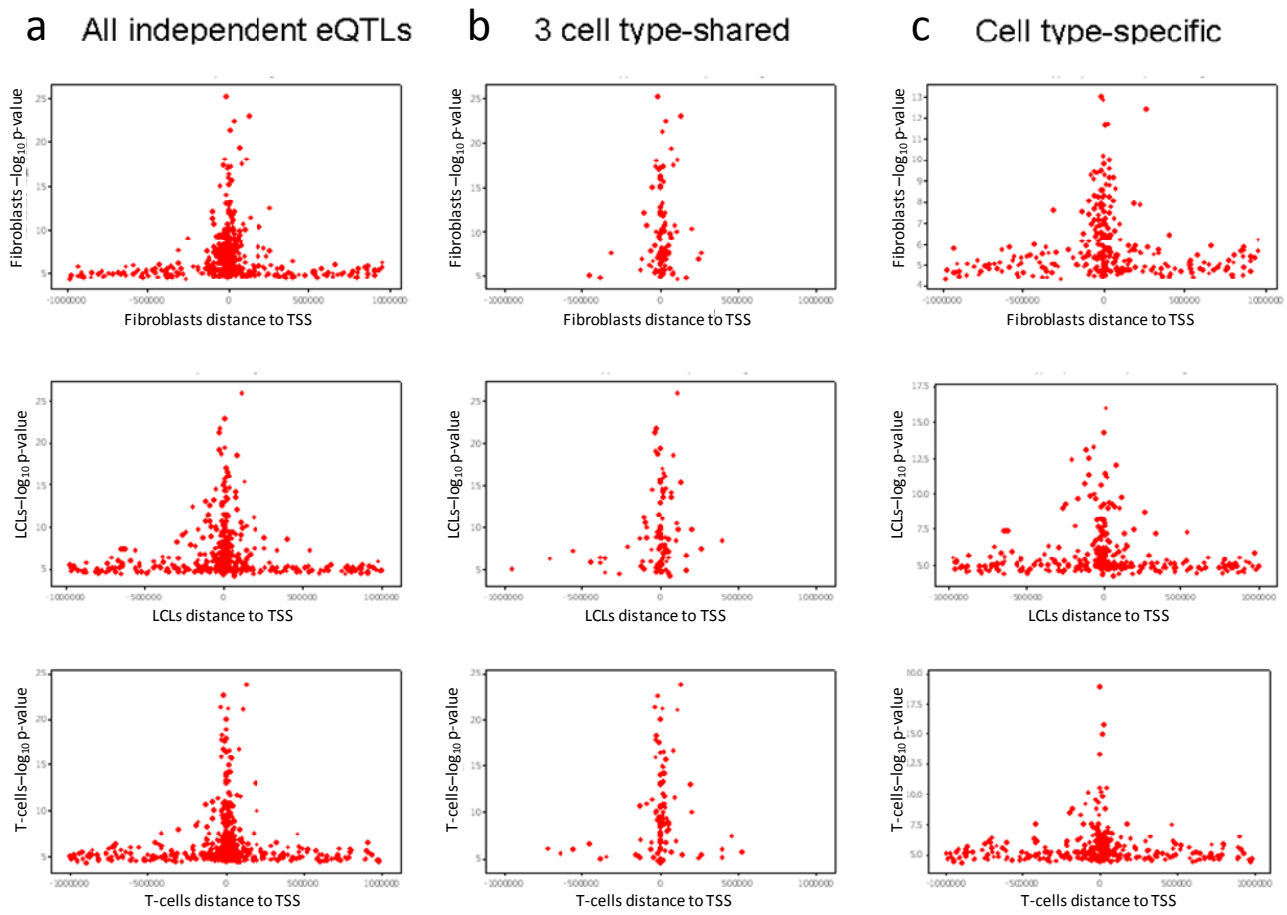


Figure 36. Localization of cis eQTLs (0.001 permutation threshold). a) Distance (in bases) to transcription start site (TSS) of all independent cis eQTLs in each cell type. **b)** Shared cis eQTLs in all three cell types. **c)** Cell type-specific cis eQTLs. Shared cis eQTLs cluster around the TSS whereas cell type-specific cis eQTLs span a wider range of distances from the TSS.

	0.01 permutation threshold			0.001 permutation threshold		
No. of Intervals	**Fibroblasts**	**LCLs**	**T-cells**	**Fibroblasts**	**LCLs**	**T-cells**
1	1,875	1,902	1,788	408	422	403
2	**237**	**217**	**230**	**18**	**16**	**27**
3	28	32	27	1	2	0
4	6	3	1	**0**	**1**	**0**
5	0	1	0	0	1	0
Total genes	2,146	2,155	2,046	427	442	430
% genes with ≥ 2 intervals	12.6	11.7	12.6	4.4	4.5	6.3

Table 15. Number of independent cis eQTLs (regulatory intervals) per gene.

This implies that for a fraction of genes and in all cell types considered, multiple cis regulatory variants act to determine expression levels. To further dissect the fine structure of regulatory variant sharing between genes, I repeated the overlap analysis but compared overlap of independent eQTLs (intervals rather than genes) across cell types. When the union of significant genes at the 0.001 permutation threshold was considered, only 6.9% of intervals were found to be shared across all three cell types. 9.7% were shared in exactly two cell types and 83.4% were cell type- specific (Figure 37 a and Table 16). The degree of interval sharing between cell types increases as genes that have shared expression associations in at least two (Figure 37 b and Table 17) and in all three cell types (Figure 37 c and Table 16) are considered.

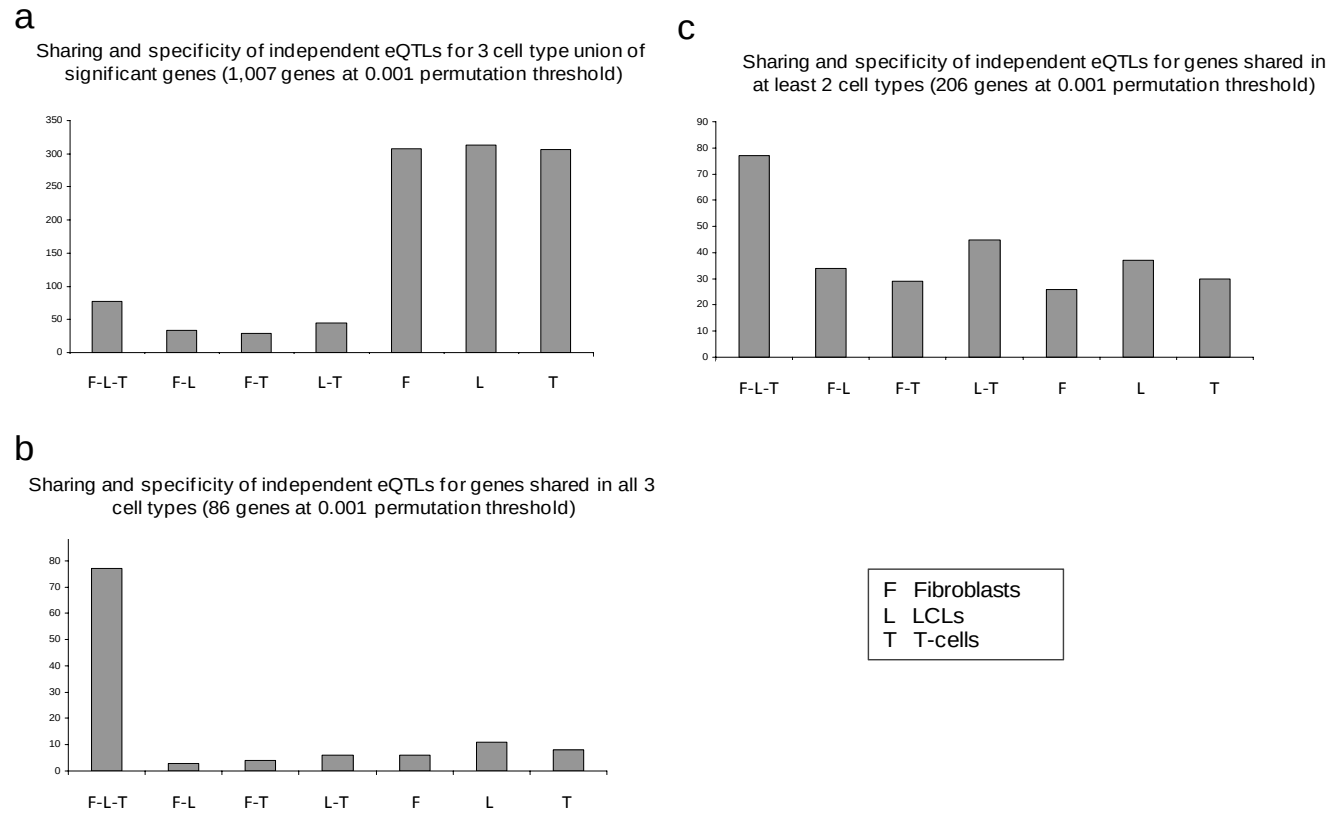


Figure 37. Fine-scale overlap of regulatory signals in three cell types (0.001 permutation threshold). Cell type-shared and specific independent intervals for **a)** the union of genes with a significant association, **b)** genes shared in at least two cell types and **c)** genes shared in all cell types.

In all cases however, there still remains a substantial fraction of cell type-specific independent eQTLs even for genes that had at least one cis eQTL in common in all three cell types.

		Independent eQTL sharing (0.001 permutation threshold)					
		3 cell type union significant genes	%	At least 2 cell type shared genes	%	3 cell type shared genes	%
Genes		1007		206		86	
3 cell type shared independent eQTLs	Fibroblasts - LCLs - T cells	77	6.9	77	27.7	77	67.0
Exactly 2 cell type shared independent eQTLs	Fibroblasts - LCLs	34	3.1	34	12.2	3	2.6
	Fibroblasts - T cells	29	2.6	29	10.4	4	3.5
	LCLs - T cells	45	4.1	45	16.2	6	5.2
cell type specific independent eQTLs	Fibroblasts	307	27.6	26	9.4	6	5.2
	LCLs	313	28.2	37	13.3	11	9.6
	T cells	306	27.5	30	10.8	8	7.0
Total		1111	100.0	278	100.0	115	100.0

Table 16. Independent eQTL (interval) sharing for significant genes (0.001 permutation threshold).

I further evaluated the distribution of independent eQTLs with respect to the TSS and their effect size, conditioning on sharing and specificity across cell types. Cell type-shared eQTLs tend to be of higher significance and larger effect size (Spearman rho) and cluster tightly around the TSS (Figure 36 for significance and Figure 38 for effect size). On the contrary, cell type-specific eQTLs tend to be of lower effect size and are more widely distributed around the TSS (Figure 36). This is in agreement with recent studies (Heintzman, Hon et al. 2009; Visel, Blow et al. 2009) showing that enhancer elements, which are found at greater distances from the gene, are more tissue-specific than basic regulatory elements such as promoters which map close to the TSS. Furthermore, the count of independent eQTLs per gene was significantly correlated with the number of transcripts per gene, for genes with significant cis eQTLs (Pearson's correlation coefficient = 0.049, p-value = 0.117 for the 0.001 permutation threshold, and Pearson's correlation coefficient = 0.105, p-value < 0.0001 for the 0.01 permutation

threshold). This suggests that regulatory complexity is correlated with transcript complexity raising the possibility that some of the regulatory variant signals may mediate genotype-specific choices for alternative TSSs or alternative splicing.

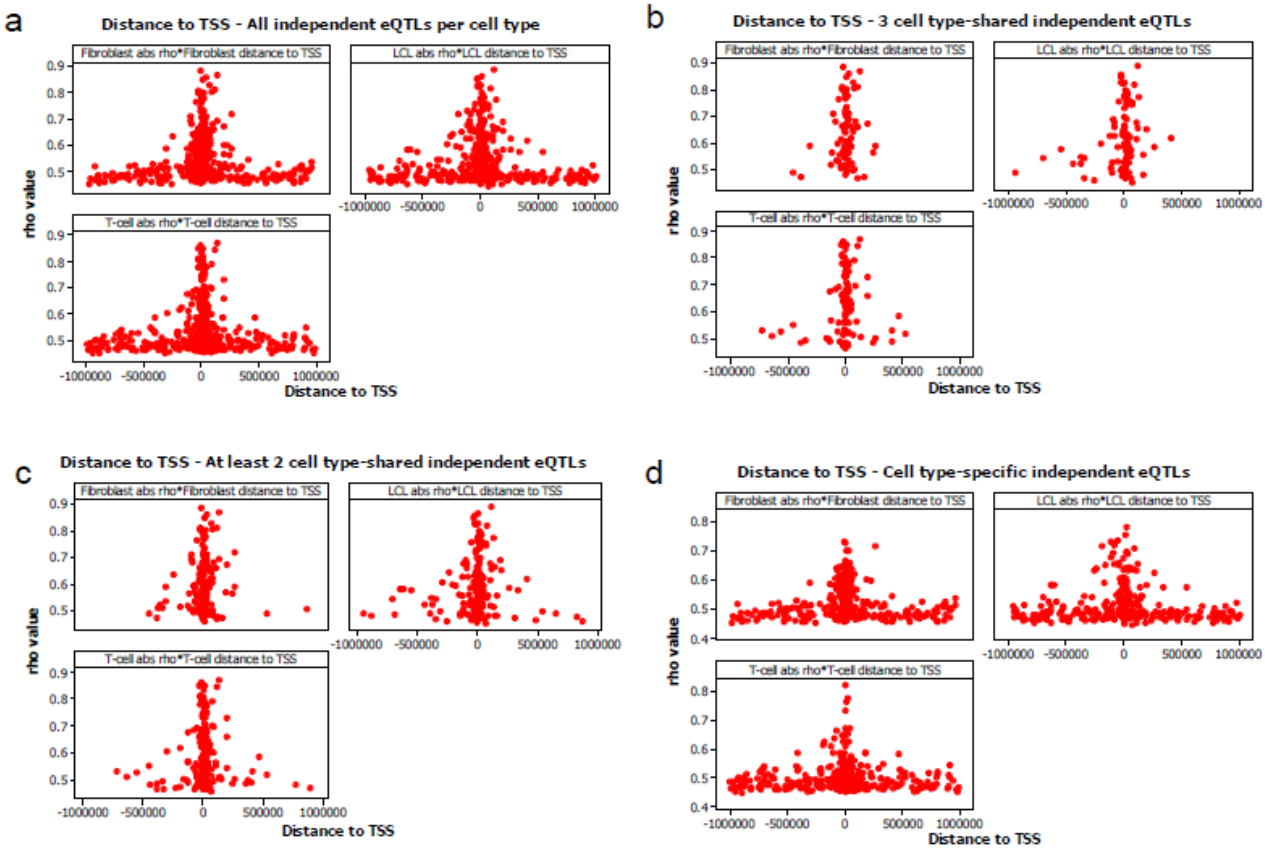


Figure 38. Effect size (Spearman rho) of independent cis eQTLs (0.001 permutation threshold) as a function of the distance (in bases) to a gene's transcription start site (TSS). a) shows all independent cis eQTLs discovered in each cell type, **b)** and **c)** show three cell type-shared and two cell type-shared independent cis eQTLs respectively and **d)** shows independent cis eQTLs specific to one cell type only.

The complexity of the regulatory landscape is illustrated in the case of TSPO, an outer mitochondrial membrane protein of peripheral tissues (Papadopoulos, Baraldi et al. 2006) with a role in cholesterol transport, immunomodulation and apoptosis (Casellas, Galiegue et al. 2002) (Figure 39). At the 0.01 permutation threshold six

independent intervals were identified for this gene: one shared in all three cell types, three fibroblast-specific and two LCL-specific intervals. Additionally, four alternate transcriptional splice variants, encoding different isoforms, have been characterized for this gene and TSPO receptors are found in many tissues of the human body.

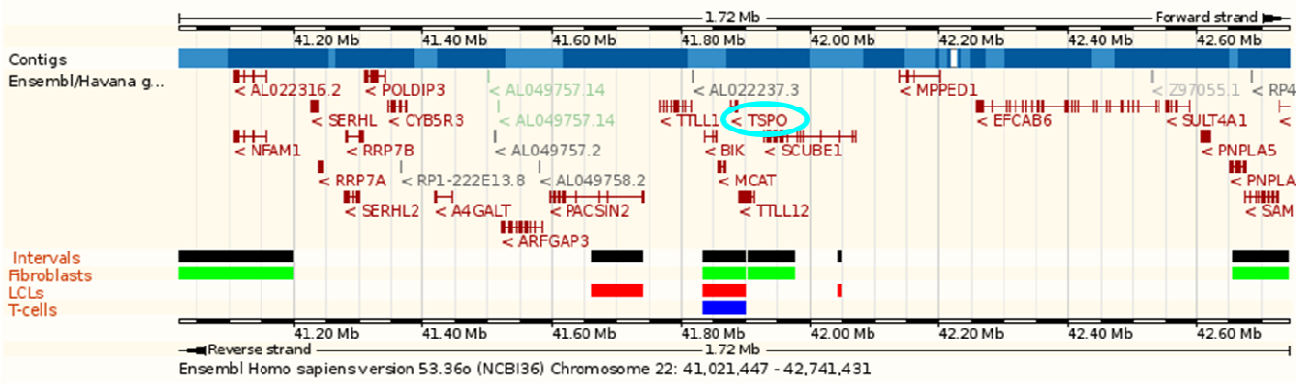


Figure 39. Complex genetic architecture around the *TSPO* gene. *TSPO* (blue oval) encodes for an outer mitochondrial membrane protein with a role in cholesterol transport, immunomodulation and apoptosis. Six independent intervals have been identified for this gene in three cell types: one shared in all three cell types, and 5 cell type-specific. Figure created using the Ensembl genome browser (http://www.ensembl.org).

Regulatory complexity also takes the form of a single independent interval regulating the expression of multiple genes (interval pleiotropy). I explored the number of associated genes per independent interval and found that at the 0.001 permutation threshold over 6% of intervals are associated with the expression of more than one gene (this number increases to almost 19% at the 0.01 permutation threshold). An example of a single eQTL influencing eight genes is shown in Figure 40. In such cases it may be interesting to explore whether the genes influenced by a common regulatory interval are components of the same pathway or network. The multidimensionality caused by cell type specificity, regulatory region promiscuity and genetic variation highlight the challenges to be faced when a wider range of conditions and context-dependent effects (cell types, tissues, developmental stages) are considered.

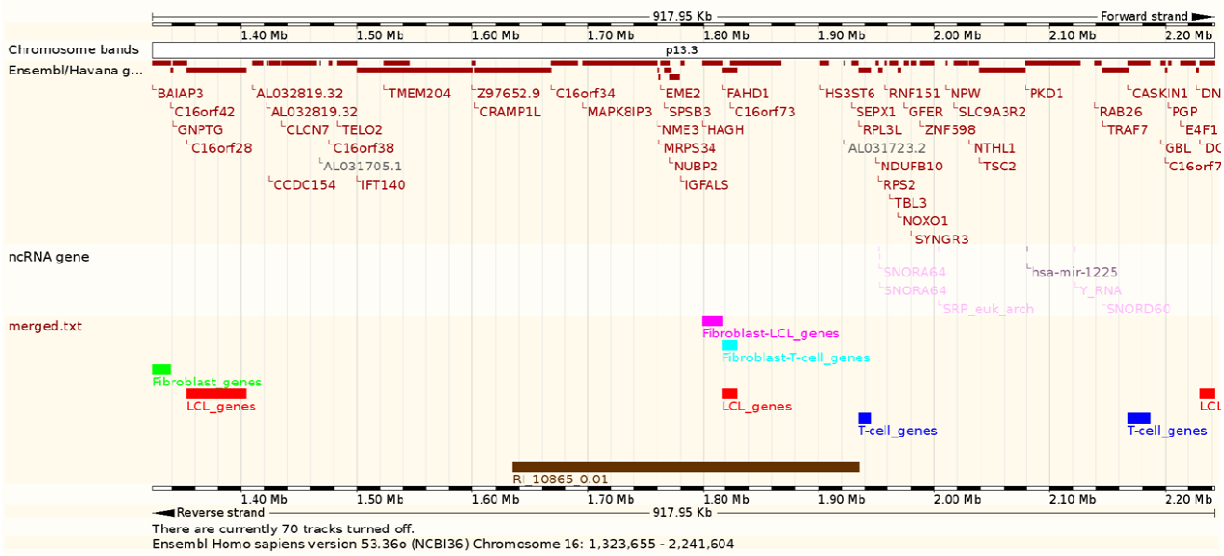
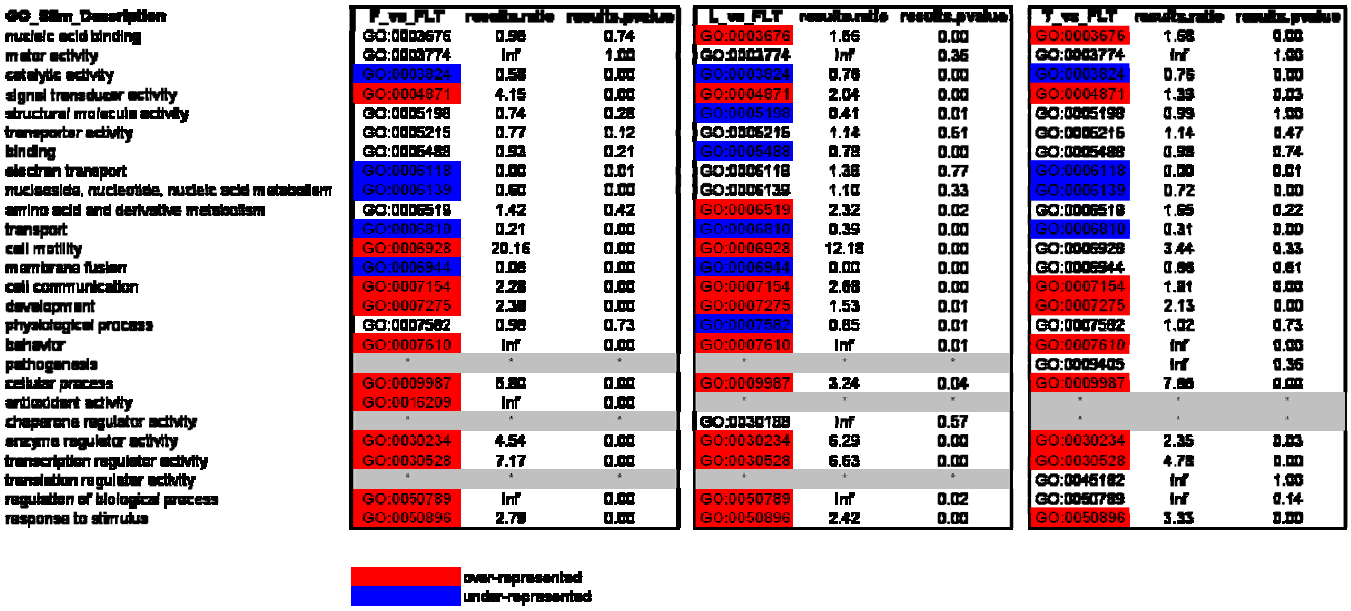


Figure 40. A single independent eQTL in the brown interval (denoted at RI_10865) affects the expression of a total of 8 genes (0.01 permutation threshold). RI_10865 affects the expression of three LCL-specific, two T-cell specific and one fibroblast-specific genes, as well as a gene with a significant association shared in fibroblasts and LCLs and a gene with a significant association shared in fibroblasts and T-cells (find genes). Figure created using the Ensembl genome browser.

5.7 BIOLOGICAL PROPERTIES OF SHARED AND CELL TYPE-SPECIFIC EQTLS

Gene Ontology (GO) (Ashburner, Ball et al. 2000) terms were used to compare biological properties of cell type-specific and shared genes. For cell type-specific associations, I detected an over-representation of properties linked to signal transducer activity, cell communication, development, behaviour, cellular process, enzyme regulator activity, transcription regulator activity and response to stimulus, reflecting properties likely to sculpt cell type-specific profiles. For associations shared in all cell types I found an over-representation of catalytic activity and transport properties (Fisher's exact test p-value < 0.05) (Table 17).



GO_Slim_Description	F_vs_FLT	results.ratio	results.pvalue	L_vs_FLT	results.ratio	results.pvalue	T_vs_FLT	results.ratio	results.pvalue
nucleic acid binding	GO:0003676	0.98	0.74	GO:0003676	1.66	0.00	GO:0003676	1.68	0.00
motor activity	GO:0003774	Inf	1.00	GO:0003774	Inf	0.35	GO:0003774	Inf	1.00
catalytic activity	GO:0003824	0.58	0.00	GO:0003824	0.78	0.00	GO:0003824	0.76	0.00
signal transducer activity	GO:0004871	4.15	0.00	GO:0004871	2.04	0.00	GO:0004871	1.39	0.03
structural molecule activity	GO:0005198	0.74	0.28	GO:0005198	0.41	0.01	GO:0005198	0.99	1.00
transporter activity	GO:0005215	0.77	0.12	GO:0005215	1.14	0.51	GO:0005215	1.14	0.47
binding	GO:0005488	0.93	0.21	GO:0005488	0.78	0.00	GO:0005488	0.98	0.74
electron transport	GO:0006118	0.00	0.01	GO:0006118	1.38	0.77	GO:0006118	0.00	0.01
nucleoside, nucleotide, nucleic acid metabolism	GO:0006139	0.60	0.00	GO:0006139	1.10	0.33	GO:0006139	0.72	0.00
amino acid and derivative metabolism	GO:0006519	1.42	0.42	GO:0006519	2.32	0.02	GO:0006518	1.65	0.22
transport	GO:0006810	0.21	0.00	GO:0006810	0.39	0.00	GO:0006810	0.31	0.00
cell motility	GO:0006928	20.16	0.00	GO:0006928	12.18	0.00	GO:0006928	3.44	0.33
membrane fusion	GO:0006944	0.06	0.00	GO:0006944	0.00	0.00	GO:0006944	0.66	0.61
cell communication	GO:0007154	2.28	0.00	GO:0007154	2.68	0.00	GO:0007154	1.81	0.00
development	GO:0007275	2.38	0.00	GO:0007275	1.53	0.01	GO:0007275	2.13	0.00
physiological process	GO:0007582	0.98	0.73	GO:0007582	0.85	0.01	GO:0007582	1.02	0.73
behavior	GO:0007610	Inf	0.00	GO:0007610	Inf	0.01	GO:0007610	Inf	0.00
pathogenesis	*	*	*	*	*	*	GO:0009405	Inf	0.36
cellular process	GO:0009987	5.80	0.00	GO:0009987	3.24	0.04	GO:0009987	7.86	0.00
antioxidant activity	GO:0016209	Inf	0.00	*	*	*	*	*	*
chaperone regulator activity	*	*	*	GO:0030188	Inf	0.57	*	*	*
enzyme regulator activity	GO:0030234	4.54	0.00	GO:0030234	6.29	0.00	GO:0030234	2.35	0.03
transcription regulator activity	GO:0030528	7.17	0.00	GO:0030528	6.63	0.00	GO:0030528	4.78	0.00
translation regulator activity	*	*	*	*	*	*	GO:0045182	Inf	1.00
regulation of biological process	GO:0050789	Inf	0.00	GO:0050789	Inf	0.02	GO:0050789	Inf	0.14
response to stimulus	GO:0050896	2.78	0.00	GO:0050896	2.42	0.00	GO:0050896	3.33	0.00

over-represented
under-represented

Table 17. GO Slim term comparison for cell type-specific vs. three cell type-shared genes (0.001 permutation threshold). Fisher's exact test significant p-value < 0.05. Biological properties over-represented in cell type-specific genes are shown in red and include signal transducer activity, cell communication, development, behaviour, cellular process, enzyme regulator activity, transcription regulator activity, and response to stimulus. Biological properties under-represented in cell type-specific genes are shown in blue and include catalytic activity and transport.

Entropy of expression for each gene was calculated as an indication of cell type specificity, with lower entropy values reflecting higher specificity. I used data from cell types (tissues) included the GNF/Novartis atlas of gene expression (Su, Wiltshire et al. 2004) (Table 18) and compared entropy between genes with shared vs. cell type-specific cis eQTLs. Genes with fibroblast-specific eQTLs showed consistently and significantly lower entropy values (i.e. were more cell type-specific) compared to shared associations (M-W p-value = 0.0047). This signal was in the same direction, but less prominent, for the other two cell types. This may be due to the fact that fibroblasts are biologically more distant to LCLs or T-cells, or to potential tissue sampling biases in the GNF/Novartis collection.

tissue	description
adipocyte	fat
adrenal cortex	perimeter of the adrenal gland
adrenal gland	endocrine glands on kidneys
amygdala	groups of neurons located within medial temporal lobes of brain
appendix	part of digestive system, blind-ended tube connected to cecum
bone marrow	tissue in the hollow interior of bones, produces new blood cells
bronchial epithelial	lung epithelium
caudate nucleus	nucleus located in basal ganglia of brain, role in learning and memory
cerebellum peduncles	region of brain, role in the integration of sensory perception
ciliary ganglion	parasympathetic ganglion located in the posterior orbit
dorsal root ganglion	nodule on dorsal root (afferent sensory root of spinal nerve)
heart	heart
hypothalamus	small nuclei in brain linking nervous to endocrine system, located above brain stem
kidney	kidney
liver	liver
lung	lung
lymph node	organ consisting of multiple cell types, part of the lymphatic system
ovary	ovary
pancreas	pancreas
pituitary	pituitary
prostate	prostate
salivary gland	salivary gland
skeletal muscle	skeletal muscle
skin	skin
smooth muscle	smooth muscle
spinal cord	spinal cord
superior cervical ganglion	largest of the cervical ganglia, supplies sympathetic innervation to the face
testis	testis
thymus	thymus
thyroid	thyroid
tongue	tongue
tonsil	tonsil
trachea	trachea
trigeminal ganglion	sensory ganglion of the trigeminal nerve (5th cranial nerve)
uterus	uterus
uterus corpus	endometrium
whole blood	whole blood
whole brain	whole brain

Table 18. Tissues used for entropy calculation (GNF/Novartis atlas of gene expression).

5.8 CONCLUSIONS

This study provides a direct comparison of the impact of regulatory variants in a cell type-dependent context. Having controlled for all other confounders such as experimental design, sampling variance and differences in technology, I have demonstrated that variants affecting gene regulation largely act in a cell type-specific manner, and even cell types as closely related as LCLs and T-cells share only a minority of their cis eQTLs. Based on the three cell types tested, it is estimated that 69-80% of regulatory variants are cell type-specific. Regulatory variant complexity correlates with

transcript complexity suggesting genotype-specific effects on alternative transcript choice. In addition, cell type-specific eQTLs are of smaller effect size and tend to localize at greater distances from the TSS recapitulating enhancer element distributions. Importantly, the signal of cell type specificity is primarily due to differential use of regulatory elements of genes that are expressed in almost all cell types. This analysis is also the first to demonstrate robust replication of eQTLs in LCLs between samples collected and transformed decades apart. This is of great importance for the field of human genetics since a large number of cohorts have collections of LCLs whose value has been debated and questioned repeatedly. I argue that LCLs are likely to represent a legitimate biological system that can be used for disease interpretation or other functional studies with all the limitations of cell line specificity. As more tissues are interrogated diminishing returns in discovery of eQTLs are expected, and it is possible that there is a minimum set of tissues that will be informative for the vast majority of regulatory variants. Nevertheless, this study highlights the need for deep and wide interrogation of regulatory variation in multiple cell types and tissues in order to elucidate their differential functional properties. The pattern of cell type specificity is not expected to be limited to regulatory variants, but is likely to apply to protein-coding and other putative functional variants (e.g. epigenetic modifications).

6 DISCUSSION

6.1 GENETIC VARIATION IN GENE REGULATION

Regulation of gene expression is one of the most important cellular functions. It defines and maintains cell types, shapes higher level phenotypes in health and disease, and it is likely that a large proportion of the genetic signal associated with phenotypic variation is harboured in regulatory sequences. At the cellular level, the effects of genetic variants can be easily interpretable, but at the whole organism level these signals may be more challenging to dissect due to the large number of direct and indirect interactions occurring between DNA and phenotype (Dermitzakis 2008). Using gene expression as an intermediate step to connect DNA variation and higher level phenotypes is an important way forward. In this thesis I have explored three aspects of the impact of genetic variation on gene expression: a) effect of interactions between genetic variants on gene expression in cis and trans, b) fine-scale architecture of the cis regulatory landscape, and c) cell type specificity of regulatory variation. The following sections summarise the findings of these three studies.

6.1.1 Genetic interactions with an impact on gene expression

Although epistasis is difficult to test for, its contribution to gene regulation is emerging. I have presented a framework to test for interactions between two common types of variants in the genome: regulatory variants (eQTLs) and protein-coding variants (nsSNPs). Two distinct concepts, the level of gene expression and allelic variation of protein sequence were jointly considered and were shown to be important for downstream regulation of genes. In cis the functional impact of protein-coding variants was shown to be modulated (magnified or masked) through the action of regulatory variants nearby. Depending on the phasing of eQTL and nsSNP alleles, cis modification

can result in the production of different ratios of distinct isoforms. If the modulated protein products have downstream targets, the interaction in cis may result in true epistasis by affecting expression levels of target genes in a trans manner. Using this framework of hypothesis-driven analysis of epistasis, I have demonstrated that genetic interactions between these, and possibly other types of functional variants, contributes to shaping phenotypic variation. Detecting epistasis is crucial as it uncovers new loci affecting phenotypes. Its effects can mask the genetic impact of variants and impede replication of associations. Differential fixation of variants modulating the primary disease effect can determine the degree of penetrance of disease alleles and concequently considering interactions is a necessary next step for GWAS.

6.1.2 Fine-scale architecture of the cis regulatory landscape

Studies interrogating regulatory variation identify genomic regions likely to harbour genetic elements that control gene expression. For over half of all genes in the genome, multiple SNPs with an expression association are identified, most of which tag the same regulatory element. Identifying eQTLs is an important step in understanding gene regulation, but it is necessary to move from identification of large segments of DNA to the fine-mapping of regulatory elements. I have presented a strategy that can be employed to narrow down regions of interest and to identify markers that tag regulatory elements with an independent effect on gene expression. After controlling for the correlated structure of variants, cis eQTLs tagging independent regulatory elements (regulatory intervals) were identified. Roughly seven percent of genes possess multiple independent regulatory intervals influencing cis expression levels and in agreement with recent studies, the strength and abundance of these were greater around the TSS. When exploring independent eQTLs across populations, 35% were shared in at least two of the eight populations studied, hinting at common regulatory control for a fraction of genes. Adding to the complexity of the cis regulatory landscape,

it was shown that interactions between genetic variants in cis also influence expression levels. This study is a first step toward the dissection of cis regulatory architecture on a genome-wide scale. This type of complexity should be anticipated and considered in future studies addressing gene regulation.

6.1.3 Cell type specificity of regulatory variation

The extent to which genetic variation manifests itself as tissue-specific expression patterns and the value of exploring eQTLs across a range of cell types are only starting to emerge. In this study I have highlighted the importance of investigating multiple cell types by exploring the specificity of cis eQTLs in three cell types (fibroblasts, LCLs and T-cells). Regulatory variation was shown to control gene activity predominantly depending on cell type, with 69-80% of regulatory variants operating in a cell type-specific manner. It was found that even the same genes are largely controlled through the action of different regulatory elements depending on the cell type. Cataloguing cell type-specific regulatory variation will help connect biological pathways controlling cellular activities in health and disease (Emilsson, Thorleifsson et al. 2008; Schadt, Molony et al. 2008; Wu, Delano et al. 2008), although it is not yet clear how straightforward it will be to determine the relevant cell type for a particular disease, or how many cell types will be necessary to compile a comprehensive catalogue of regulatory variation.

6.2 Overlap of GenCord eQTLs with disease and complex trait SNPs

Integrating the results of eQTL and genome-wide disease and trait association studies can provide important clues for mechanisms that give rise to phenotypes, can point to genes that have not been reported in primary disease association studies or can strengthen and complement the role of identified candidate genes. At the time of writing, scanning the Catalog of Genome-Wide Association Studies

(http://www.genome.gov/gwastudies) revealed a number of SNPs with a significant association to a disease or trait that were also GenCord eQTLs. In cases where a disease or complex trait SNP is also an eQTL, it is plausible that the GWAS phenotype arises to a certain extent as a consequence of regulatory effects. Here I discuss a number of examples of overlapping GWAS SNPs and eQTLs to demonstrate the usefulness of integrating information from these sources.

6.2.1 Crohn disease

A strong association to CD in European-derived populations was detected for rs744166 (Barrett, Hansoul et al. 2008). This SNP maps in an intron of *STAT3*, a gene with a role in signalling pathways implicated in CD pathogenesis. The identical SNP showed a significant association (0.01 permutation threshold) with expression levels of genes *STAT5A* and *STAT5B* in fibroblasts and T-cells respectively (Figure 41).

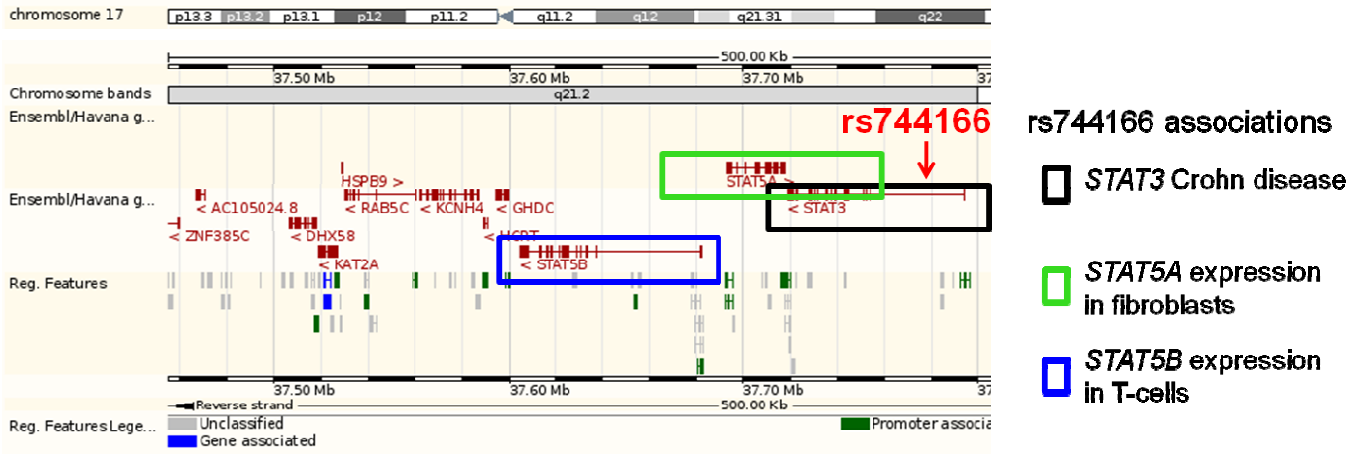


Figure 41. Overlap between disease-associated SNPs and eQTLs in the case of Crohn disease (CD). rs744166 is associated with CD and is an eQTL in fibroblasts and T-cells. *STAT3* (black box) is the GWAS-reported gene, *STAT5A* and *STAT5B* are the expression-associated genes in fibroblasts (green box) and T-cells (blue box) respectively. All genes are components of the JAK-STAT pathway which is likely to have a role in immune system-related diseases. Figure created using the Ensembl genome browser.

Both expression-associated genes, as well as the GWAS-reported gene belong to the STAT family of transcriptional activators. STAT3 has anti-apoptotic as well as proliferative effects and its constitutive activation is associated with various human cancers. It is required for self-renewal of embryonic stem cells (Takeda, Noguchi et al. 1997), is essential for differentiation of TH17 helper T-cells (Yang, Panopoulos et al. 2007) and has been implicated in a variety of autoimmune diseases including CD. STAT5A mediates the responses of cell ligands and growth hormones, has a role in tumourigenesis in myeloma and lymphoma and its mouse counterpart suggests an antiapoptotic function. STAT5B mediates signal transduction triggered by cell ligands and growth hormones and has a role in diverse biological processes including T-cell receptor signalling, apoptosis, adult mammary gland development and sexual dimorphism of liver gene expression.

STAT3, STAT5A and STAT5B interact with a number of proteins, some unique to each STAT family member, but all interact with JAK1 (Figure 42), a component of the JAK-STAT signalling pathway. This pathway is involved in regulation of cellular responses to cytokines and growth factors through signal transduction to the nucleus, where activated STAT proteins modify gene expression. It plays a central role in principal cell fate decisions, in cell proliferation, differentiation and apoptosis and is particularly important in hematopoiesis. Dysregulation of JAK–STAT signalling is associated with immune disorders including CD (Shuai and Liu 2003). In their 2003 review, Shuai and Lui stated (pg 908): "The aetiopathology of Crohn's disease is poorly understood. Mice with tissue-specific disruption of *Stat3* during haematopoiesis show Crohn's disease-like pathogenesis. In addition, constitutively tyrosine phosphorylated STAT3 is found in intestinal T-cells from patients with Crohn's disease. These results indicate that the dysregulation of STAT3 signalling might be involved in the pathogenesis of Crohn's disease. However, the exact role of STAT3 in the pathogenesis

of Crohn's disease is not understood." It may be the case that pathogenesis arises in part as a consequence of quantitative perturbations of different components of interacting proteins of the JAK-STAT pathway. Intriguingly, at the time of writing, preliminary evidence suggested that rs744166 also shows a strong association to MS (GWAS results from a European population, V.L. personal communication). Similarly to CD, MS is also an autoimmune disease, but in this case pathogenesis involves an immune response triggered by T-cells and directed at axon myelin.

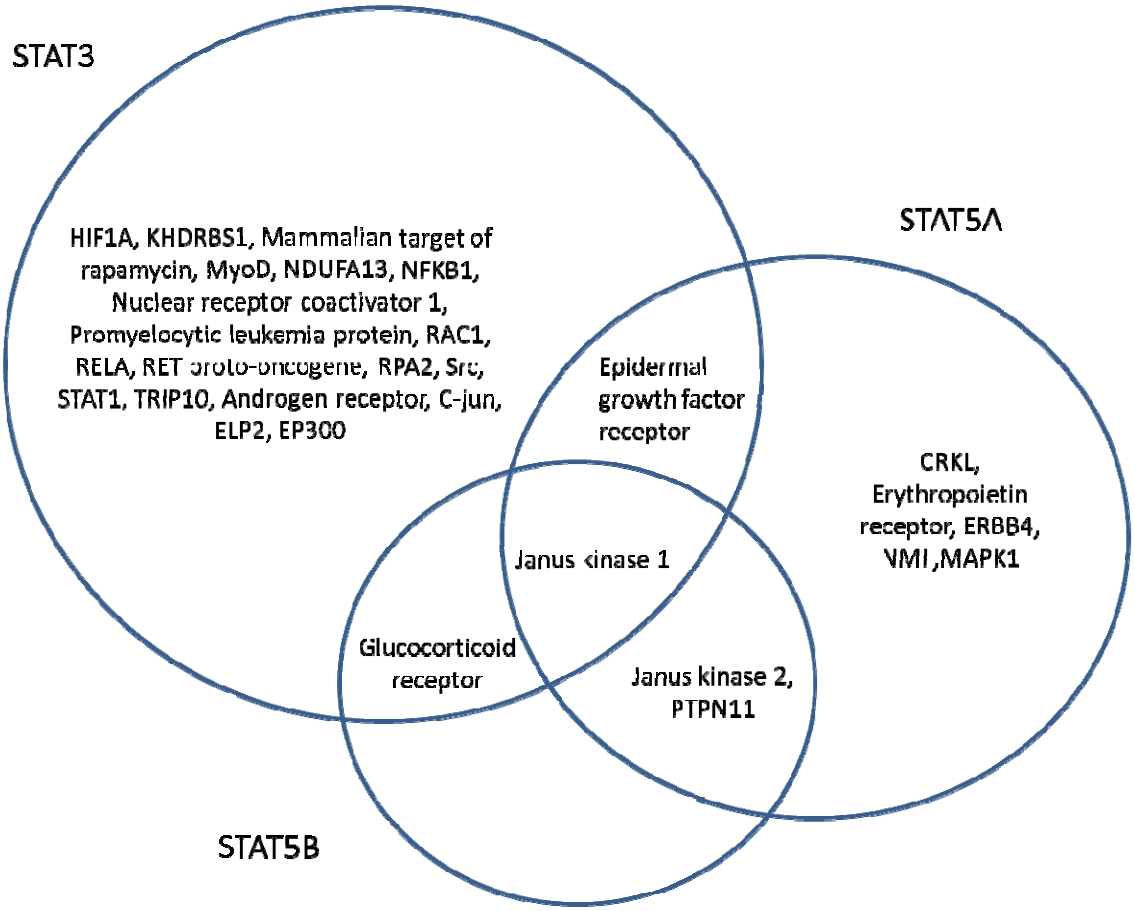


Figure 42. STAT3, STAT5A and STAT5B interacting proteins. All gene products interact with Janus kinase 1, a component of the JAK-STAT pathway. STAT3 and STAT5A interact with Epidermal growth factor receptor. STAT3 and STA5B interact with Glucocorticoid receptor. STAT5A and STAT5B interact with Janus kinase 2 and PTPN11. These interactions highligł complexity underlying disease pathogenesis and suggest that clues from different typ studies can help piece together pathogenesis mechanisms. STAT3 was identified as a cand gene from a disease GWAS. STAT5A and STAT5B were identified as genes with eQTLs.