Regulatory variation and its role in disease

Alexandra Cristina Nica

Hughes Hall College
University of Cambridge

August 2010

This dissertation is submitted for the degree of Doctor of Philosophy

To my parents

Mihaela and Marian

Declaration

This thesis describes my work undertaken at the Wellcome Trust Sanger Institute under the supervision of Prof. Emmanouil Dermitzakis and Dr. Inês Barroso, in fulfilment of the requirements for the degree of Doctor of Philosophy at University of Cambridge. This dissertation is the result of my own work and includes nothing that 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 word limit specified by the Biology Degree Committee.

Alexandra Cristina Nica
Cambridge, August 2010

Abstract

The role of regulatory variation in shaping phenotypes became apparent once significant species differences could not be explained by differences at DNA sequence level. Since then, the control of gene expression emerged as an essential process at the heart of cell-type differentiation and determination of phenotypic variance across multiple populations and tissues. Concurrent with the identification of genetic variants affecting transcript levels (eQTLs) across the human genome, large-scale genome-wide association studies (GWAS) shed light into the genetics of complex traits by detecting a multitude of susceptibility loci of modest effect-size. The goal of this thesis is to explore the role of regulatory variation in explaining genetic associations with complex traits and assess how that role differs across tissues.

To address this aim, I first developed an empirical methodology called Regulatory Trait Concordance (RTC) that integrates eQTLs and GWAS results in order to reveal the subset of association signals due to proximal eQTLs (*cis* variants). By simulating different genomic regions, I show that this method outperforms simple correlation metrics between single nucleotide polymorphisms (SNPs). I observe a significant enrichment of regulatory effects among currently known GWAS loci and I apply the RTC method to prioritize relevant genes for each of the tested complex traits. For this purpose, I use gene expression data measured in lymphoblastoid cell lines (LCLs) derived from HapMap 3 individuals and I detect several potential disease-causing regulatory effects, with a strong enrichment for immunity-related conditions. Furthermore, I present an extension of the method in *trans*, where interrogating the whole genome for downstream effects of the disease variant can be informative regarding its unknown primary biological effect.

Given that certain phenotypes manifest themselves only in certain tissues, I next explore the complexity of regulatory tissue-specificity in three human cell-types: LCLs, skin and fat. I discover an abundance of eQTLs in each of the three tissues derived from a sample set of well-phenotyped female twins and I make use of the unique study design (matched co-twins) to validate the discoveries. I highlight the challenges of comparing eQTLs between tissues and propose that continuous significance estimates and direct comparison of the magnitude of effect on the fold

change in expression are essential properties providing a biologically realistic view of tissue-specificity. Under this framework, I find evidence for extensive tissue-specificity: 30% of eQTLs are shared among the three tested tissues and of those, 10-20% have significant differences in the magnitude of fold change between homozygote genotypic classes across tissues.
Finally, I show that finding causal regulatory effects for complex disease associations is highly impacted by the tissue where expression is quantified and its relevance to the trait. I apply the RTC method on GenCord, a dataset where gene expression had been previously measured in LCLs, fibroblasts and primary T-cells derived from the same 75 individuals of Europeans descent. As expected, I find a large proportion of likely causal regulatory effects for GWAS signals to be tissue dependent (70% of all significant signals).

Altogether, my results support the informative value of gene expression in explaining a subset of GWAS signals and highlight the need to explore a variety of cell-types for enhancing our understanding of the biology behind these associations.

Publications

Publications arising during the course of the work described in this thesis:

Nica A.C., Parts L. et al. The architecture of regulatory gene variation across multiple human tissues: the MuTHER study. *In review.* (2010)

Nica A.C., Montgomery S.B. etal. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. **PLoS Genetics** 6(4) (2010)

Xue Y, Zhang X, Huang N, Daly A, …, **Nica A.C.**, …, Tyler-Smith C. Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. **Genetics** 183(3): 1065-77 (2009)

Soranzo N, Rivadeneira F, Chinappen-Horsley U, Malkina I, … , **Nica A.C.**, …, Deloukas P. Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size. **PLoS Genet** 5(4) (2009)

Loos RJ, Lindgren CM, Li S, Wheeler E, … , **Nica A.C.**, … , Mohlke KL. Common variants near MC4R are associated with fat mass, weight and risk of obesity. **Nat Genet** 40, 768-75 (2008)

Nica A.C. & Dermitzakis E.T. Using gene expression to investigate the genetic basis of complex disorders. **Hum. Mol. Gen**. 17, R129-R134 (2008)

Stranger B.E**., Nica A.C.** et al. Population genomics of human gene expression. **Nat Genet** 39, 1217-24 (2007)

Acknowledgements

Firstly, I am very grateful to my supervisor Manolis Dermitzakis, whose guidance and support have been essential for my PhD. Thank you for the exciting discussions, the encouragement to take on challenging projects and the great help with relocating to Geneva. I have learned a lot professionally and personally from you. In particular, I would like to also thank my co-supervisor Inês Barroso for kindly proofreading my manuscripts and giving me helpful comments on the thesis. Thank you for always finding the time to see me when I needed advice.

Throughout the four years, I have been fortunate to have extraordinary colleagues who helped me with my projects and with whom I shared not only an office, but also memorable times at conferences, dinner parties or coffee breaks. For their valuable support I thank Antigone Dimas, Barbara Stranger, Catherine Ingle, Christine Bird, Claude Beazley, Daniel Jaffares, James Nisbet, Maria Gutierrez-Arcelus, Stephen Montgomery, Tuuli Lappalainen and Tsun-Po Yang.

I can hardly imagine a better PhD program than the one at the Wellcome Trust Sanger Institute and I am deeply thankful to the Wellcome Trust and the Sanger Committee of Graduate Studies for caring so well for us.

For the first time while being abroad, living in Cambridge stopped feeling temporary and I owe this unexpected feeling to my dear friends. PhD06 have been great company in the pub, on the tennis court, at concerts, formals and in every other Central or rather Eastern European home country we visited together. Great thanks to Naomi, Andrei, Dan and Mircea for the benefits of missing a boat in Yvoire, the gems on the bookshelf, the most tasty meals with a post-rock twist and caring Millroadian cycling.

My beginning in Geneva would have been even harder if not for Tamara, Ouarda and Djamel cheering me up so many times.

Carmen’s inspiring interludes since early school days make me believe now too in the astral hours of our encounter.

Finally, they say you don’t get to choose your relatives, but that feels inaccurate in my case. For as long as I can remember, my parents have given me confidence to aim high while holding the safety net with love, humour and honesty. This thesis is dedicated to them.

Table of Contents

1 Introduction

1.1 An overview of gene expression

Gene expression is a fundamental cellular process by which a gene gives rise to a functional product and thus produces an observable phenotype. The phenotypic manifestation of genes is ensured by the synthesis of proteins and functional RNA molecules (e.g. rRNAs, tRNAs, microRNAs). These products are the consequence of a regulated flow of genetic information happening almost exclusively one way (Crick 1958): information is first transferred from DNA to RNA (transcription) followed by the transfer of information from RNA to protein in the case of protein-coding genes (translation) (Figure 1.1). A brief overview of these processes is presented in the following section.

Transcription

RNA polymerase II transcribes all eukaryotic protein-coding genes. To initiate transcription, the enzyme requires a set of additional proteins (transcription factors -TFs) which guide its positioning at the promoter and aid in pulling apart the two DNA strands, one of which acts as a template for RNA synthesis. The assembly of the transcription initiation machinery onto DNA is facilitated by the concomitant recruitment of chromatin-modifying enzymes, allowing access to the tightly chromatin-packaged DNA molecule. Following transcription initiation, other TFs guide the RNA polymerase into elongation mode. The single stranded pre-mRNA (primary transcript including both exons and introns) is synthesized in a 5' to 3' direction by adding ribonucleoside monophosphate residues to the free hydroxyl group at the 3′ end of the growing RNA chain (Strachan and Read 2004).

RNA processing

Eukaryotic transcription elongation is tightly coupled to RNA processing (McCracken, Fong et al. 1997). The first modification of the pre-mRNA is the addition of a 5' cap (a modified guanine) to the emerging transcript. This ensures that the cell distinguishes mRNAs from other types of RNA molecules and aids their transport to the cytosol. Following this processing step, introns are excised from the pre-mRNA by endonucleolytic cleavage and exons joined together through the process of RNA splicing.

Alternative splicing - mediated by a large RNA-protein complex called the spliceosome - can give rise to various polypeptide products (isoforms) resulting from different combinations of joined exons. This ability to produce multiple proteins from the same gene increases immensely the coding potential and complexity of eukaryotic genomes (Modrek, Resch et al. 2001; Modrek and Lee 2003). The 3' end of the RNA molecule is also processed, by the addition of a stretch of ~200 A nucleotides (poly-A tail), which helps direct the synthesis of the protein on the ribosome.

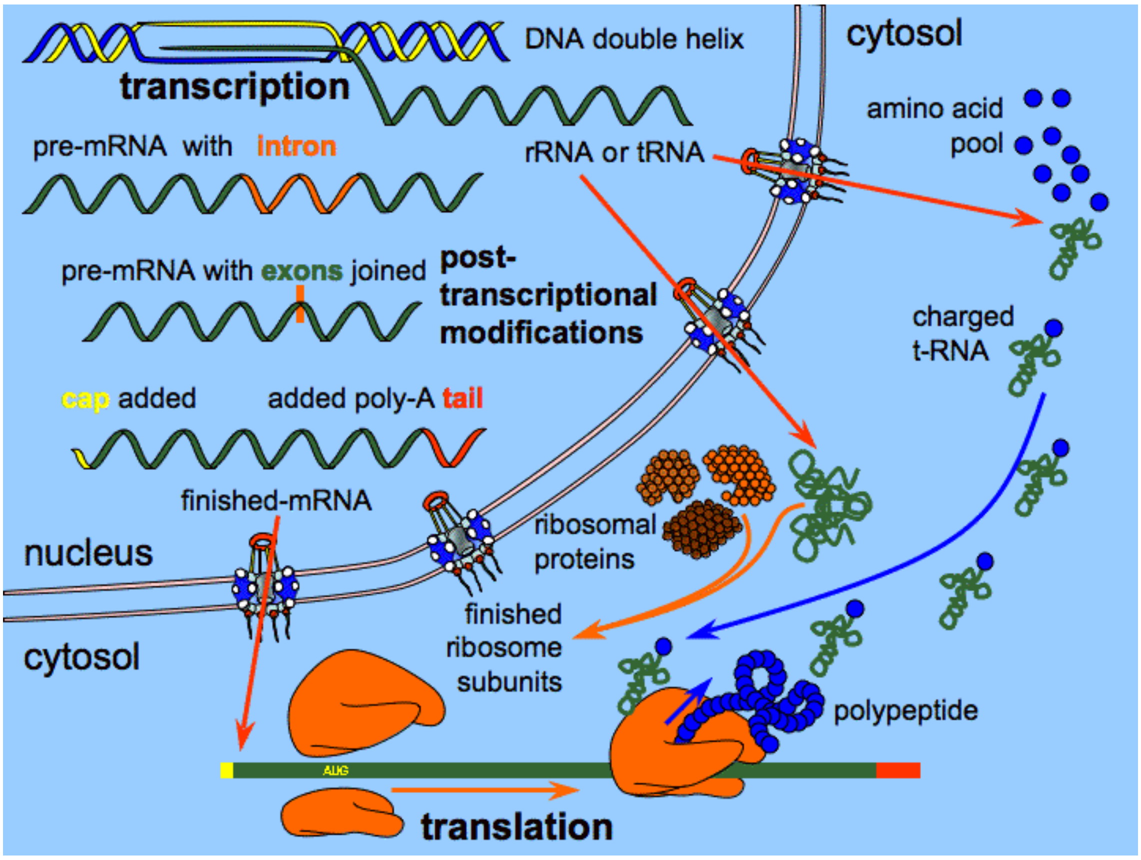


Figure 1.1 The process of eukaryotic gene expression. Inside the nucleus, RNA in transcribed from the DNA template into a primary RNA molecule (pre-mRNA). The pre-mRNA synthesis is followed by a series of processing steps including removal of intronic sequences (RNA splicing), 5' capping and 3' polyadenylation. The resulting processed mRNA molecule is exported into the cytoplasm where it engages with the translational machinery to give rise to the corresponding protein product. Modified from http://plantphys.info/plant_physiology/basiccytology1.shtml.

mRNA transport and translation

For some genes (e.g. rRNA genes, tRNA genes), RNA is the final gene product. Selected mature mRNA molecules however are transported through the nuclear pore into the cytoplasm, where they interact with the translational machinery and engage protein synthesis. Typically, the core of the mRNA is translated, while the flanking 5' and 3' sequences (UTRs – untranslated regions) are copied from the terminal exons to assist the stable binding of the mRNA to the ribosome and start polypeptide synthesis (Strachan and Read 2004). The assembly of the polypeptide is achieved by decoding the mRNA sequence as dictated by the triplet genetic code (three successive nucleotide sequences – codons – specify the corresponding amino acids). The decoding process is mediated by tRNAs bearing specific trinucleotide sequences (anticodons) and covalently bound amino acids which are subsequently inserted in the growing polypeptide chain. Once a stop codon is encountered, translation is terminated and the complete polypeptide released. Post-translational modifications involve attachment of functional groups (e.g. phosphoryl, carbohydrate), proteolytic cleavages or changing the chemical nature of selected amino acids (Mann and Jensen 2003).

1.2 Mechanisms of gene regulation

The brief overview of gene expression presented above highlights the complexity of the process and the multitude of steps involved in its completion. Any of these steps can be regulated to ensure the proper functioning of cells. Specifically, a cell can control the gene products it makes by (a) controlling when and how much of a given gene is transcribed (*transcriptional control*), (b) controlling how the RNA transcript is processed and spliced *(RNA processing control)*, (c) selecting which messenger RNAs are exported into the cytoplasm and where they should be localized *(RNA transport and localization control)*, (d) selecting which mRNAs should be translated *(translational control)*, (e) selectively destabilizing certain mRNA molecules *(mRNA degradation control)*, or (f) activating, inactivating or degrading specific protein products *(protein activity control)* (Alberts 2002). For most genes though, the most common regulatory control point is the initiation of transcription (Guenther, Levine et al. 2007). Given this, I emphasize below the most common points of transcriptional control and note that for the purpose of this thesis, transcript abundance (mRNA levels) was considered a proxy to gene expression.

1.2.1 Transcriptional control

Transcription of eukaryotic genes relies on two fundamental components: 1) stretches of defined DNA sequence in the gene's vicinity and 2) gene regulatory proteins (e.g. TFs) that recognize and bind to these sequences in order to recruit and activate the RNA polymerase. The sequence elements serving as recognition signals for the transcriptional apparatus are referred to as *cis*-acting, whereas gene regulatory proteins typically encoded elsewhere remotely in the genome (few megabases away from the gene or on another chromosome) are called *trans*-acting. Characterizing the full repertoire of regulatory elements is important and projects like the Encyclopedia of DNA elements (ENCODE - (Birney, Stamatoyannopoulos et al. 2007) have made important progress towards this goal. The pilot ENCODE project characterized in detail 1% of the human genome (~30 Mb) representing 44 carefully selected regions of variable gene content or containing functional elements revealed by comparative sequence analysis. The authors highlighted the pervasively transcribed nature of our genome by reporting that the majority of the human DNA sequence is represented in primary transcripts, many of which overlap considerably and include non-protein-coding regions. A multitude of new transcription start sites (TSS) was identified and the distribution of regulatory elements surrounding them was refined as being symmetrical, with no bias towards 5' regions as previously thought (Zhang, Paccanaro et al. 2007). The project offered also new insights into the relationship between chromatin structure and transcriptional control by showing that chromatin accessibility and patterns of histone modifications are predictive of transcriptional activity (Koch, Andrews et al. 2007).

1.2.1.1 *Cis* regulatory elements

The current standard view on transcription regulation involves the interplay of five major *cis*-regulatory elements (Maston, Evans et al. 2006):

1) Promoters, short stretches of DNA sequence immediately upstream of a gene, typically within 200 base pairs (bp) of the TSS. They are composed of different regulatory sequences (core promoter and nearby proximal regulatory elements) which function as a docking site for the basic transcriptional machinery (RNA polymerase and a set of general and promoter-specific TFs).

2) Enhancers, long-distance transcriptional control elements functioning as binding sites for activators, a class of TFs that increase the basal level of transcription initiated at the promoter. They control transcription in a spatial and temporal manner and are hence at the basis of tissue-specific gene expression (Visel, Blow et al. 2009). A consensus DNA-looping model explains how the long physical distances between enhancers and the genes they regulate are overcome: the DNA between the enhancer and the core promoter loops out bringing the enhancer-bound proteins in proximity to the basal transcription complex.

3) Silencers, binding sites for TFs that reduce or repress transcription (repressors). They have been shown to act by blocking the binding of an activator (Harris, Mostecki et al. 2005), competing for an activator binding site (Li, He et al. 2004), or recruiting chromatin-modifying factors and thus blocking access to the promoter (Srinivasan and Atchison 2004).

4) Insulators, sequence boundary blocks (0.5-3 kb long) that prevent genes from being inappropriately regulated by neighbouring transcriptional elements. These DNA segments can preclude undesirable interactions between a distal enhancer and a promoter when situated in between the two (Geyer and Corces 1992; Kellum and Schedl 1992) or can act as barriers against spreading of repressive chromatin (heterochromatin), which might otherwise silence expression (Pikaart, Recillas-Targa et al. 1998).

5) Locus Control Regions (LCRs), groups of multiple *cis* regulatory elements acting upon an entire locus or cluster of genes (Li, Peterson et al. 2002). Each element in an LCR (enhancers, silencers, etc) affects expression differentially and only their cooperative activity determines its spatial/temporal expression properties. Moreover, LCRs seem to provide an open-chromatin domain for the gene cluster they regulate. DNase I hypersensitive sites - chromatin regions often preceding active promoters and having a high sensibility to cleavage by the DNase I nuclease - have been often observed in the proximity of LCRs (Lowrey, Bodine et al. 1992).

1.2.1.2 *Trans* regulatory elements

An abundance of distal protein regulators of transcription (*trans*-regulatory elements) further increases regulatory complexity. These proteins are mostly TFs (sequence specific DNA-binding proteins that mediate transcriptional activation or repression) or elements of chromatin modification complexes which assist the transcriptional apparatus to navigate through chromatin (Levine and Tjian 2003). TFs are broadly classified as general (factors such as the TATA box-binding transcription factor II D - TFIID - which assemble at the core promoter to form the preinitiation complex and are required for transcription of most genes (Workman and Roeder 1987)) and tissue-specific (factors that ensure that certain genes are expressed only in certain tissues: e.g. the hepatic nuclear factor no 5 (HNF-5) modulating liver specific expression (Grange, Roux et al. 1991) or the epidermal-enriched factor KER1 (Leask, Rosenberg et al. 1990) controlling keratinocyte specific expression patterns).

Overall, the interplay of the multiple *cis* regulatory elements with the combinatorial activity of available TFs in specific chromatin-accessible genomic regions determines whether a transcript is being generated and if so, its level of steady-state expression (mRNA abundance). Mutations in any of these numerous components of the transcriptional machinery as well as any other post-transcriptional changes affecting mRNA stability (e.g. miRNA regulatory effects (Selbach, Schwanhausser et al. 2008)), splicing, cell signalling or protein-level modifications (Chen and Rajewsky 2007) influence gene activity.

1.3 Genetics of global gene expression

Gene expression underlies cellular and higher-order phenotypes by determining and maintaining proper transcript levels for each gene in a given cell-type. Understanding the genome-wide properties of regulatory control has been the focus of genetical genomics (Jansen and Nap 2001), a recent field of genetic analysis linking global gene expression with natural sequence variation. In this section, I present the main developments in the field and the methods employed to enhance the current knowledge on the genetics of gene expression.

1.3.1 Gene expression is a heritable quantitative trait

Despite their central role in shaping phenotypes, gene expression levels have been observed to differ significantly among individuals. These differences were first observed in model organisms such as yeast (Brem, Yvert et al. 2002) or mouse (Cowles, Hirschhorn et al. 2002), followed by similar observations in humans (Cheung, Conlin et al. 2003; Schadt, Monks et al. 2003; Morley, Molony et al. 2004).

Furthermore, evidence for familial aggregation of human expression profiles was found (Yan, Yuan et al. 2002; Cheung, Conlin et al. 2003; Monks, Leonardson et al. 2004), suggesting a heritable component of gene expression. Heritability estimates (h^2) capture the proportion of phenotypic variance among individuals in a population attributable to genotypic differences. Therefore, evidence of heritability of a trait makes it amenable for genetic analysis. The lymphoblastoid transcriptome was the first to be estimated as variable among individuals, using pedigree analysis of samples from the Centre d' Etude du Polymorphisme Humain (CEPH) panel of lymphoblastoid cell lines (LCLs) (Cheung, Conlin et al. 2003; Monks, Leonardson et al. 2004; Morley, Molony et al. 2004). The initial heritability estimates differed between these studies, probably mostly due to small sample sizes and experimental artefacts that introduce additional expression variability. Nevertheless, they concurred that a large percentage of genes exhibit significant heritability levels (h^2). Monks etal. analyzed LCLs from 15 families and reported 762 genes of the 2430 (31%) differentially expressed as significantly heritable, with a median h^2 of 0.34 (Monks, Leonardson et al. 2004). Later on, Goring et al. examined expression in lymphocytes isolated from 1240 individuals from 30 large families and estimated that up to 85% of the 19,648 detected transcripts were significantly heritable (median h^2 of 0.23 among all expressed transcripts) (Goring, Curran et al. 2007). The authors also draw attention on the considerable influence on gene expression of environmental factors and the physiological state of an individual at the time of sample collection (e.g. time of blood draw). Overall, the studies above indicated that most transcript levels are influenced by an individual's genetic makeup and justified the upcoming efforts trying to identify the genetic determinants of gene expression variation.

1.3.2 Mapping expression quantitative trait loci (eQTLs)

The quantification of gene expression in numerous individuals from a population made it possible to treat the expression profile of each gene as a quantitative trait (Jansen and Nap 2001). This realization, together with the confirmation of a genetically determined component of gene expression, encouraged a series of efforts to map those regions of the genome that contribute to variation in transcript abundance (eQTLs) (Rockman and Kruglyak 2006).

Initially, small-scale experiments on the genetics of gene expression were performed. Allele-specific expression (ASE) assays confirmed that allelic differences in gene expression are common in autosomal non-imprinted genes (Yan, Yuan et al. 2002). Yan et al. compared relative expression levels of the two alleles for 13 genes in 96 individuals from the CEPH families. They observed allele-specific differences in six of the 13 tested genes, with a 1.3-4.3 fold difference between alleles. Reporter gene assays were also informative with respect to the impact of genetic variation on gene expression (Hoogendoorn, Coleman et al. 2003). Hoogendoorn etal. screened for common polymorphisms the first 500 bp of the 5' flanking region of 170 genes and measured each promoter's ability to promote transcription in three human cell lines. The authors estimated that around a third of the promoter variants tested could significantly alter gene expression levels.

It was the development of microarray platforms however, that made it possible to shift from small-scale quantifications to genome-wide measurements, where transcript abundance of thousands of genes is determined simultaneously in a single experiment. These, combined with genetic variation information, allowed the identification of an abundance of loci with functional effects on gene expression. Two traditional approaches reviewed below have been used for eQTL mapping: linkage and association analysis (Hirschhorn and Daly 2005; Gilad, Rifkin et al. 2008).

1.3.2.1 Linkage mapping

Linkage mapping identifies genetic regions likely containing a causal variant by tracking the transmission pattern of chromosomes through families. The aim is to identify markers co-segregating with the trait of interest, as these are linked to the functional loci driving the phenotype. The main advantage of linkage studies is that they can be performed using a relatively low density of markers (<1000 microsatellites or slightly larger number of single nucleotide polymorphisms (SNPs) in humans) (Gilad, Rifkin et al. 2008). Some of the early genetical genomics studies identified regions controlling gene expression by using genome-wide linkage mapping in cell lines from individuals of the CEPH pedigrees (Monks, Leonardson et al. 2004; Morley, Molony et al. 2004). However, linkage mapping is only successful in detecting rare variants with high penetrance, such as those underlying monogenic 'Mendelian' disorders where the segregating causal allele is found in the same 10-20 cM region within each family (Hirschhorn and Daly 2005). In fact, in their study measuring expression of 23,499 genes in LCLs derived from 15 CEPH families, Monks et al. are only powered to detect eQTLs for 33 genes at a pointwise significance level of .000005. These are, as expected, large effects accounting for > 50% of the expression variance and having very high heritability (75% of the 33 eQTLs have a heritability > 0.76) (Monks, Leonardson et al. 2004). Typically, the regulatory regions uncovered by linkage mapping are also quite large. Morley etal. detect linkage peaks

>5Mb for approximately 1,000 expression phenotypes of the 3,554 expressed genes in LCLs from 14 CEPH families (Morley, Molony et al. 2004). Fine mapping of these large regions is very challenging and depends on the occurrence of recombination events within families. For detecting common variants (minor allele frequency (MAF) ≥ 5%) with smaller effect size on expression, association studies are much better powered and more suitable.

1.3.2.2 Association mapping

Association studies use phenotypes measured in collections of unrelated individuals and dense marker information from the same samples (typically > 500,000 genotyped SNPs in humans) in order to detect statistically significant correlations between marker genotypes and the analyzed trait (transcript abundance in the case of eQTL studies). Again, the assumption is that the causal locus is linked, or correlated with the markers showing statistical associations with the phenotype. The extent of this correlation is determined by linkage disequilibrium (LD), a property of the genome describing the non-random association of alleles at different loci in a population (Rockman and Kruglyak 2006). The preferential association of allelic combinations is reflected in the haplotype structure of the genome (Figure 1.2.), whereby a set of highly correlated genetic markers (LD blocks) undisrupted by recombination mechanisms are inherited together through generations (Paigen and Petkov 2010). The size of the LD blocks across the human genome is variable but nevertheless, they provide a much better resolution than linkage maps, with causal variants typically to be found within windows of few tens or hundreds of kilobases (kb) (Dawson, Abecasis et al. 2002).

Taking advantage of the correlation structure in the genome, the International HapMap project was launched in 2002 as a large-scale collaborative effort with the goal to identify and catalogue most of the common human genetic variation (Consortium 2003). The purpose of a detailed haplotype map (HapMap) of the human genome was to serve as a public resource of genetic markers and facilitate subsequent association studies with various phenotypes. The project was a success and its scale has been growing considerably throughout the years, reaching now its third phase.

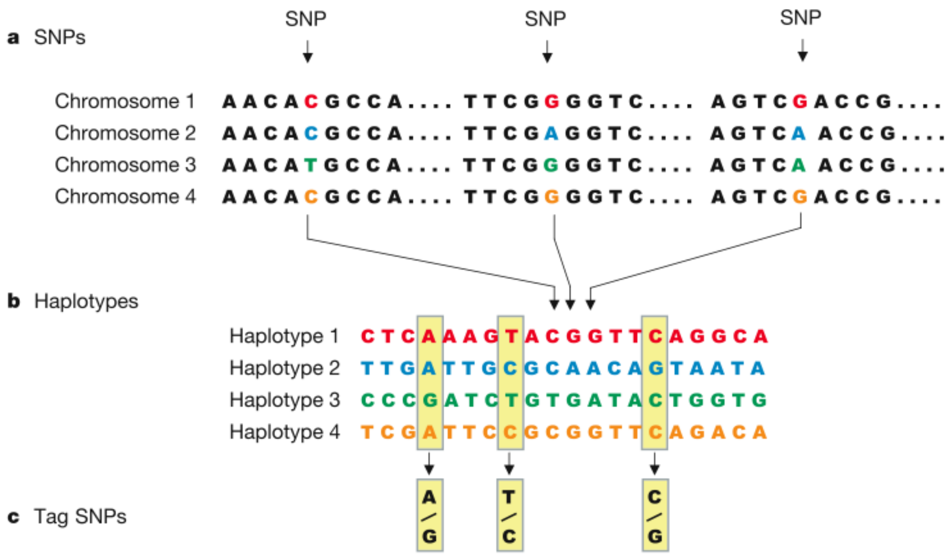

a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G

Figure 1.2. The haplotype structure of the human genome. a) A short DNA sequence from the same chromosome is shown in four different individuals. The nucleotides are identical for most genomic positions except at three variable loci (SNPs) b) A particular combination of alleles observed in a population is called a haplotype. In this example, four haplotypes capture the sequence variation in the population of the DNA region in panel **a**. Only the 20 variable loci of the total 6,000 DNA bases represented are shown. c) To uniquely identify these four haplotypes, it is sufficient to genotype three tag SNPs out of the 20 variants. For example, Haplotype 1 can be recognized in any individual having the A-T-C pattern at the three tag SNPs. Typically, many chromosomes would carry the common haplotypes in the population. Figure adapted from the International HapMap Project, *Nature* 426, 789-796 (2003).

Phase 1 of the HapMap project aimed at genotyping at least one common SNP every 5 kb across the genome in each of the 269 samples belonging to four geographically distinct populations. Additionally, ten ENCODE regions, each 500 kb long, were sequenced in 48 individuals and all SNPs in these regions were genotyped in the full sample set The 269 samples consisted of: 90 individuals (30 parents-offspring trios) of northern and western European ancestry living in Utah from the Centre d'Etude du Polymorphisme Humain collection (abbreviated CEU), 90 individuals (30 trios) from the Yoruba in Ibadan, Nigeria (YRI), 45 Han Chinese from Beijing, China (CHB) and 44 Japanese from Tokyo, Japan (JPT). At this stage, a total of approximately 1.3 million SNPs were genotyped in each population (Consortium 2005).

In its second phase (HapMap 2), 270 individuals (the 269 Phase 1 individuals plus an additional JPT sample) were genotyped for a further 2.1 million SNPs. The resulting denser SNP map (approximately one common SNP per kb) contains an estimated 25-35% of the total 10 million SNPs expected across the human genome (Frazer, Ballinger et al. 2007).

The current and largest phase of the project (HapMap 3) involves an extension of the genotyped sample set, both by supplementing the initial four-population collection with more individuals and also by adding samples from seven other populations (http://hapmap.ncbi.nlm.nih.gov/). As such, over 4 million SNPs were genotyped from individuals of the Phase 1 and 2 populations (180 CEU, 90 CHB, 91 JPT, 180 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).

In combination with large-scale expression data, these well-documented common genetic variation maps enabled the success of detecting eQTLs using population association studies (Cheung, Spielman et al. 2005; Stranger, Forrest et al. 2005). Building on their previous whole-genome linkage work, Cheung et al. used > 770,000 HapMap 1 SNP markers and mRNA abundance measurements from 57 CEU individuals to map eQTLs for previously identified expression phenotypes. Among the chosen 27 phenotypes with significant linkage evidence ($P < 3.7 \times 10^{-5}$), the authors confirmed 70% as having significant evidence of association ($P < 0.0001$). For all the concordant signals between the two methods, they were also able to narrow down the candidate functional regions, making thus use of the better resolution conferred by LD (Cheung, Spielman et al. 2005). Stranger etal. further explored the power of association studies by performing a genome-wide analysis of 630 genes in LCLs from 60 unrelated CEU individuals genotyped in HapMap 1. For the subset of 374 expressed genes, the authors detected eQTLs for up to 40 genes, with the majority of the eQTLs identified mapping in the proximity of the genes they associate with. Laying the ground for future genome-wide expression studies, Stranger etal. paid special attention to the multiple-testing problem and evaluated three statistical correction methods to reduce false positives, namely Bonferroni (Miller 1981), false discovery rate (FDR) (Storey and Tibshirani 2003) and permutations (Churchill and Doerge 1994). The Bonferroni method is prone to conservative estimates of significance since it does not account for the dependence of SNPs due to LD and treats each SNP – gene test as independent. The authors nevertheless report a generally good concordance among the different multiple-testing correction methods. Based on the highest enrichment of significant discoveries, the results favoured permuting the expression values on the genotypes as a very suitable statistical correction strategy (Stranger, Forrest et al. 2005).

Besides multiple-testing, another caveat of genome-wide association mapping is the occurrence of false positives because of population substructure. In this case, allele frequency differences due to systematic ancestry differences between individuals having a dissimilar profile of interest (gene expression pattern or disease status) can cause spurious associations. Careful consideration of this issue lead to the development of appropriate statistical correction methods such as principal component analysis, modelling explicitly the inter-individual differences in ancestry prior to association testing (Price, Patterson et al. 2006). Altogether, these statistical advancements contributed unequivocally to the success of genome-wide association mapping.

In addition to single base variations (SNPs), structural DNA variations greater than 1 kb and present at variable copy number compared to the reference genome (copy number variants - CNVs) have also been successfully mapped. Initial observations suggested that CNVs are commonly present across the human genome and alluded to their substantial contribution to genetic variation in the population (Iafrate, Feuk et al. 2004; Sebat, Lakshmi et al. 2004). Consequently, their likely substantial effect on phenotypic variation resulting from gene dosage alteration, disruption of coding sequences or perturbation of long-range interactions (Kleinjan and van Heyningen 2005) elicited considerable attention. The contribution of CNVs to gene expression variation was assessed first in LCLs derived from the HapMap 1 samples (Stranger, Forrest et al. 2007). The association analysis between expression levels of 14,925 transcripts and correspondingly typed SNP and CNV variants in the 210 unrelated HapMap individuals revealed a series of *cis* effects: a total of 888 nonredundant genes associated with at least one SNP and 238 nonredundant genes associated with at least one autosomal CNV. The authors estimated that 83.6% of the expression variation is attributable to SNPs while CNVs capture 17.7% of the total expression variation in the current samples, with little overlap between them. Recently however, the higher CNV resolution enabled by tiling-array comparative genome hybridization (CGH) approaches indicates that CNV effects are much less dramatic than initially suspected. In conjunction with the denser SNP maps available, it was concluded that the contribution of common CNVs to phenotypic variance (including thus mRNA levels) is already captured to a great extent by neighbouring SNPs (McCarroll, Kuruvilla et al. 2008; Conrad, Pinto et al. 2010).

1.3.2.3 *Cis* and *trans* eQTLs

One of the major advantages of eQTL mapping using the genome-wide association approach is that it permits the identification of new functional loci without requiring any previous knowledge about specific *cis* or *trans* regulatory regions. Typically in the eQTL mapping literature, regulatory variants have been characterized as either *cis* or *trans* acting, depending on the physical distance from the gene they regulate (Figure 1.3). In this thesis variants within one megabase (Mb) on either side of a gene's TSS were called *cis*, while those at least five Mb downstream or upstream of the TSS or on a different chromosome were considered *trans*-acting.

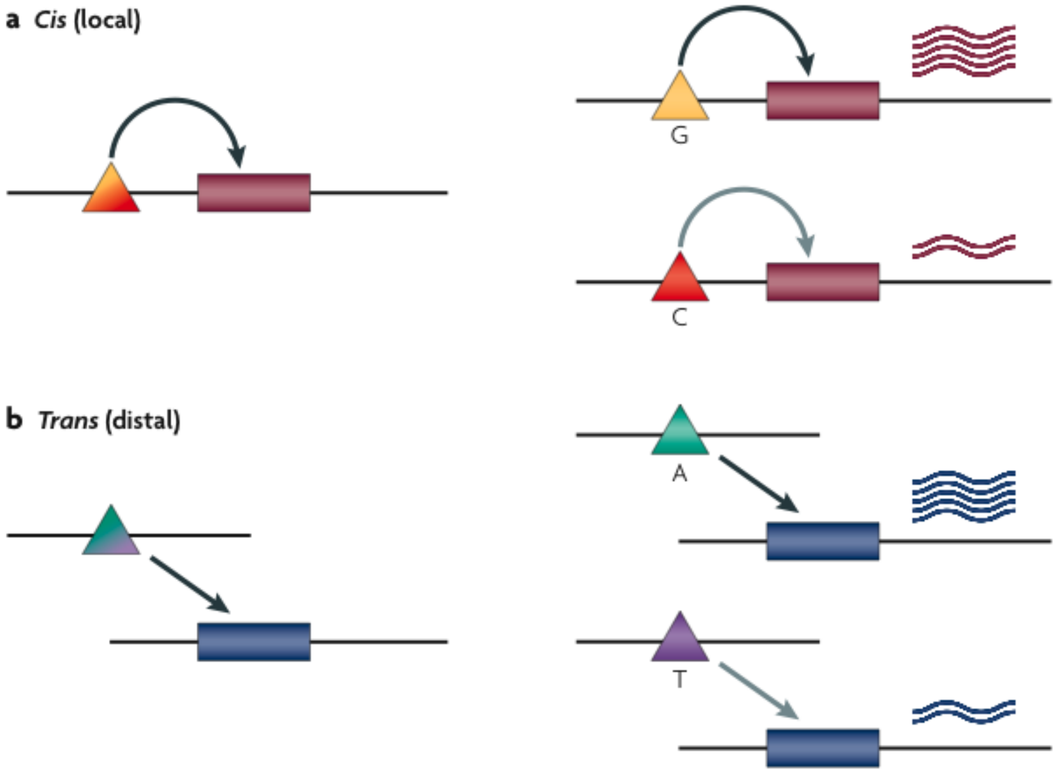


Figure 1.3. *Cis* and *trans* effects on transcript levels. Polymorphic regulatory variants (SNPs) affecting variation in a gene's transcript levels in *cis* (a) or in *trans* (b). The *cis* variant is located close to the gene it regulates. Individuals with the G allele at the *cis* eQTL have higher expression levels than individuals with the C allele. *Trans* variants are located at a much further genomic distance from the gene they regulate. There too, a particular allele (A in this case) drives high expression levels as opposed to the T allele determining lower mRNA levels. Modified from (Cheung and Spielman 2009).

Studies so far explain most of the variance in gene expression locally, by sequence variants in the vicinity of the associated genes. In a large-scale expression study where lymphocytes in 1,240 individuals were profiled, the authors identified 1,345 *cis*-regulated transcripts at an FDR rate of 5% of the total 19,658 tested (Goring, Curran et al. 2007). A study in our lab detected numerous *cis* effects in transformed B-cells. The analysis of 270 lymphoblastoid cell lines derived from the HapMap 2 individuals and genotyped for 2.2 million common SNPs revealed 831 genes of the 13,643 tested as having a significant *cis* eQTL (Stranger, Nica et al. 2007). Since power increases with the availability of larger sample sizes, the number of genes detected to have eQTLs is also expected to increase. Finding *trans* eQTLs has been less successful so far, mainly because interrogating the whole-genome for potential regulatory effects is a daunting statistical and computational task, requiring the correction for millions of tests. Whether the current enrichment of *cis* versus *trans* eQTLs reflects biological reality and is not just attributable to low power in *trans* is still under debate (Wray 2007; Wittkopp, Haerum et al. 2008).

1.3.3 Population differentiation of gene expression

Several studies have analyzed expression data in populations of different ancestry and revealed substantial differences at many loci. A study on 16 individuals of European and African descent estimated that 17% of genes were differentially expressed between populations (Storey, Madeoy et al. 2007). Differences were found also between European and Asian–derived populations for 1,097 of 4,197 genes tested (Spielman, Bastone et al. 2007). Larger scale studies confirmed the initial estimates. The eQTL study on 270 individuals of the four HapMap 2 populations of European (CEU), Asian (CHB, JPT) and African (YRI) descent reported that 17-29% of loci have significant differences in mean expression levels between population pairs (Stranger, Nica et al. 2007). While some of these observations are due to environmental factors (Idaghdour, Storey et al. 2008), genetics plays an important role in shaping the observed differences. Price etal. provide evidence for population differentiation due to genetic effects using cell lines derived from an admixed African American population (Price, Patterson et al. 2008). They estimated a mean value of 0.2 and a median of 0.12 in the proportion of gene expression variation attributable to population differences. A large proportion of the genetically determined variation in gene expression across populations has been explained by different allele frequencies (Spielman, Bastone et al. 2007), suggesting that regulatory mechanisms are probably not fundamentally different between populations.

1.3.4 Multiple-tissue studies

So far, the majority of human eQTL studies have been performed exclusively on blood-derived cells or cell lines. This relatively easily accessible cell-type has been very useful in understanding the genetics of gene expression and continues to be a great resource. However, as gene expression signatures are cell-type specific (Alberts 2002), the question arises whether regulatory control of steady-state expression is also cell-type dependent. Estimates vary depending on the tissues being compared and the eQTL methods used, but generally, a significant tissue-specific component of *cis* regulation has been systematically reported.

Myers et al. analyzed for the first time the genetics of gene expression variability in the human brain. After expression profiling and genotyping 193 neuropathologically normal human brain samples, the authors estimated that 58% of the transcriptome is cortically expressed and identified significant eQTLs for 21% of the expressed transcripts (2,975 of the total 14,078 tested). A comparison of the cortical results with eQTLs previously identified in LCLs from CEPH individuals resulted in barely any overlap. While some degree of brain-specific control of gene expression is expected, the marked lack of overlap observed here is exacerbated by the different microarray platforms used and the distinct samples profiled in the two experiments (Myers, Gibbs et al. 2007). In a study comparing adipose and blood expression patterns between two Icelandic cohorts of considerable sample size (673 and 1,002 individuals respectively), 50% of the *cis* eQTLs detected were shared (Emilsson, Thorleifsson et al. 2008). Another study overlapping eQTLs identified in 93 autopsy-derived cortical tissue samples and 80 peripheral blood mononucleated cell samples outlined the distinct genetic control of expression in the two tissues, reporting <50% sharing (Heinzen, Ge et al. 2008). Finally, a study in our lab compared the regulatory landscape in three tissues (fibroblasts, LCLs and primary T cells) derived from the same set of 75 European individuals. Unlike previous studies, this unique dataset properly accounts for confounding factors such as differences in population samples, array platforms or statistical methods. The authors reported that 69-80% of *cis* eQTLs are cell-type specific, augmenting thus the need to study multiple tissues to determine the full spectrum of regulatory variants (Dimas, Deutsch et al. 2009).

1.3.5 Environmental and epistatic effects on expression

Gene expression is a complex trait shaped, in addition to genetic factors, by environmental conditions. Lifestyle (e.g. diet, smoking), geographic conditions or age have been shown to have a considerable impact on expression, sometimes even larger than that attributed to genetic effects (Idaghdour, Storey et al. 2008). Moreover, experimental treatment of cells can markedly change their expression patterns (Choy, Yelensky et al. 2008). Exposing cells to different perturbations also revealed that individuals differ in their response to external stresses (e.g. ionizing radiation) and lead to the identification of DNA variants that influence this differential response (Smirnov, Morley et al. 2009). Further such studies will be very useful for understanding the genetics of differential toxin response in view of improving drug administration. This has been the focus of pharmacogenomics, a field aiming to identify genetic determinants of drug response, i.e. polymorphisms influencing the activity of drug metabolizing genes (Evans and Relling 1999). Differential tissue and organ response to other disease relevant stimuli (e.g. insulin) is also likely to influence disease status and will need appropriate consideration in future association studies.

Interactions between genetic factors (Brem, Storey et al. 2005; Dimas, Stranger et al. 2008), but also between genetic and environmental factors (Gibson 2008) have an effect on gene expression as well. However, in the absence of good hypotheses for which particular combinations of DNA variants to test and under which model, detecting epistatic effects is statistically still very challenging.

1.4 Gene expression shapes cellular and high-order phenotypes

The role of gene regulation in shaping phenotypes has been pointed out early on, starting with the landmark paper of King and Wilson (King and Wilson 1975). Through comparative protein analysis, they were able to demonstrate that humans and chimpanzees are 99% genetically identical and thus suggested that significant species differences are probably due to gene regulatory variation. Since then, gene expression has been implicated in associations with a wide range of cellular and high-order phenotypes. A succinct review of a number of key examples certifying the role of gene expression variation in shaping phenotypes is presented below.

1.4.1 The role of expression in defining and maintaining cell-specificity

The control of gene expression is fundamental for the formation of specialized differentiated eukaryotic cells. Precise spatial and temporal gene regulation during development determines cell fate and helps maintain differentiated cell-specific signatures through subsequent cell generations (Alberts 2002). One example of a highly coordinated regulatory genetic switch mechanism is that involved in the formation of muscle cells during embryonic development. The development of muscle cells depends on the expression of myogenic proteins (MyoD, Myf5, myogenin, and Mrf4), a family of helix-loop-helix regulatory proteins. These bind to specific regulatory sequences surrounding muscle-specific genes and activate their transcription (Weintraub, Davis et al. 1991). Through a series of positive feedback loops, myogenic proteins further stimulate transcription of other gene regulatory proteins involved in muscle cell development. The deterministic role of gene expression in cell-type formation was additionally confirmed by observing the ability of myogenic proteins to trigger muscle differentiation in other cell-types (e.g. the human myogenic factor Myf5 induces myogenic phenotypes such as formation of multinuclei and synthesis of sarcomeric myosin heavy chains when transiently expressed in embryonic mouse fibroblasts (Braun, Buschhausen-Denker et al. 1989)).

Once cells differentiate into a certain cell-type, they remain specialized and transmit their specific expression signatures to daughter cells. This is attained by feedback loops wherein key gene regulatory proteins activate their own transcription or that of cell-type specific genes they interact with (Alberts 2002). Chromatin signatures also ensure the faithful propagation of cell-type specific expression, as unexpressed genes are packaged into compact chromatin forms, inaccessible to the transcriptional apparatus (Boyle, Davis et al. 2008). Specificity of gene expression in different cell-types has been extensively observed and Adams et al. were among the first to report this at a genome-wide scale (Adams, Kerlavage et al. 1995). In their study sampling 30 tissues with more than 1000 ESTs (expression sequence tags) each, they detected only eight genes matched by ESTs in all 30 tissues and 227 genes represented in at least 20 tissues. A following large-scale study on 46 human and 45 mouse tissues, organs and cell lines reported only 6% ubiquitously expressed genes and ascertained that tissue-specific gene clusters can be found in nearly all tissues examined (Su, Cooke et al. 2002).

1.4.2 Gene expression shapes complex phenotypes in the natural and disease range

Despite the essential role of regulatory control in ensuring normal functioning of cells, most biological systems are remarkably robust, showing abundant gene expression variation (Oleksiak, Churchill et al. 2002; Gilad, Oshlack et al. 2006). Much of the natural phenotypic variation including numerous adaptive features of various organisms has been associated with changes in gene expression. One of the early examples included the differential expression of the Hox gene *Ultrabithorax*, which has been shown to pattern fine hair outgrowths (trichomes) on the posterior femur of the second leg in *Drosophila* (Figure 1.4). The evolution of *cis*-regulatory elements of this gene, rather than the protein itself was indicated as responsible for these adaptive morphological changes (Stern 1998). Similarly, evolution of a *cis*-regulatory element in the *yellow* gene has been shown to contribute to the gain of a male-specific pigmentation spot in *Drosophila biarmipes* (Gompel, Prud'homme et al. 2005) while comparative analysis of expression patterns of growth factors in Darwin’s Finches revealed that differential expression of *Bmp4* has a major role in determining beak morphology (Abzhanov, Protas et al. 2004).

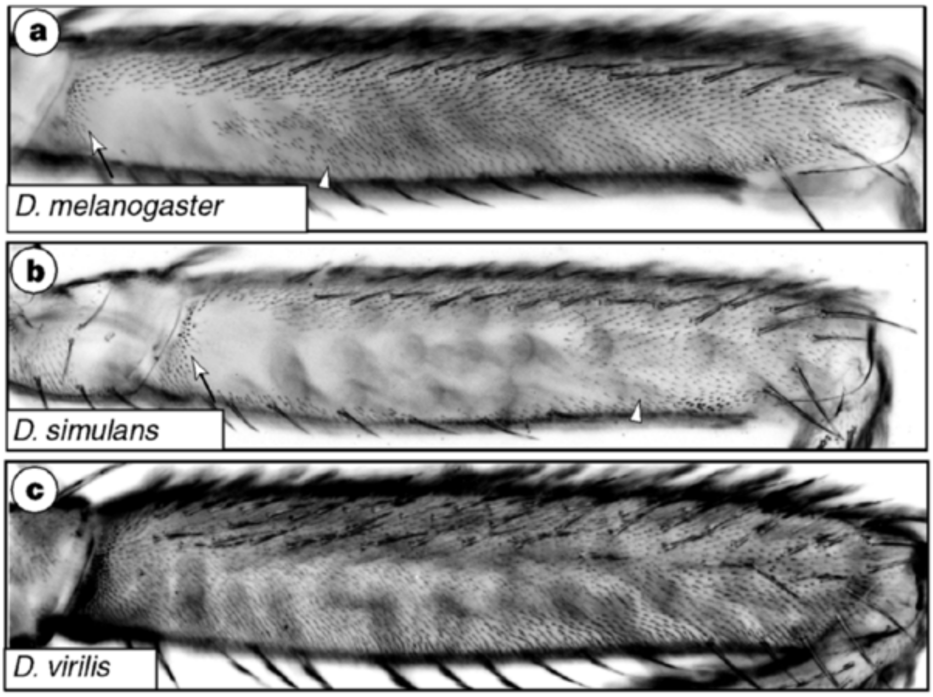


Figure 1.4. Different trichome patterns among *Drosophila* species. The posterior second femur of three *Drosophila* species is shown. The morphological differences observed are due to differential regulation of the Hox gene *Ultrabithorax.* Modified from (Stern 1998).

In humans, regulatory variation has been also associated with a series of phenotypic changes. A common well-studied example is that of lactase persistence (Ingram, Mulcare et al. 2009). Digestion of lactose, the sugar essential for nourishment of newborn mammals, is facilitated by lactase, a small intestinal enzyme encoded by the *LCT* gene. After weaning, the production of lactase decreases significantly in humans resulting in the inability to digest milk (lactose intolerance). Some humans however, are able to express *LCT* in adulthood (lactase persistence) and this has been especially observed in regions with traditional practice of milking (Figure 1.5), suggesting that the locus has been subject to strong positive selection (Holden and Mace 1997). The worldwide differential lactase expression has a genetic determinant and that was shown to be a *cis*-regulatory element (Wang, Harvey et al. 1995).

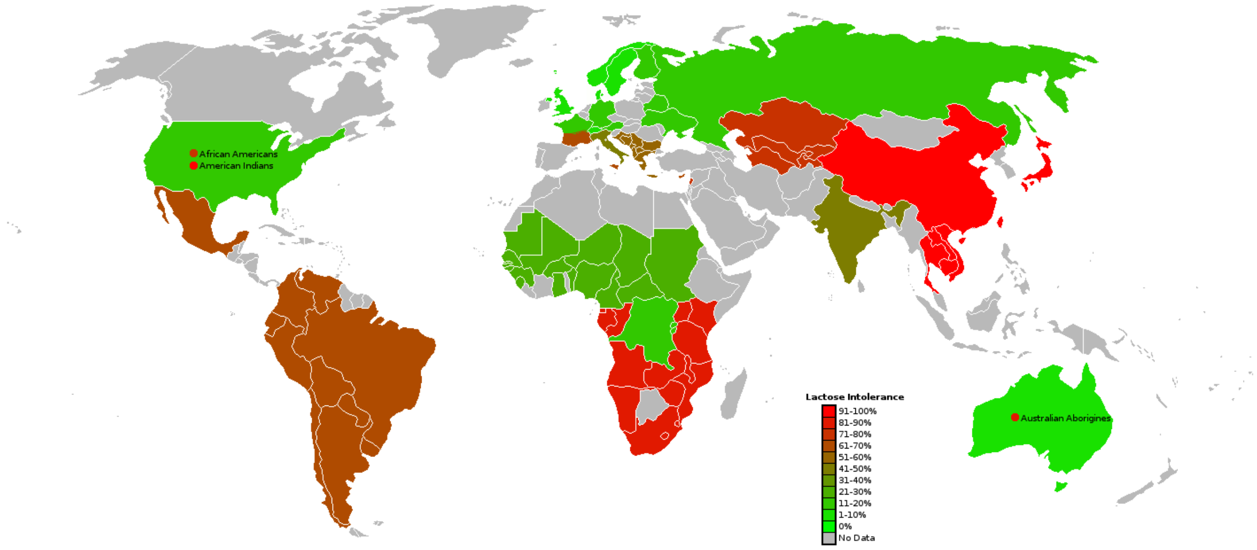


Figure 1.5. Geographic distribution of lactose intolerance. World map showing the distribution of lactose intolerance by region. Red indicates a high intolerance percentage and green a low percentage of lactose intolerance. The regions of low intolerance (or lactase persistence) coincide with areas of known cattle farming tradition. Image from http://en.wikipedia.org/wiki/Lactose_intolerance.

Gene expression changes beyond a tolerance limit can have more serious phenotypic consequences and prove detrimental. Decreased or complete loss of α-globin expression has been associated with α-thalassaemia (Weatherall 1998) and a regulatory SNP mapping in between the α-globin gene cluster and its upstream regulatory elements has been linked to the disease by causing significant down-regulation of the α^D, α2 and α1 genes (De Gobbi, Viprakasit et al. 2006). Low levels of adenomatous polyposis coli (*APC*) expression predispose to hereditary colorectal cancers (Yan, Dobbie et al. 2002) and over-expression of *C-MYC* (v-myc myelocytomatosis viral oncogene homolog) can lead to Burkitt's lymphoma (Boxer and Dang 2001). Specific regulatory polymorphisms inducing differential gene expression associated with complex traits have also been identified. Progression of coronary atherosclerosis has been associated with reduced expression of human stromelysin-1, regulated by a common *cis*-regulatory variant at the promoter (Ye, Eriksson et al. 1996). Susceptibility to autoimmune disorders has also been attributed to changes in expression: variants in the noncoding 3' region of the cytotoxic T lymphocyte antigen 4 gene (*CTLA4*) correlating with lower mRNA levels of a *CTLA4* splice variant were identified as disease determinant candidates for Graves' disease, autoimmune hypothyroidism and type 1 diabetes (Ueda, Howson et al. 2003). Taken together, these examples highlight the considerable range of phenotypic changes attributable to gene expression variation.

1.5 Genetics of complex diseases

Concomitant with the progress in understanding the genetics of gene expression, new insights into the genetic causes of common diseases have also been obtained. In the current section I briefly outline the developments that lead to this progress and the challenges that still exist in this area.

1.5.1 The road to genome-wide association studies (GWAS)

During the last twenty years, genetic mapping techniques allowed the elucidation of numerous rare monogenic disorders (Jimenez-Sanchez, Childs et al. 2001). Identifying genetic determinants of common diseases on the other hand, was lagging behind. The hitherto most common methods for uncovering disease genes, candidate gene approaches and family studies using linkage analysis, have been both failing to identify causative loci for complex traits.

Candidate gene studies are impractical since they are conditioned by often-unavailable information about disease biology. Even when a proposed relationship with a complex phenotype exists, finding causative variants within candidate genes has been unsuccessful, mainly because of the small sample sizes and the lenient statistical criteria by which associations were deemed causal. As such, many of the genotype-phenotype associations reported based on candidate-gene approaches failed to replicate in independent studies (Ioannidis, Ntzani et al. 2001).

Linkage approaches were not far more successful, unsurprisingly given the properties of complex traits. Common complex disorders do not follow simple Mendelian inheritance patterns and are in addition characterized by multiple gene-gene and gene-environment interactions (Lander and Schork 1994). Each of these factors has a relatively small individual contribution to the determination of the ultimate disease phenotype. A number of additional aspects explain why it is improbable to find variants that impact common diseases and also co-segregate in families: (1) incomplete penetrance, whereby not all individuals inheriting the predisposing allele manifest the phenotype (2) locus and allelic heterogeneity, when mutations in any of several genes and different mutations within a gene respectively may give rise to the same phenotype or (3) pleiotropy, occurring when a single gene has multiple parallel phenotypic effects (Lander and Schork 1994). Under these circumstances, only few attempts to uncover complex disease loci using linkage analysis were successful (e.g. *NOD2* associated with Crohn's disease susceptibility (Hugot, Chamaillard et al. 2001; Ogura, Bonen et al. 2001)). Numerous other putative candidates were not further replicated, limiting in this way our understanding of complex diseases. In a classical paper in the field, Risch and Merikangas explained that the failure of replication was due to the limited power of linkage analysis to detect small genetic effects (Risch and Merikangas 1996). Thus, many of the proposed candidates were false positives and in fact, an unachievable sample size of more than 2500 families would be required to detect loci having low genotypic relative risk (typically ≤ 2 for complex disorders) with a minimum 80% power. Despite technical limitations at the time, the authors proposed genome-wide association studies (GWAS) as an alternative powerful approach.

1.5.2 The GWAS revolution

In the past few years, the ability of GWAS to help understand the genetic basis of complex disorders has become apparent (WTCCC 2007). The outburst of successful GWAS is owed to the availability of well-documented common human genetic variation maps (e.g. HapMap project (Frazer, Ballinger et al. 2007)), large patient samples with accurately recorded phenotypic information as well as appropriate statistical methods to assess significance (Rice, Schork et al. 2008). This approach has revealed a multitude of disease-susceptibility loci, now stored in an updated catalogue at the National Human

Genome Research Institute (NHGRI) (Hindorff, Sethupathy et al. 2009). For several common disorders such as type 1 (Hakonarson, Grant et al. 2007; Todd, Walker et al. 2007) and type 2 diabetes (Scott, Mohlke et al. 2007; Sladek, Rocheleau et al. 2007; Zeggini, Weedon et al. 2007; Zeggini, Scott et al. 2008), prostate cancer (Eeles, Kote-Jarai et al. 2008; Thomas, Jacobs et al. 2008) or inflammatory bowel disease (Parkes, Barrett et al. 2007; Rioux, Xavier et al. 2007), a multitude of predisposing loci has been reported. Most of these studies featured case-control designs, whereby a group of selected individuals diagnosed with a disorder of interest (cases) is compared to a group of people not ascertained for that phenotype (controls). The goal is to detect susceptibility alleles having marked frequency differences between the two groups. This design requires population stratification corrections and careful case/control selection in order to avoid misclassification biases (classification of cases as non-diseased), which can all decrease the power to detect associations (McCarthy, Abecasis et al. 2008).

Most recently, GWAS have been performed on population-based cohorts, offering insights into the genetics of continuous traits, be they anthropomorphic like height (Weedon, Lettre et al. 2007; Lettre, Jackson et al. 2008) or disease relevant (e.g. fat mass (Frayling, Timpson et al. 2007; Loos, Lindgren et al. 2008), lipids (Saxena, Voight et al. 2007; Willer, Sanna et al. 2008; Teslovich, Musunuru et al. 2010)). The discovery of multiple susceptibility variants per complex trait, each of small effect size, as well as their considerable pleiotropic overlap (Stratton and Rahman 2008) is gradually shifting the focus from viewing disease as a dichotomous trait towards a quantitative view, where common disorders are the extremes of a spectrum of quantitative traits (Figure 1.6) (Dermitzakis 2008; Plomin, Haworth et al. 2009).

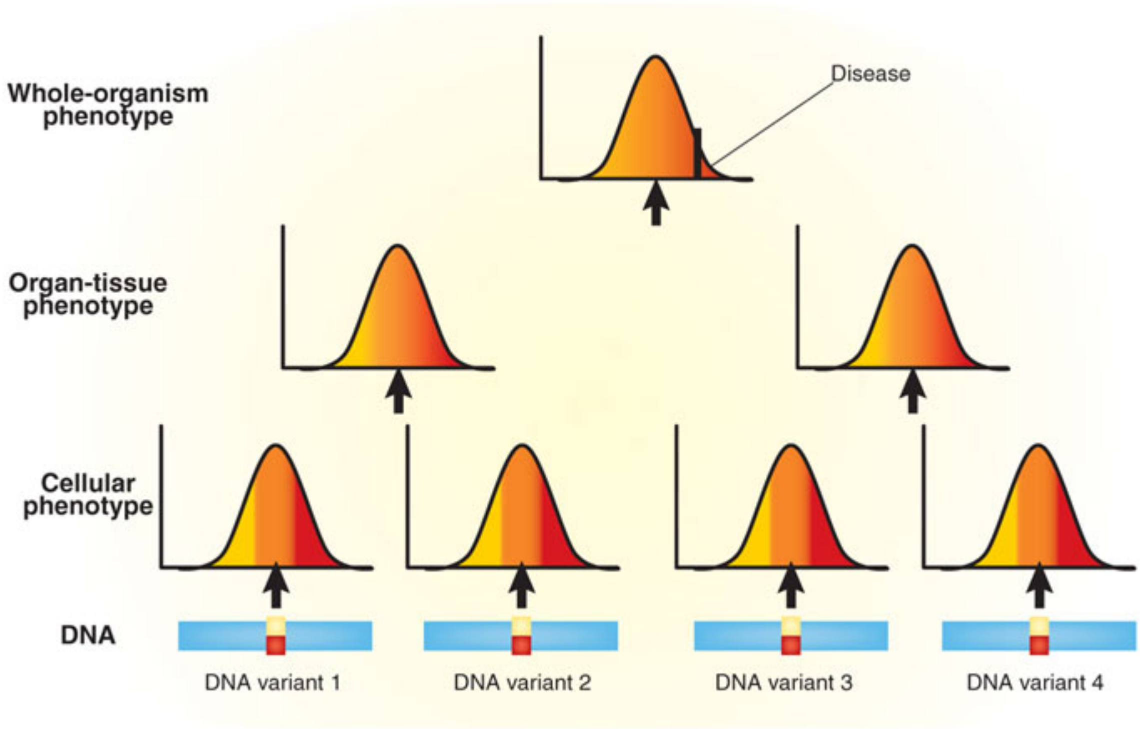


Figure 1.6. Common disorders as quantitative traits. Disease state can be viewed as the tail of a spectrum of continuous phenotypes**.** **I**n this schematic example**,** four unlinked DNA variants determine the whole-organism phenotype through changes at the cellular level, which in turn affect intermediate organ-tissue phenotypes. Yellow to red gradients represent the effect of each of the four red and yellow DNA variants at the different ends of the phenotypic spectra. At the cellular level, these effects are easy to interpret and to detect, whereas at organismal level, the power to detect them is reduced owing to the large number of direct and indirect intermediate interactions. Adapted from (Dermitzakis 2008).

A wide range of quantitative trait data of potential disease relevance is nowadays being collected and combined into large-scale meta-analyses. With greater power, such efforts identify disease affecting loci and the intermediate quantitative traits underlying them (e.g. Prokopenko etal. and Dupuis et al. look at fasting glucose levels as a continuous trait, find variants that associate with glucose concentrations and subsequently identify type 2 diabetes susceptibility loci (Prokopenko, Langenberg et al. 2009; Dupuis, Langenberg et al. 2010)).

1.6 Promise of eQTL studies for disease genetics

Despite the impressive success of GWAS, there is a substantial gap between the susceptibility variants discovered and understanding how those respective loci contribute to disease. Frequently such loci map to genomic regions of no apparent function (non-coding) or the genome’s tight correlation structure (LD) does not permit firm conclusions about functional effects (i.e. which is the causal variant and which gene function does it affect). Under these circumstances, the need to incorporate additional information for interpreting GWAS results became evident. The direct link between DNA polymorphisms (usually SNPs) and variable transcript levels along with the increasing role attributed to regulatory variation in shaping phenotypic differences, nominated gene expression as an important mechanism underlying complex traits. Subsequently, I describe the main results obtained so far in support of this hypothesis.

1.6.1 GWAS SNPs can be strong eQTLs

Comparing expression levels of individual genes between cases and controls may not be sufficiently powered to detect significant differences (Cookson, Liang et al. 2009) and discriminating between causal and reactive expression changes would be a tough challenge. However, genetic markers simultaneously associated with disease status and eQTLs are very interesting: if one allele is more frequent in cases than controls and at the same time it is causal for gene expression effects of a nearby gene, which is itself important for the disease, then it is likely that causality can be established. Several recent studies have shown the value of this principle by incorporating eQTL analyses with GWAS results and thus proposing candidate disease genes. Moffatt etal*.* identified a series of strongly correlated SNPs in a 200 kb region of chromosome 17q23 associated with childhood asthma (Moffatt, Kabesch et al. 2007). The association region contained 19 genes, none of which had an evident disease role. Expression analysis on lymphoblastoid cell lines derived from the same families showed that the most significant GWAS SNPs also explained ~29.5% of the variance in transcript levels of one of those 19 genes, *ORMDL3* (ORM1-like 3), now the best candidate for further functional studies. Expression data has helped interpret some of the association signals for Crohn’s disease as well. Initial findings of a recent GWAS included multiple susceptibility loci mapping to a 1.25 Mb gene desert region on chromosome 5 (Barrett, Hansoul et al. 2008). eQTL data showed that one or more of these loci act as long-range *cis* regulators of *PTGER4* (prostaglandin E receptor 4), a gene 270 kb away from the associated region whose homologue has been implicated in phenotypes similar to Crohn’s disease in the mouse (Libioulle, Louis et al. 2007). Other similar examples for height (Gudbjartsson, Walters et al. 2008), systemic lupus erythematosus (Hom, Graham et al. 2008), type 1 diabetes (Hakonarson, Grant et al. 2007) or bipolar disorder (WTCCC 2007) support the use of eQTL data in aiding the interpretation of GWAS results.

However, not all cases are so straightforward, as shown by the association of the *SH2B1* (SH2B adaptor protein 1) locus to body mass index (BMI) (Willer, Speliotes et al. 2009). In this case, a non-synonymous genome-wide significant SNP in *SH2B1* was associated also with differential expression of two other genes (*EIF3C* – eukaryotic translation initiation factor 3, subunit C and *TUFM* – Tu translation elongation factor, mitochondrial). Functional evidence from mice, where mutating a *SH2B1* homologue leads to extreme obesity (Ren, Li et al. 2005) and from humans, where a chromosomal deletion encompassing *SH2B1* associates with severe early-onset obesity (Bochukova, Huang et al. 2010) strengthen the hypothesis that the missense SNP is the actual functional variant. This SNP is then most probably in high LD with a different causal regulatory variant, which affects *EIF3C* and *TUFM* expression. This is a typical example of a coincidental overlap of GWAS and eQTL results, which must be carefully distinguished from causal cases where both the GWAS SNP and the eQTL tag the same functional variant (Figure 1.7).

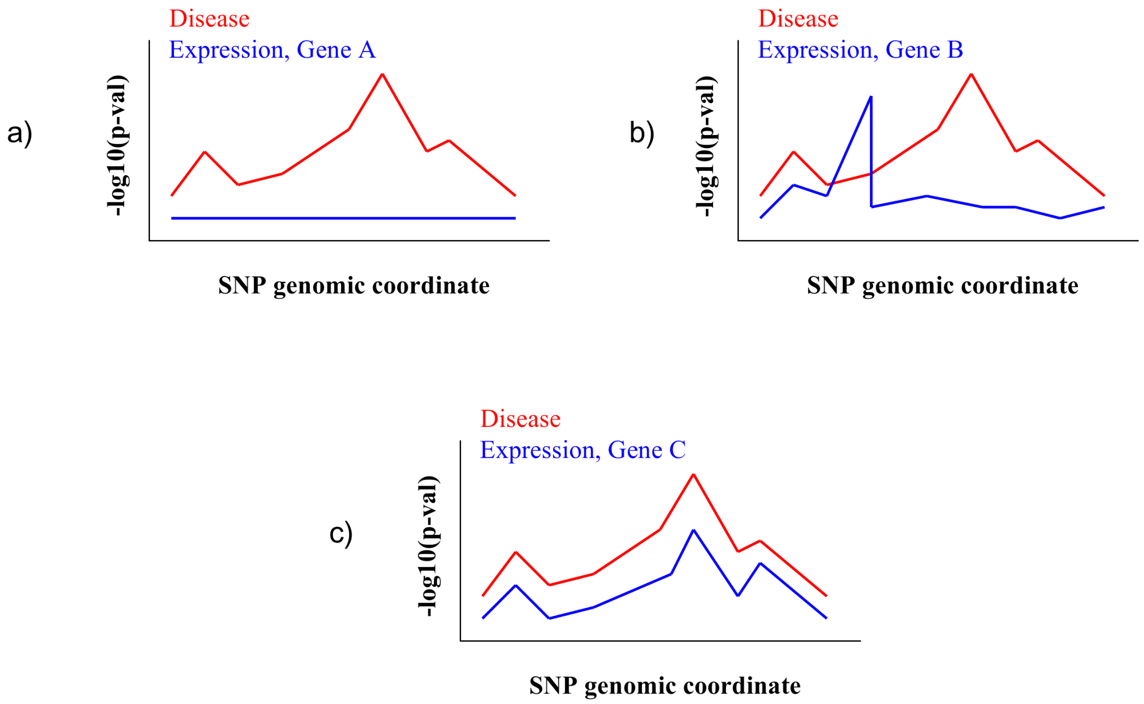


Figure 1.7. Interpreting GWAS results with eQTL data. Schematic representation of a genomic interval where same SNPs have been tested independently for associations with a disease (red) and transcript levels of a set of genes (blue). Three nearby genes are investigated for potential causal regulatory effects: (a) Gene A can be ruled out as it has no significant eQTL in the interval (b) The disease - associated interval harbours an eQTL for gene B, but the eQTL and the disease marker tag different functional variants (c) The GWAS SNP is a strong eQTL for Gene C and they likely tag the same functional effect as reflected by the similar association patterns at other tagging SNPs in the interval (Nica and Dermitzakis 2008).

Given the ubiquitous nature of regulatory variants (Stranger, Nica et al. 2007) and hence the high probability of such coincidental overlaps, integrative methods pinpointing true causal regulatory effects are desirable (Nica and Dermitzakis 2008). Nevertheless, since many traits manifest themselves only in certain tissues, such methods are only informative if expression measurements from disease-relevant cell-types are compared. Defining disease relevance of a cell-type is however yet another challenging task. The pathology of diseases is largely tissue-specific, but it remains mostly unknown how tissue-wide germline mutations lead to tissue-restricted disease effects (Lage, Hansen et al. 2008). Moreover, substantial overlap has been observed between pathways involved in progression of different diseases (Bentires-Alj, Kontaridis et al. 2006) and sometimes this overlap is not intuitive (Swanberg, Lidman et al. 2005; Torkamani, Topol et al. 2008). Therefore, confidently defining tissue relevance to a complex trait is yet unrealistic and expression datasets from seemingly irrelevant tissues should not be discarded at this stage, as they could be informative of disease biology.

1.6.2 Gene regulatory networks

The large-scale disease studies performed so far have uncovered multiple variants of small effect sizes affecting multiple genes. This suggests that common forms of disease are most probably not the result of single gene changes with a single outcome, but rather the outcome of perturbations of gene networks which are affected by complex genetic and environmental interactions (Schadt 2009). The numerous genetic factors involved in disease predisposition appear randomly distributed across the genome, but the expectation is that they are functionally linked and that these functional interactions are useful in prioritizing disease genes (Franke, van Bakel et al. 2006). DNA sequencing of tumour samples from pancreatic and brain cancer respectively, provided supporting evidence for this principle by identifying candidate genes belonging largely to core pathways involved in tumorigenisis or tumour progression (Jones, Zhang et al. 2008; Parsons, Jones et al. 2008). Recently, analysis of gene regulatory networks has offered important insight into complex disease mechanisms. In a study integrating co-expression networks and genotypic data from an F2 intercross population, Chen etal. identified a liver and adipose macrophage-enriched sub-network (MEMN) associated with metabolic syndrome relevant traits (Chen, Zhu et al. 2008). Three genes in this network, lipoprotein lipase (*Lpl*), lactamase β (*Lactb*) and protein phosphatase 1-like (*Ppm1l*) were validated by gene knockouts as causal obesity genes, strengthening the association of MEMN to phenotypes characteristic to metabolic syndrome. A parallel study in humans identified a homologous transcriptional network constructed from adipose data, having substantial overlap with MEMN sub-modules and being enriched for genes involved in inflammatory and immune response (Emilsson, Thorleifsson et al. 2008). Subsequent eQTL mapping identified *cis*-regulatory variants affecting specific genes in this network and the joint analysis of the strongest *cis* eQTLs revealed substantial enrichment for variants associated to obesity related clinical traits. Classical genetic approaches would not be able to detect such variants with small individual effects. Identifying them as a group affecting gene networks which - when perturbed - result in a disease state, is in this case much better powered.

1.6.3 Candidate gene approach via transcriptome profiling

The major challenge when using transcriptome data for interpreting disease effects is distinguishing between causal and reactive changes in gene expression. An interesting approach to address this issue has been taken recently by Naukkarinen and colleagues, who use it to detect potential obesity candidate genes (Naukkarinen, Surakka et al. 2010). The authors made use of genome-wide expression data from adipose tissue and a unique collection of samples (a set of 13 monozygotic - MZ - twin pairs discordant for BMI) in order to devise an original candidate gene prioritization strategy. An additional cohort of 77 non-related individuals having a wide and representative BMI range had been profiled for adipose expression. The authors compared significant expression differences between lean and obese individuals in the MZ twins and the separate cohort, the rationale being that expression differences in the genetically identical twins represent likely reactive effects to obesity, whereas expression differences in the non-related individuals are a combination of causal and reactive determinants. The difference of the two sets would constitute a plausible collection of genes causally implicated in obesity risk. Variants in these genes (197 SNPs in 27 genes) showed a significant excess of low P-values when tested for association with BMI in a large cohort. Among the top associated SNPs, seven mapped to the same gene, F13A1 (coagulation factor XIII, A1 polypeptide), a newly proposed obesity susceptibility candidate. Variants in this gene were replicated in another independent cohort of ~2,000 samples, yet further validation of the associations with BMI is still required in larger independent cohorts. The choice of the candidate cell-type for expression quantification is an obvious issue. The authors acknowledge that for example, MC4R, a known obesity gene has been excluded, as it is not expressed in fat. Given the poor candidate tissue knowledge, choosing a relevant tissue for such experiments is currently very challenging for many traits. Furthermore, identical twins discordant for a phenotype of interest, while very informative, are a rare sample collection. Nevertheless, this informed candidate gene strategy is an interesting example of how to identify further disease variants when limited by the sample sizes available. For many complex traits, identifying additional genetic associations beyond the initial ‘low-hanging fruits’ requires sampling of a vast set of individuals. For example, the six new loci associated with body mass index (BMI) recently reported by the GIANT consortium involved the analysis of more than 91,000 samples (Willer, Speliotes et al. 2009). Similarly large or larger sample sizes are unrealistic for other traits and novel strategies like the example presented here could be helpful for uncovering smaller genetic effects.

1.7 Thesis aims

The genetics of global gene expression has been extensively studied in recent years and it is now unquestionable that regulatory variants affecting transcript levels are ubiquitously distributed throughout the human genome. Concurrently, large-scale GWAS have shed light into the genetics of human complex traits and identified a multitude of susceptibility loci of modest effect-size each. Despite the statistical success in revealing DNA variant - trait associations, by themselves, these results alone don’t necessarily lead to the identification of causal disease mechanisms. The goal of my thesis is to further the understanding of regulatory variation particularly with a focus on its role in complex diseases. Specifically, I address this by: (a) developing an empirical methodology (RTC) that directly combines eQTL and GWAS results in order to detect causal regulatory effects and prioritize candidate disease genes (Chapter 3) (b) exploring the complexity of regulatory tissue-specificity in multiple primary tissues derived from a set of twins (Chapter 4) (c) analyzing the implications of tissue-dependency in detecting causal regulatory effects for complex traits (Chapter 5). Taken together, the results presented here underline the informative value of expression phenotypes for explaining the biological properties behind genetic associations with complex traits and highlight the need to explore regulatory complexity in a variety of relevant cell-types.

2 Materials and methods

2.1 Resources

The data presented and analysed in this thesis has been derived from samples belonging to three major resources, briefly outlined below: the HapMap, MuTHER and GenCord projects. Table 2.1 summarizes the number of available samples, SNPs and transcripts per resource and the respective thesis chapters where they have been analyzed.

Chapter	Resource	Samples by Tissue	SNPs	Mapped Probes	Mapped Genes
3	HapMap 3 (CEU)	109 LCL	1,186,075	21,800	18,226
4	MuTHER	156 LCL, 160 SKIN, 166 FAT	865,544	27,499	18,170
5	GenCord	75 LCL, 75 fibroblasts, 75 T-cells	1,428,314	26,651	17,945

Table 2.1. Summary of resources (samples, SNPs and transcripts) used throughout the thesis.

2.1.1 HapMap

The International HapMap project is a large-scale collaboration launched in 2002 to identify and catalogue common human genetic variation (Consortium 2003). DNA from LCLs derived from individuals of different population ancestry has been genotyped in an attempt to discover the vast majority of common human SNPs (MAF ≥ 5%). HapMap 3, the current and largest phase of the project (http://hapmap.ncbi.nlm.nih.gov/) is comprised of over 4 million SNPs genotyped from individuals of the Phase 1 and 2 populations (180 CEU, 90 CHB, 91 JPT, 180 YRI) and approximately 1.5 million SNPs genotyped in 760 individuals of seven new populations (90 ASW, 100 CHD, 100 GIH, 100 LWK, 90 MEX, 180 MKK, 100 TSI).

In this thesis, I analysed data from the subset of unrelated HapMap 3 CEU individuals (N=109) in the study described in Chapter 3.

2.1.2 MuTHER

The MuTHER (Multiple Tissue Human Expression Resource) project was funded by the Wellcome Trust in 2007 as a coordinated program of analysis aiming to enhance our knowledge about common trait susceptibility. By generating detailed genetic (genotyping and resequencing) and genomic (mRNA expression, methylation status) information from a range of tissues collected from ~1000 twins, the MuTHER project will constitute a major resource for understanding the relationships between sequence variation and disease phenotypes (http://www.muther.ac.uk/).

LCLs, fresh lymphocytes, fat, muscle and skin biopsies have been obtained from a maximum of 855 twins (318 monozygotic, 537 dizygotic) from the well-characterised Twins UK Resource (Spector and Williams 2006). This sample of volunteers was recruited by media campaigns without selecting for particular diseases or traits. All twins received a series of detailed disease and environmental questionnaires and the majority of individuals have been clinically assessed at several time points for hundreds of phenotypes related to common diseases or intermediate traits. All individuals recruited in this study were Caucasian female twins aged between 39 and 70 years old.

At the time of writing, whole-genome genotyping and expression profiling of the full set of 855 twins was underway. A sample subset representing the pilot phase of the MuTHER project had been profiled in advance in three tissues: LCL, skin and fat. Skin punch biopsies (N=196) were taken from a relatively photo-protected area adjacent and inferior to the umbilicus. The fat sample was then carefully dissected from the same skin biopsy incision. A peripheral blood sample to generate lymphoblastoid cell lines (LCL) was taken contemporaneously. The biopsies were performed by Daniel Glass at KCL following the technique steps described in Appendix 1.

Chapter 4 describes the analysis I performed on the MuTHER pilot project data.

2.1.3 GenCord

The GenCord project was initiated at the University of Geneva Hospital and consists of a collection of cell lines derived from the umbilical cords of 85 individuals of Western European origin. The primary goal of the project was to serve as a resource facilitating discovery and comparison of eQTLs across multiple tissues while controlling for confounding factors such as different population samples or differences in technological and statistical methods employed. Umbilical cord was chosen due to its accessibility and the potential of harvesting multiple tissues from the same sample. Following appropriate consent and ethical approval (Dimas, Deutsch et al. 2009), cord blood and cord tissue was obtained per each sample in order to derive three cell-types: primary fibroblasts, EBV-immortalized lymphoblastoid cell lines (LCL) and primary T-cells. All pregnancies were full or near full term (38-41 week) ensuring age homogeneity of the samples.

GenCord LCL data was used in the control experiment I describe in Chapter 3. GenCord data from LCLs, fibroblasts and T-cells was used in the analysis I present in Chapter 5.

2.2 SNP genotyping

Genetic variation data (SNP genotypes) from HapMap 3, MuTHER and GenCord has been analysed throughout the course of my PhD, primarily to identify associations with gene expression variation (eQTL discovery, section 2.4).

SNP detection has been performed mostly on Illumina's whole-genome genotyping platforms using the Infinium HD technology. This enables dense, uniform genome coverage by typing a representative set of tag SNPs. The Infinium II assay workflow is described in Figure 2.1.

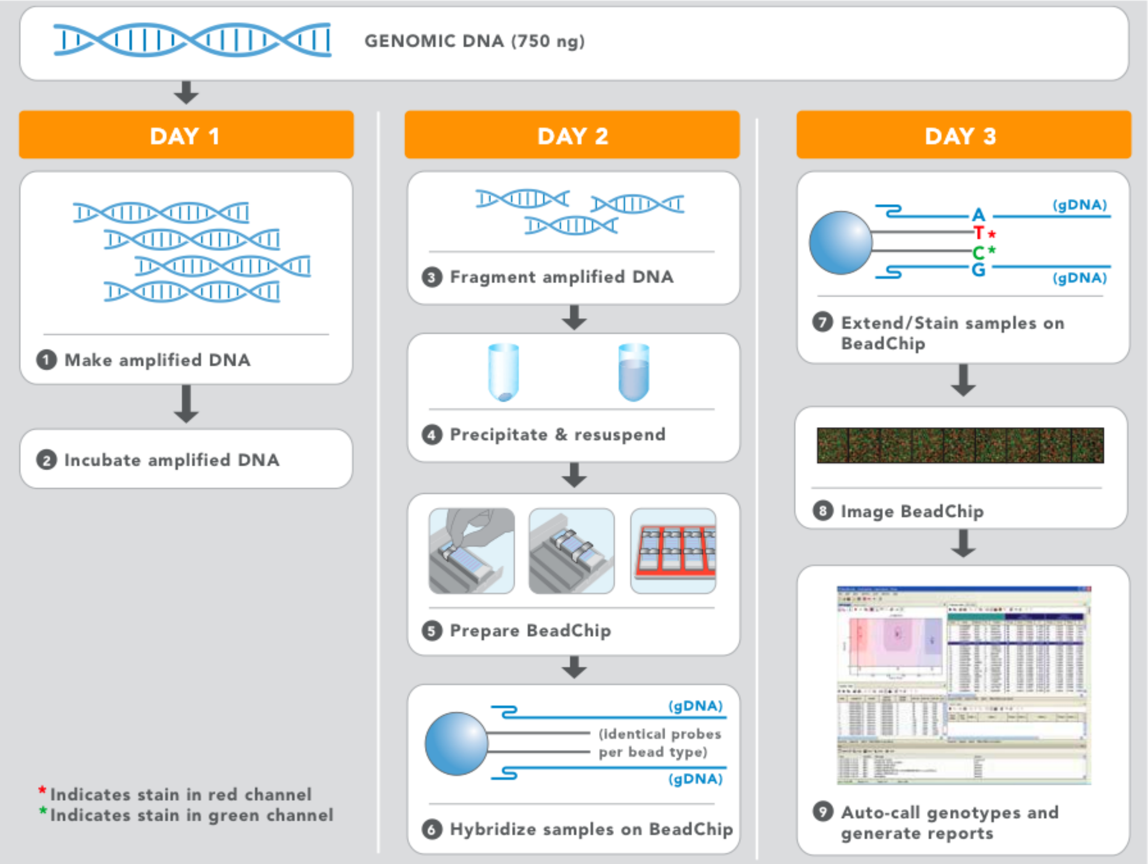

GENOMIC DNA (750 ng)
DAY 1
DAY 2
DAY 3
1 Make amplified DNA
2 Incubate amplified DNA
3 Fragment amplified DNA
4 Precipitate & resuspend
5 Prepare BeadChip
(gDNA)
(identical probes per bead type)
(gDNA)
6 Hybridize samples on BeadChip
(gDNA)
A
T*
C*
G
(gDNA)
7 Extend/Stain samples on BeadChip
8 Image BeadChip
9 Auto-call genotypes and generate reports
*Indicates stain in red channel
*Indicates stain in green channel

Figure 2.1. Illumina II assay protocol. The Infinium II whole-genome genotyping assay uses a single bead type and dual colour channel approach. During Step 1 and Step 2, a DNA sample of relatively low required quantity (750 ng suffice for assaying 500,000 SNPs) is amplified and incubated overnight. The amplification has no appreciable allelic partiality. Following the amplification, the product is fragmented in an enzymatic process (Step 3). After precipitating and resuspending the DNA (Step 4), the BeadChip is prepared for hybridization (Step 5). The DNA samples are applied onto the BeadChips and incubated overnight, thus allowing the fragmented DNA to hybridize to locus-specific 50-mers on the chips which are covalently linked to one of the > 500,000 chip bead types (Step 6). One bead corresponds to each allele per SNP locus. After hybridization, an enzymatic base extension process ensures allelic specificity and the products are subsequently fluorescently stained (Step 7). Finally, the BeadArray Reader (Step 8) detects the fluorescence bead intensities, which are in turn analyzed by calling algorithms and translated into genotypic information (Step 9). Figure and assay protocol description from www.illumina.com

HapMap

HapMap genotypes have been generated by the International HapMap Consortium and are publicly available on the HapMap website (http://hapmap.ncbi.nlm.nih.gov/). The release used in this thesis (HapMap version 27, NCBI Build 36) contains SNP genotype data generated from 1,301 HapMap 3 samples collected using two platforms: the Illumina Human1M (by the WTSI) and the Affymetrix SNP 6.0 (Broad Institute). Data from the two platforms have been merged and the subset of SNPs passing the following QC criteria kept: 1) Hardy-Weinberg p-value > 10^{-6} per population; 2) genotype missingness < 0.05 per population; 3) <3 Mendel errors per population; 4) SNP must have an rsID and map to a unique genomic location. For the analysis presented in Chapter 3 I have used all common (MAF ≥ 5%) autosomal SNPs from the unrelated CEU HapMap 3 individuals (N=109). This dataset amounts to 1,186,075 SNPs.

MuTHER

The pilot MuTHER samples have been genotyped at WTSI using in parallel Illumina's 1M-Duo and 1.2M-Duo custom chips on different subsets of individuals. Before further filtering, there were 106 samples with call rate (CR) ≥ 0.90 on the 1.2M and 88 samples with CR ≥ 0.90 on the 1M chip. Combined intensity files were created for Illuminus (Teo, Inouye et al. 2007) by retaining on a per-chromosome basis only SNPs common to both chips. Additionally, any SNPs that moved position between the two chips were removed. Following further quality checks (Hardy- Weinberg p > 10^{-4}, MAF > 1%), 865,544 SNPs were kept for analysis. The QC analysis was performed by Simon Potter at WTSI.

The set of successfully genotyped samples was overlapped with individuals having corresponding expression data available. This amounted to the following sample set per tissue: 156 LCL, 160 skin and 166 fat individuals (Chapter 4).

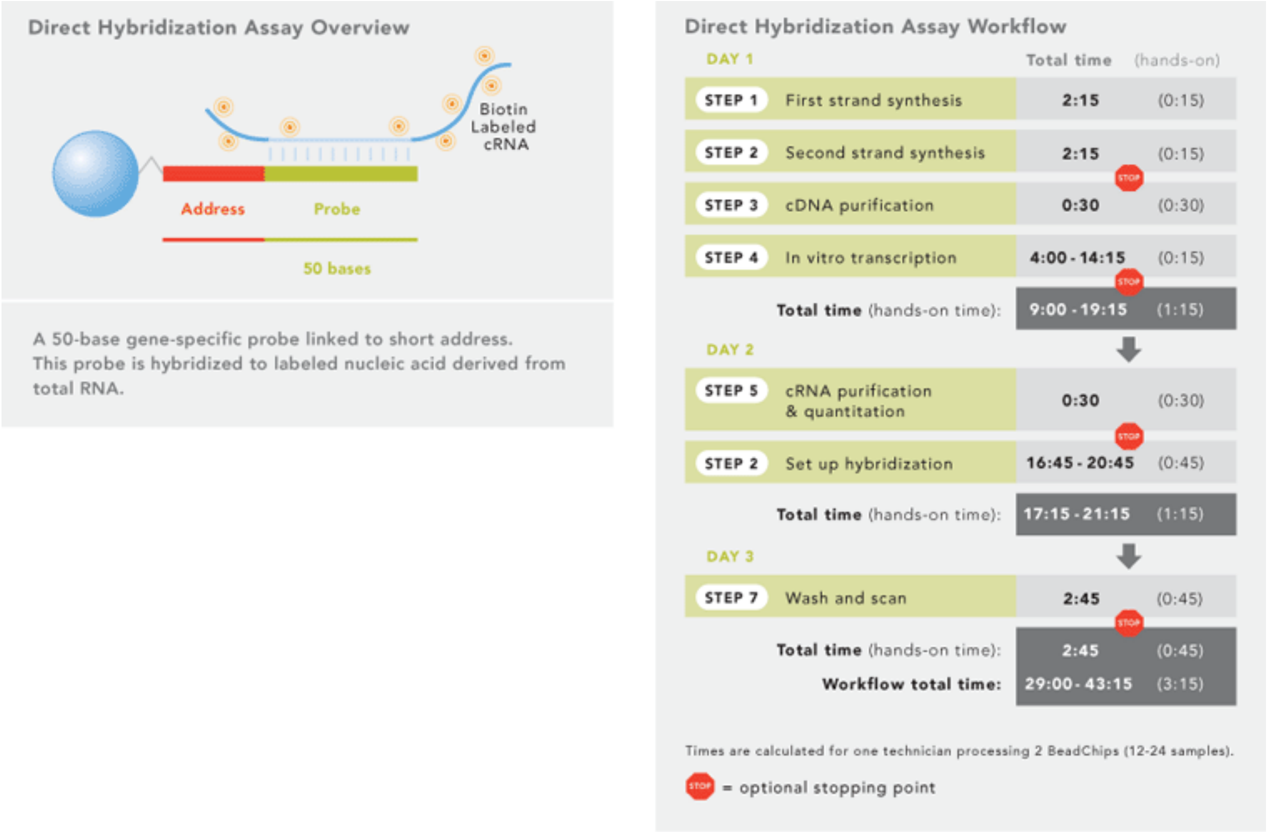
GenCord

The 85 GenCord individuals were genotyped for approximately half a million SNPs each using Illumina's 550K SNP array. DNA samples were extracted from cord tissue LCLs with the Puregene cell kit (Gentra-Qiagen, Venlo, The Netherlands). This work was carried out by Samuel Deutsch and colleagues in Stylianos Antonarakis' lab at UGMS. Principal component analysis (PCA) was performed on the genotype data to detect potential outliers. Following this analysis performed by Stephen Montgomery at the WTSI, ten individuals were removed. After further QC analysis (removing SNPs with missing data), 394,651 SNPs with MAF ≥ 5% were kept for analysis (Chapter 3).
To increase the power to detect associations with expression, GenCord genotypes were imputed onto the reference HapMap 2 data using the BEAGLE software (Browning and Browning 2007). Following imputation, QC was performed whereby SNPs with imputation quality scores < 0.9 (24,7078 SNPs) and those failing MAF (<5%) or Hardy-Weinberg equilibrium checks (total of 67,718 SNPs) were removed. This work was performed at UGMS by Eugenia Migliavacca (imputation) and Tuuli Lappalainen (QC). A final set of 1,428,314 SNPs in 75 individuals was used for the analysis In Chapter 5.

2.3 Gene expression quantification

Transcript levels in HapMap (LCL), GenCord (LCLs, fibroblasts, T cells) and MuTHER (LCL, skin, fat) samples were quantified at WTSI using Illumina's whole-genome gene expression arrays. HapMap and GenCord data are also publicly available at http://www.sanger.ac.uk/resources/software/genevar/.

Whole-genome expression profiling is based on the direct hybridization technology developed by Illumina (Figure 2.2).



A 50-base gene-specific probe linked to short address. This probe is hybridized to labeled nucleic acid derived from total RNA.

Direct Hybridization Assay Workflow

DAY 1		Total time	(hands-on)
STEP 1	First strand synthesis	2:15	(0:15)
STEP 2	Second strand synthesis	2:15	(0:15)
STEP 3	cDNA purification	0:30	(0:30)
STEP 4	In vitro transcription	4:00 - 14:15	(0:15)
	Total time (hands-on time):	9:00 - 19:15	(1:15)
DAY 2			
STEP 5	cRNA purification & quantitation	0:30	(0:30)
STEP 2	Set up hybridization	16:45 - 20:45	(0:45)
	Total time (hands-on time):	17:15 - 21:15	(1:15)
DAY 3			
STEP 7	Wash and scan	2:45	(0:45)
	Total time (hands-on time):	2:45	(0:45)
	Workflow total time:	29:00 - 43:15	(3:15)

Times are calculated for one technician processing 2 BeadChips (12-24 samples).

STOP = optional stopping point

Figure 2.2. Direct hybridization assay overview and workflow. Figure from www.illumina.com

The protocol features first the amplification of the starting RNA material via first- and second-strand reverse transcription, followed by a single in vitro transcription (IVT) amplification that incorporates biotin-labelled nucleotides. The resulting cRNA is purified, hybridized to the array and labelled with Cy3-streptavidin (Amersham Biosciences, Little Chalfont, UK). The fluorescence emission by Cy3 is scanned and quantified with Bead Station (Illumina).

More than 48,000 unique bead types (one for each of the 47,294 transcripts plus controls) are represented on the array. Each bead contains several hundred thousand copies of gene-specific 50mer probes covalently attached. The probes are derived from the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) and UniGene databases. The beads are assembled into 3 µm diameter wells, generating an average 30-fold redundant information for each probe. These background-corrected values for a single bead type are summarized by Bead Studio (Illumina software) and outputted to the user as a set of 47,294 intensity values for each individual hybridization.

HapMap

Total RNA was extracted from LCLs derived from the HapMap 3 individuals (Coriell). Gene expression was quantified using Illumina's commercial array Sentrix Human-6 Expression BeadChip version 2. For each RNA extraction, two one-quarter scale Message Amp II reactions (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 were hybridized to the whole-genome expression array. Six arrays were run in parallel on each individual BeadChip. The experimental work was carried out by Catherine Ingle, James Nisbet and Magdalena Sekowska at the WTSI.

To combine information from the two replicate hybridizations, raw data was normalized on a $\log_2$ scale by quantile normalization (Bolstad, Irizarry et al. 2003) across replicates of a single individual followed by median normalization across all individuals from a single population. Normalization was performed by Stephen Montgomery at WTSI.

Of the >48,000 probes represented on the array, only a trustable subset was chosen for further analysis. The Sentrix Human-6 Expression BeadChip version 2 array covers over 24,000 unique, curated RefSeq genes, as well as genes with less well-established annotation. Only probes corresponding to well-annotated RefSeq genes were kept at this point. Additionally, probes were matched to corresponding Ensembl genes (Ensembl 49 NCBI Build 36) using SSAHA (Sequence Search and Alignment by Hashing Algorithm) (Ning, Cox et al. 2001). Following the SSAHA run 22,512 probes were mapped to 19,862 Ensembl genes. Probes mapping to multiple Ensembl genes were removed, as well as ones mapping to sex chromosomes. After filtering, a non-redundant set of 21,800 probes (corresponding to 18,226 Ensembl genes) was used for association analysis. Mapping and selection of probes for final analysis was carried out by Antigone Dimas at WTSI.

MuTHER

RNA was extracted from LCLs, skin and fat samples derived from the pilot MuTHER individuals. Gene expression was measured using Illumina's HumanHT-12 version 3 whole-genome array, as explained previously (in this case, each sample had three technical replicates). The experimental work was carried out by James Nisbet and Magdalena Sekowska at WTSI and by Amy Barrett and Mary Travers at WTCHG.

Log_2-transformed expression signals were normalized separately per tissue as follows: quantile normalization was performed across the 3 replicates of each individual followed by quantile normalization across all individuals.

The >48,000 probes targeting more than 25,000 genes are derived from RefSeq (Build 36.2, Rel 22) and UniGene (Build 199). To select probes corresponding to well-annotated genes, Illumina's v3 probes were mapped to unique Ensembl gene IDs by combining and cross-checking two methods. The first approach used probe annotations to RefSeq IDs provided by Illumina, which were further queried with BioMart (Ensembl 54) for corresponding Ensembl genes IDs. RefSeq IDs mapping to multiple Ensembl Genes were excluded, and only autosomal genes retained. This step was performed with the help of Tsun-Po Yang at WTSI. The second approach used BLAT (Kent 2002) to map the 50-mer probe sequences to Ensembl transcripts and to extract genomic locations matching all 50 bases of the probe sequence. Probes with unique perfect match to the genome and corresponding transcripts matching to the same genes were kept. This approach was performed by Josine Min at WTCHG. The union of the two mapping approaches after excluding 196 conflictingly matching probes resulted in 27,499 probes corresponding to 18,170 autosomal genes available for association analysis.

GenCord

Total RNA was extracted from LCLs, fibroblasts and T-cells of the 85 GenCord individuals. 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. Each RNA sample had two technical replicates. This work was carried out by Catherine Ingle, James Nisbet, and Magdalena Sekowska at the WTSI.

The expression raw data was normalized independently for each cell type as follows: the intensity values were $\log_2$ transformed, quantile normalized per sample replicates and median normalized 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.

The WG-6 v3 Expression BeadChip array covers over 27,000 unique coding transcripts. For some of them, well-established annotation exists (7,000 transcripts have provisional annotation). In addition, the array covers non-coding transcripts, as well as experimentally confirmed mRNA sequences aligning to EST clusters. Again, only probes with good or provisional annotation (mapping to RefSeq genes) were selected of the total 48,000 probe set 36,156 probes with Refseq IDs were queried for their corresponding Ensembl gene IDs in Biomart (Ensembl 50, NCBI Build 36). Of these, 22,651 probes had a uniquely assigned Ensembl gene ID and did not map to either chromosomes X or Y. These probes corresponding to 17,945 RefSeq genes and 15,596 Ensembl genes respectively were used for subsequent analysis. Selection of the final probe list was done by Antigone Dimas at WTSI.

2.4 eQTL discovery

Associations between SNP genotypes and normalized expression values were run using Spearman Rank Correlation (SRC) and additive linear regression (LR). SRC was exclusively used to detect eQTLs (Chapter 4) while LR was used to quantify the proportion of expression variance unexplained by the SNP genotypic classes (Chapter 3, Chapter 5). I considered SNPs within a 1Mb window on either side of a gene’s transcription start site (TSS) as *cis*-acting while SNPs located further than 5 Mb away either side of a gene’s TSS or SNP-gene pairs on different chromosomes as *trans*-acting.

2.4.1 Association analysis

Before association, the SNP genotypes were numerically encoded (0, 1 or 2) to represent the counts of alphabetically sorted alleles at each locus (e.g. counting the number of G alleles for an A/G SNP: AA = 0, AG = 1, GG = 2) (Figure 2.3).

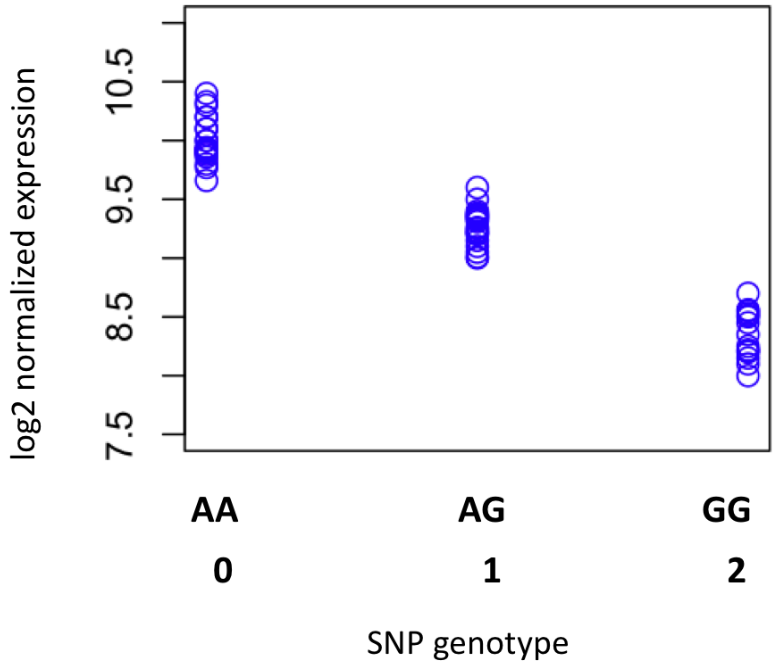


Figure 2.3. SNP-gene association example. The A/G SNP in this schematic example is plotted against a gene's corresponding normalized $\log_2$ expression values. In this case, the A allele at the SNP locus predisposes individuals to have higher expression values of the respective gene.

2.4.1.1 Spearman Rank Correlation (SRC)

SRC is a non-parametric test assessing the degree of statistical dependence between two variables (X and Y). A monotonic function is fitted to describe the correlation between X and Y (e.g. X = genotype, Y = expression). No other assumption is made about the relationship between the two variables, which are rank-ordered. In our case for example, expression values are ordered low to high and ranked accordingly (1..n), irrespective of their actual numerical value. This makes sure that outliers do not have a high impact on estimating the correlation between X and Y. The degree and direction of this correlation is reflected in the ρ (rho) coefficient, calculated as below, where *n* is the number of observations and d_i is the difference between the ranks of each observation on the two variables ($d_i = x_i - y_i$):

$$\rho = 1 - \frac{6\sum d_i^{\,2}}{n(n^2 - 1)}$$

When two observations for the same variable are equal (tied), they are each assigned the average corresponding rank. A perfect Spearman correlation ($\rho = 1$ or $\rho = -1$) occurs when each of the variables is a perfect monotone function of the other. The sign marks the direction of the correlation: $\rho > 0$ (positive correlation) if Y tends to increase when X increases and $\rho < 0$ (negative correlation) if Y tends to decrease when X increased. A nominal p-value for the association test is also reported.

2.4.1.2 Additive linear regression (LR)

In a LR model, the relationship between two variables is explored by fitting a linear equation to the observed values. For the work presented in this thesis, the following main effects additive model was used to test for SNP-gene expression associations:

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

Here, the dependent variable Y_i is a probe's normalized $\log_2$ expression value quantified in individual *i* (*i* = 1..n) and the explanatory variable X*i* is the corresponding numerically encoded genotype. ε_i are independent normally distributed random variables with mean 0 and constant variance (Stranger, Forrest et al. 2005). b_i is the slope of the fitted regression line ($b_i = 0$ if there is no association between the genotype and the expression values). How well the regression model fits the data can be estimated from the inspection of the residuals i.e. the vertical distances of each point from the regression line. The residuals quantify the proportion of the variance in the dependent variable (Y - expression) that cannot be accounted for by the explanatory variable (X - genotype). As such, the most common regression technique employs minimizing the sum of squared residuals.

2.4.2 Multiple testing correction

The statistical significance of associations between SNP genotypes and gene expression levels was assessed using permutations (Churchill and Doerge 1994; Doerge and Churchill 1996). The $\log_2$ normalized expression values of each probe were permuted 10,000 times relative to the genotypes of the SNPs in the tested window (2MB in *cis*). The minimal p-value association of each run was retained generating thus a distribution of 10,000 values corresponding to the best random SNP-probe associations. Significance was assessed for different threshold levels (0.5, 0.01, 0.001 and 0.0001) by comparing the tail of the distribution of the 10,000 minimal p-values for each gene to the observed association p-value (e.g. an association was considered significant at the 0.0001 threshold if the nominal observed p-value was lower than the 0.0001 tail of the distribution of minimal permuted p-values) (Stranger, Forrest et al. 2005; Stranger, Nica et al. 2007).

2.5 Recombination hotspot mapping and LD filtering

To restrict the search space for causal regulatory effects and refine eQTL signals, I have made use of the genome's correlation structure (LD). Specifically, I used recombination hotspot coordinates derived from the statistical analysis of the variation data generated by the HapMap 2 project (Release 22, Build 36) (McVean, Myers et al. 2004) (Myers, Bottolo et al. 2005). The recombination hotspots inferred are typically 1-2 kb long and are surrounded by much larger regions (defined here as recombination hotspot intervals) essentially devoid of recombination (Paigen and Petkov 2010). All autosomal SNPs in HapMap 3 CEU, MuTHER and GenCord have been mapped to recombination hotspots and hotspot intervals. The mapping serves both to restrict the search for functional regulatory variants explaining GWAS signals (Chapter 3, Chapter 5) and also for refining eQTL signals by identifying independent regulatory effects and comparing them across multiple tissues (Chapter 4).

In Chapters 3 and 5, GWAS results are tested for explanatory regulatory effects. For this purpose, given any GWAS SNP, I focus on the recombination hotspot interval where it resides and where also at least one eQTL co-localizes. Limiting the search space for causal effects to these intervals with independent recombination history is a reasonable approach, as few or no recombination events are expected between the reported associated SNPs and the functional variants they are tagging.

In Chapter 4, I aim to characterize in detail the landscape of regulatory variation across LCLs, skin and fat. For this reason, I refine the discovered eQTL signals to likely independent effects per gene. The strategy employed is the following: after mapping significant eQTLs to recombination hotspot intervals, the most significant SNP per gene per interval is kept. Furthermore, to avoid long-range correlations which can extend over recombination hotspots, an additional LD filtering step is performed so that for each pair of significant eQTLs with D' > 0.5, the least significant SNP is ignored. The choice of D' over r^2 as LD filtering metric is based on their distinctive properties. Both metrics relate to D, the basic unit of LD measuring the deviation of haplotype frequencies from equilibrium state (Lewontin and Dunn 1960). For two SNPs with alleles (A,a) and (B,b) respectively:

$$D = f(AB) - f(A)f(B)$$

where f(X) is the frequency of the X allele. If D is significantly different from 0, LD occurs. D', calculated as below, ranges from 0 to 1, with D' =1 denoting complete LD while values towards 0 indicating linkage equilibrium, i.e. historical genetic independence.

$$\begin{aligned} if D \geq 0, \quad D' &= \frac{D}{D_{\max}} \\ if D < 0, \quad D' &= \frac{D}{D_{\min}} \end{aligned}$$

r^2 is the statistical coefficient of determination, or the measure of correlation between a pair of variables (SNP genotypic classes in this case). Also in the range of 0 to 1, $r^2 = 1$ indicates that one SNP is directly predictive of the other (perfect correlation) and lower values denote the decay of their correlation (r^2 approaches 0) (Wang, Barratt et al. 2005).

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}$$

While r^2 quantifies the statistical correlation between two variants, D' is a measure of their historical relationship which is biologically more meaningful. For example, two correlated SNPs in between which no recombination event occurred (D' =1) but which have different MAFs (low r^2) can be tagging the same functional effect (e.g. a single independent regulatory variant residing in the respective hotspot interval). The stringent D' threshold (which corresponds to an even lower r^2) provides thus a more suitable method to filter for historically independent effects. When comparing across tissues, this filtering ensures that true shared effects (interval-gene combinations) are contrasted and not just genes, which would be inaccurate in cases when the same gene is regulated by different functional variants in different tissues.

2.6 RTC scoring scheme (Chapter 3, Chapter 5)

The Regulatory Trait Concordance (RTC) method was developed in order to detect the subset of GWAS signals which could be explained by significant regulatory effects and identify the genes whose expression levels they mediate. For this purpose, I used expression data from two resources: HapMap 3 and GenCord. The whole-genome expression quantification experiments on the MuTHER pilot samples were performed towards the end of my PhD and were not available for analysis at that time. eQTLs discovered in LCLs derived from HapMap 3 CEU and GenCord individuals were tested in Chapter 3, while eQTLs detected in the three GenCord tissues (LCLs, fibroblasts, T-cells) were overlaid with GWAS results in Chapter 5. I next describe the RTC method and the main experiments it has been used for.

2.6.1 Method overview

I assess the likelihood of a shared functional effect between a GWAS SNP and an eQTL by quantifying the change in the statistical significance of the eQTL after correcting for the genetic effect of the GWAS SNP. The correction is performed using a LR model. The GWAS SNP is first regressed against normalized expression values of the gene for which an eQTL exists. The residuals capture the remaining unexplained expression variance after the removal (correction) of the GWAS SNP effect. This resulting pseudo phenotype is used to redo the SRC association with the eQTL genotype. It is expected that if the GWAS SNP mediates the disease effect through a change in gene expression due to a regulatory variant (eQTL) then correcting out the GWAS SNP effect will have a marked consequence on the eQTL i.e. the eQTL SNP – gene association p-value after correction will be much less significant than the association p-value before correction. The p-value estimates however, are affected also by the LD structure of the investigated region: the correlation between the eQTL and the GWAS SNP but also between each of the two and the actual functional variants (most often unknown) influence the correction outcome. Given that part of the change in the p-values will be attributed to LD, it is necessary to account for this correlation in each interval of interest.

I account for the LD structure in each hotspot interval separately by ranking (Rank $_{\text{GWAS SNP}}$) the impact on the eQTL (quantified by the adjusted association P-value after correction) of the GWAS SNP correction to that of correcting for all other SNPs in the same interval. The rank denotes the number of SNPs which when used to correct the expression data, have a higher impact on the eQTL (less significant adjusted P-value) than the GWAS SNP (i.e. Rank$_{GWAS\ SNP}$ = 0 if the GWAS SNP is the same as the eQTL SNP, Rank$_{GWAS\ SNP}$ = 1 if of all the SNPs in the interval, the GWAS SNP has the largest impact on the eQTL etc). By taking into account the total number of SNPs in the interval (N$_{\text{SNPs}}$), this ranking can be compared across different genes and intervals. For this purpose, the RTC score is defined as follows:

$$RTC = \frac{N_{SNPs} - Rank_{GWAS\ SNP}}{N_{SNPs}}$$

The RTC score ranges from 0 to 1, with values closer to 1 indicating causal regulatory effects. The highest RTC statistic (RTC = 1) is obtained for the lowest correction ranking (Rank$_{GWAS\ SNP}$ = 0) corresponding to cases when the GWAS SNP is identical to the eQTL. As expected in these instances, correcting the eQTL SNP with itself removes the largest possible amount of variance, more so than with any other SNP in the region. Cases when the eQTL and GWAS SNP are identical are impossible to resolve with the RTC or any other method. They are however still informative, indicating that the pattern of association between the SNPs in that region and the disease phenotype and gene expression respectively are identical.

2.6.2 RTC properties under simulations

Before applying it to large-scale expression datasets, I investigated the properties and robustness of the RTC score with respect to D' and r^2, the two most common LD metrics. Both possible scenarios were tested: the null hypothesis (H_0) when a GWAS disease SNP (dSNP) and a co-localizing eQTL would tag two different causal variants and the alternative hypothesis (H_1) when the eQTL and dSNP tag the same functional variant. For this purpose, I have simulated causal SNPs (cSNP), eQTLs and dSNPs under different scenarios varying the LD levels between them as well as the LD pattern of the hotspot interval where they reside. The dSNP emulates the most significant trait-associated SNP typically reported by GWAS studies, while the cSNP represents the actual functional variant, unknown most of the times. For each simulated case, the cSNP was first masked, then the RTC was calculated and its performance evaluated. I used the HapMap 3 CEU *cis* eQTLs (315 genes at 10^{-3} permutation threshold) to create the list of cSNPs.

For the H_0 test, the cSNPs were called causal eQTL SNPs (c-eQTLs). For each c-eQTL, I sampled a different causal disease SNP (c-dSNP) from the same recombination hotspot interval, with the requirement that its MAF comes from a distribution identical to that of the GWAS SNPs downloaded from NHGRI (976 GWAS variants) (website accessed 02.03.09). Subsequently, I sampled up to five eQTL-dSNP pairs per interval where the eQTLs and dSNPs are the topmost correlated (r^2) SNPs with the c-eQTL and the c-dSNP respectively. These imitate the typical tagging SNPs reported as having a significant association with gene expression and disease phenotypes respectively. After sampling, I excluded cases where the eQTL and dSNP are identical, as these contradict the H_0. c-eQTL-c-dSNP-eQTL-dSNP quartets mapping to 287 unique hotspot intervals were sampled and tested under H_0. The RTC score was calculated for all simulated eQTL-dSNP pairs in each of the 287 hotspot intervals. The predictive value of the RTC score was compared against standard measures of LD (r^2, D') between the eQTL and the dSNP.

Under the H_1, the cSNP represents the untyped causal variant mediating the disease association via significant changes in gene expression levels. In this case, both the eQTL and the dSNP tag the same effect. Therefore, up to five eQTL-dSNP pairs were sampled for each hotspot interval harbouring a cSNP under H_1 as follows: the eQTLs were chosen as the top most significant SNPs per eQTL gene - excluding the cSNP; the dSNPs were randomly sampled from the same hotspot interval such that the r^2 between each of them and the cSNP was in the range [0.5,0.9]. Perfectly correlated SNPs ($r^2 = 1$) were excluded, as such cases cannot be resolved. In addition, at any stage of the 5-step iteration process per cSNP, the dSNP was selected to be different from the cSNP and the eQTLs sampled up to that point. cSNP-eQTL-dSNP trios mapping to 290 unique hotspot intervals throughout the genome were sampled and tested under the H_1. For all simulated eQTL-dSNP pairs per each hotspot interval (N = 290), the RTC score was calculated and its predictive value compared against the correlation level (r^2, D’) between the eQTL and the dSNP.

Finally, the effect of a region’s overall LD pattern on estimating the RTC score was explored. For this purpose, the extent of LD per hotspot interval was calculated as the median r^2 of all pairwise SNP combinations available per interval. Under both H_0 and H_1, the relationship between the median r^2 of a hotspot interval and the RTC was investigated.

The RTC properties as revealed by these analyses are described in Chapter 3.

2.7 MuTHER eQTL analysis (Chapter 4)

2.7.1 Factor analysis

eQTL analysis on the MuTHER pilot data was performed using the discovery framework presented in Section 2.4 of this chapter (Methods). Additionally, eQTL analysis was conducted after accounting for experimental noise and global environmental conditions, which are also known to impact gene expression in a global manner. For this purpose, a Bayesian factor analysis (FA) model (Stegle, Parts et al. 2010) was applied to the expression data in each tissue. This approach uses an unsupervised linear model to account for global variance components in the data, and yields a residual expression dataset that can be used in further analysis.

A wide range of parameter settings was tested for the model, controlling the amount of variance explained by it. This was achieved by setting the parameters of the prior distributions for gene expression precision (inverse variance) and factor weight precision. These random variables are modelled using Gamma distributions, thus their natural exponential family parameters (the prior mean and number of prior observations) were varied. The prior mean was varied from 10^{-6} to 10^{-2} and the number of prior observations from $N*10^{-3}$ to N, where N is the number of observations from data. 120 latent factors were thus learned. For each tissue, the residual dataset that gave the best eQTL overlap between co-twin samples was used in the subsequent eQTL analyses. The prior values used for each dataset are given in Table 2.2. The FA was developed and carried out by Leopold Parts at WTSI.

	Weight prior		Noise prior	
	Mean	**Observations**	**Mean**	**Observations**
LCL	10^{-6}	23	10^{-3}	10
SKIN	10^{-6}	23	10^{-1}	100
FAT	10^{-6}	23	10^{-3}	10

Table 2.2. Factor analysis weight and noise prior values applied to each tissue. Analysis performed on MuTHER pilot samples.

Following FA, the eQTL analysis on the corrected expression data was performed identically to the original detection strategy: SRC followed by multiple-testing correction using permutations.

2.7.2 Estimation of proportion of true positives (π_1)

Overlapping eQTL discoveries at the same threshold is very sensitive to power, as thresholds are driven by statistical significance. Given this, eQTL replication and tissue sharing was quantified also in a continuous way with Storey's qvalue statistic (Storey and Tibshirani 2003). The QVALUE software implemented in the R package qvalue 1.20.0 was used under the default recommended settings. The program takes a list of p-values and computes their estimated π_0 - the proportion of features that are truly null - based on their distribution (the assumption used is that alternative cases tend to be close to zero, while p-values of null features will be uniformly distributed among [0,1]). The quantity π_1 = 1- π_0 estimates the lower bound of the proportion of truly alternative features, i.e. the proportion of true positives (TP). Replication and sharing between two samples is reported as the proportion of TP (π_1) estimated from the p-value distribution in the second sample of independent eQTLs initially discovered in the first sample (exact snp-probe combinations are used).

3 RTC – empirical method for integrating regulatory variants with complex trait associations

The biological interpretation of the plenitude of GWAS signals (WTCCC 2007; Eeles, Kote-Jarai et al. 2008; Zeggini, Scott et al. 2008) is very challenging since most candidate loci fall either in gene deserts or in regions with many equally plausible causative genes. Following the concurrent progress in understanding the genetic basis of regulatory variation (Cheung, Spielman et al. 2005; Dixon, Liang et al. 2007; Goring, Curran et al. 2007; Stranger, Forrest et al. 2007), differential gene expression has been proposed as a promising intermediate layer of information to aid this interpretation (Emilsson, Thorleifsson et al. 2008). Most commonly, interrogating the GWAS SNPs themselves for significant associations with gene expression has been employed to explain some of the GWAS results (Moffatt, Kabesch et al. 2007; Barrett, Hansoul et al. 2008). However, the ubiquity of regulatory variation throughout the human genome (Dixon, Liang et al. 2007; Stranger, Nica et al. 2007) makes coincidental overlaps of eQTLs and complex trait loci very likely. This likelihood is a direct consequence of the correlation structure in the genome (linkage disequilibrium - LD), which makes functionally unrelated variants statistically correlated.

As sample sizes increase, allowing the discovery of larger numbers of eQTLs of smaller effect size and as the expression experiments will be performed in a larger variety of tissues, we can envisage that almost every gene will have an associated eQTL under a certain condition. Consequently, the probability that any of these will map to a genomic region where a GWAS SNP also resides is very high. Therefore, it is important to emphasize that while it is very tempting to infer potential causal mechanisms based on such overlaps, this would be a naïve inference in the absence of additional supporting evidence for causality. In the long run, this will not only be an issue for gene expression, but also for any other cellular phenotype. Association studies for intermediate phenotypes with possible relevance to complex traits are underway and their results will overlap some of the GWAS signals. The biological meaning of these overlaps will again need to be evaluated in the context of the genome’s correlation structure.

It is not evident though, how to model each genomic region with overlapping association signals in the absence of information about the history of the region. Accounting for the historical parameters of a region under the coalescent, while desirable, is computationally and practically not feasible since the human population history is too complex to properly model and small deviations or slightly incorrect assumptions could create false signals or reduce power.

In order to distinguish such accidental co-localizations (Chen, Zhu et al. 2008; Plagnol, Smyth et al. 2009) from true sharing of causal variants, I propose here an empirical methodology instead. This directly combines eQTL and GWAS data while accounting for the LD of the region harbouring the GWAS SNP. In this chapter, I demonstrate the value of the approach by predicting the regulatory impact of several GWAS variants in *cis* and *trans* and I also show that the correlation strength (r^2, D') between the GWAS SNP and the eQTL is not a sufficient predictor of regulatory mediated disease effects. This work has been described in (Nica, Montgomery et al. 2010).

3.1 Current GWAS signals are enriched for regulatory variants

To identify likely causal effects (not variants since full sequencing data is not available at this point), I took advantage of published association data catalogued in the NHGRI database (Hindorff, Sethupathy et al. 2009) and gene expression data generated in LCLs derived from HapMap 3 individuals (see Methods). In this study, I limited the expression analysis to the 109 CEU individuals (European origin), as they are the closest in ancestry to the majority of individuals in published GWAS studies. I used the NHGRI database (accessed 02.03.09) to extract 976 GWAS SNPs with minor allele frequency (MAF) > 5% that were also genotyped in the HapMap 3 CEU, thus allowing to test the exact GWAS SNPs for associations with differential gene expression in LCLs. In total 17673 genes were examined. To discover eQTLs, I used Spearman Rank Correlation (SRC). This method captures the vast majority of associations discovered with standard linear regression (LR) models, with the additional advantage that it's not affected by outliers and hence has more power and allows direct comparison of nominal P-values (Stranger, Nica et al. 2007). I looked for both proximal (*cis*) and distal (*trans*) effects as follows: variants within 1Mb on either side of the transcription start site (TSS) of a gene are considered to be acting in *cis,* while those at least 5 Mb downstream or upstream of the TSS or on a different chromosome are considered to be acting in *trans*.

In order to assess the overall impact of the currently known GWAS SNPs on expression, I contrasted their *cis* and *trans* effects to those of a random set of SNPs, representing the null. In a QQ plot (Figure 3.1), I compared the distributions of the best *cis* and *trans* association p-values per SNP for the 976 GWAS SNPs (observed) to 1000 sets of most significant p-values of 976 random SNPs each (expected). The 1000 random sets of 976 SNPs were sampled to have identical MAF distribution to the GWAS SNPs.

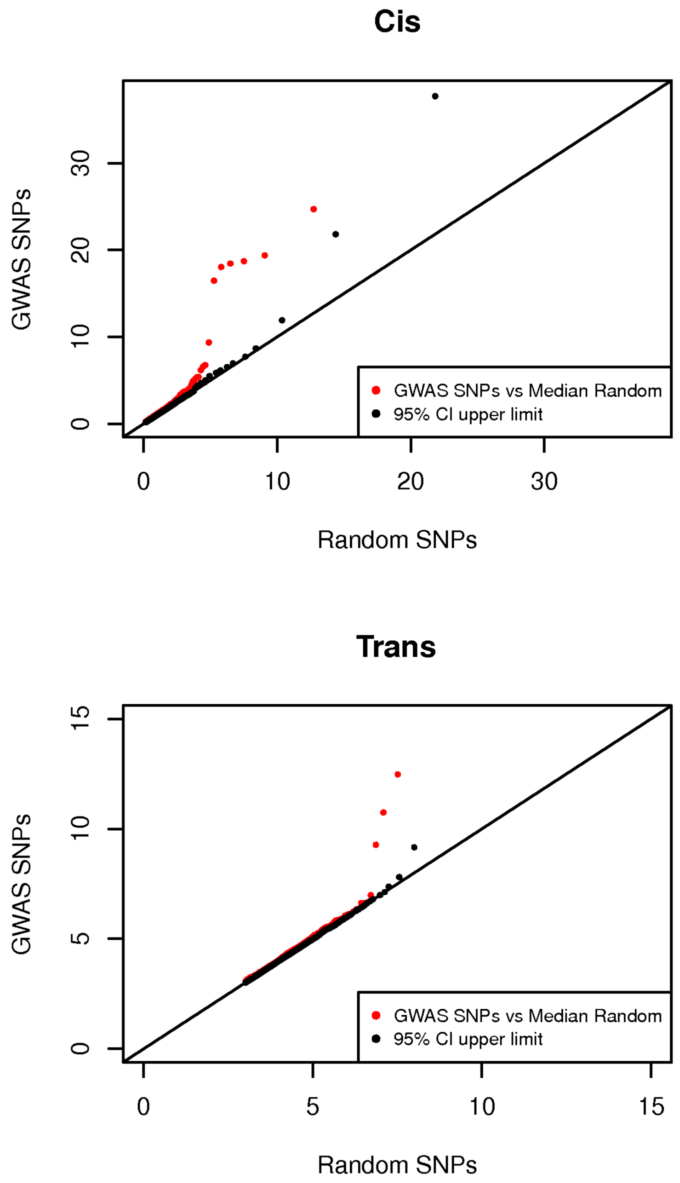


Figure 3.1. Excess of regulatory variants among GWAS signals. QQ plot depicting the excess of significant regulatory signal in GWAS data (976 NHGRI SNPs). For both the *cis* and *trans* analyses, the $-\log_{10}$(P-value) of the best associations per SNP are plotted. In red, the distribution of these values for GWAS SNPs is compared to that of the median of 1,000 sets of 976 random SNPs with same MAF distribution. In black, the estimated upper limit of the 95% confidence interval is plotted.

In *cis*, I observe a much stronger regulatory signal in the GWAS data compared to random (Figure 3.1). The significant difference between the two becomes apparent above a –$\log_{10}$(P-value) = 4. In *trans,* I also detect a more significant regulatory signal for GWAS SNPs compared to random, however not as strong as in *cis*. This is to be expected given that the much greater statistical space explored in *trans* limits the power to detect such effects.

Nevertheless, despite their confinement to one tissue type - LCLs, these comparisons support the overall explanatory potential of regulatory variation for the biological effects of GWAS variants. As expected given the nature of the tissue, the phenotypes responsible for this enrichment are immunity related (Figure 3.2).

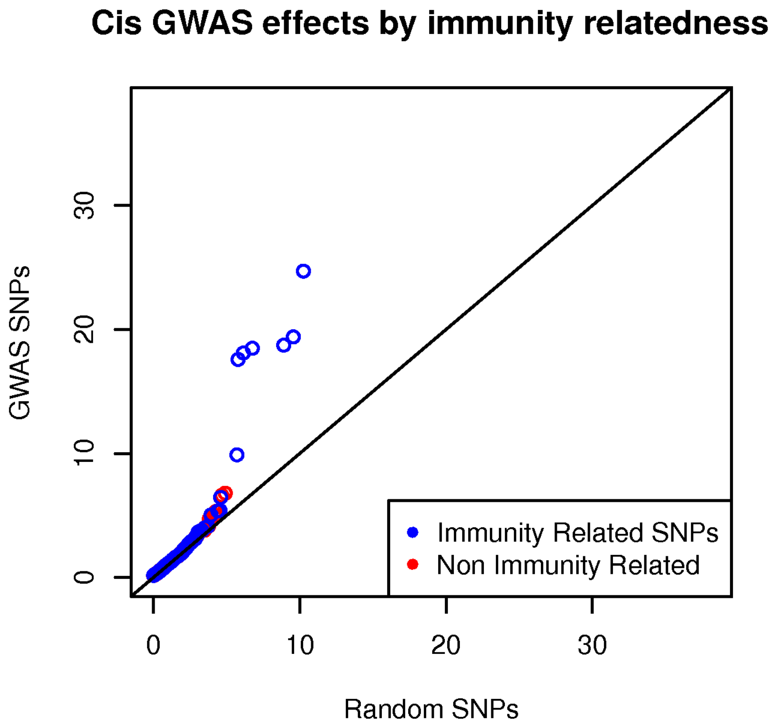


Figure 3.2. *Cis* regulatory enrichment stratified by immunity relatedness. The −$\log_{10}$(P-value) of the best associations per GWAS SNPs and a set of random SNPs are plotted. As expected given the tissue (LCLs), immunity related phenotypes are mainly responsible for the enrichment.

3.2 RTC score to distinguish between causal effects and coincidental overlaps

To identify the subset of causal effects from the regulatory enrichment observed, I focused only on the genomic regions harbouring either *cis* or *trans* eQTLs. I split the genome into recombination hotspot intervals based on genome-wide estimates of hotspot coordinates from McVean et al. (McVean, Myers et al. 2004). Limiting the search space for causal effects to these intervals is a reasonable conventional approach, as the lack of recombination events between the reported associated SNPs and the functional variants they are tagging enabled the discoveries through GWAS in the first place.

Given the abundance of *cis* eQTLs in the human genome, mere interval overlap is not sufficient to claim that a co-localized *cis* eQTL and a GWAS SNP are tagging the same functional variant. However, if the GWAS SNP and the eQTL do tag the same causal SNP, it is expected that removing the genetic effect of the GWAS SNP will have a marked consequence on the eQTL association. Starting from this hypothesis, I developed an empirical method to uncover regulatory mediated associations with complex traits. For all genes with a significant *cis* eQTL (0.05 permutation threshold as defined in Methods) (Stranger, Nica et al. 2007) in a given interval, I created corrected phenotypes from the residuals of the standard LR of the GWAS SNP against normalized expression values of the gene for which an eQTL exists. The residuals capture the remaining unexplained expression variance after the removal of the GWAS SNP effect. The SRC analysis was redone, this time with the pseudo phenotype, and the adjusted association P-value retained. Depending on the internal LD structure of the hotspot interval, the correlation between the GWAS SNP and the eQTL will vary, hence so will the P-values after and before correction. One way to assess the relevance of the GWAS SNP to the eQTL is to compare its correction impact to that of all other SNPs in the interval. For this purpose, I defined a Regulatory Trait Concordance (RTC) Score for each gene-GWAS SNP combination as a ratio taking into account the ranking of the correction with respect to all SNPs in the interval (Rank $_{\text{GWAS SNP}}$) and the total number of tested SNPs (N_{SNPs}) (see Methods).

$$RTC = \frac{N_{SNPs} - Rank_{GWAS\,SNP}}{N_{SNPs}}$$

The rank denotes the number of SNPs which when used to correct the expression data, have a higher impact on the eQTL (less significant adjusted P-value) than the GWAS SNP. As such, the RTC score will always be in the range (0,1], with values close to 1 indicating that the GWAS effect is the same as the eQTL effect.

The RTC score captures the LD structure of each tested region by taking into account the correction at all SNPs for every recombination hotspot interval. In addition, this ensures that RTC estimates are not up weighted in intervals with low number of SNPs (e.g. an extreme hypothetical case would be an interval with two SNPs only, the eQTL and the GWAS SNP; in this case the ranked correction at the eQTL would be high - $\text{Rank}_{\text{GWAS SNP}}$ = 1, as there is no other SNP in the interval to test; nevertheless, given just the 2 SNPs in the interval, the RTC score would only be 0.5 = (2 – 1) / 2). While this is not a problem for overestimating confident RTC scores, a caveat of the method is that intermediate values are equally discarded when in fact estimations derived from intervals with more SNP information should be up scaled (i.e. an RTC = 0.7 in an interval with 150 SNPs is more considerable than an RTC = 0.7 in an interval with 10 SNPs). Adjusting the value of the RTC score based on the SNP content of each region is a pending further development of the method. Meanwhile, one way to maximize the information content in each interval would be to include imputed SNP data. Given that the p-value associations prior to and after GWAS SNP correction are calculated with a non-parametric ranked test (SRC), it would be possible to use the estimates of allele dosage instead of the direct genotypes. This strategy has been shown to have comparable results to methods that take genotype uncertainty into account (Guan and Stephens 2008) and along with the SRC test as well as the permutations-based eQTL assignment, it should not be sensitive to outliers. A thorough evaluation of the use of imputed data to estimate RTC scores remains to be performed as a further improvement of the test.

3.3 RTC properties

The properties and robustness of the RTC score were investigated under the null hypothesis (H_0: eQTL and GWAS are tagging two different causal SNPs) and the alternative hypothesis (H_1: same causal SNP). For this purpose, I have simulated causal SNPs (cSNP), eQTLs and dSNPs (see Methods) varying the LD levels between them as well as the LD pattern of the hotspot interval where they reside. The cSNPs were then masked and subsequently, the RTC score was calculated under these different LD scenarios for both hypotheses.

The RTC score is uniformly distributed under the null, when the simulated causal eQTL SNP (c-eQTL) and the causal disease SNP (c-dSNP) are different (Figure 3.3, left panel). Under the H_1 on the other hand, the RTC score is right skewed, with a clear enrichment for values close to 1 recovering the single causal SNP effect (Figure 3.3, middle panel).

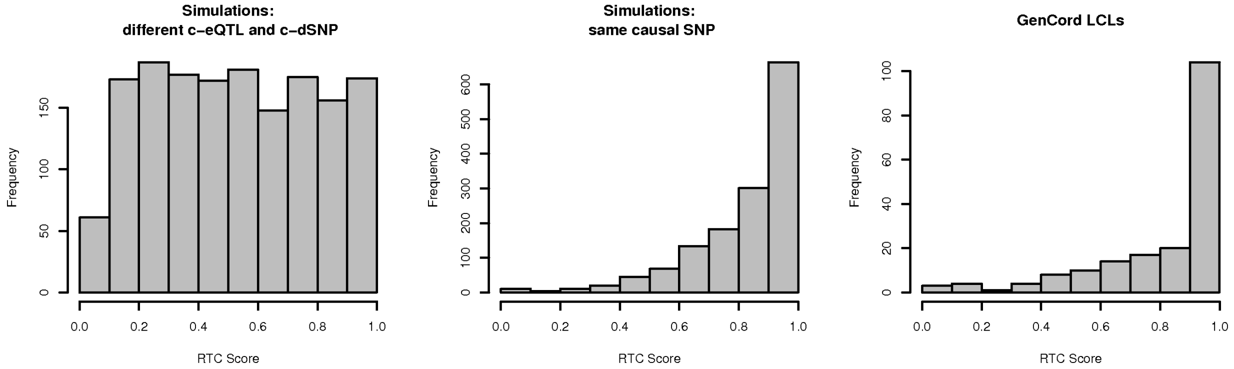


Figure 3.3. RTC score distribution following simulations. The RTC score is uniformly distributed for simulated eQTLs and dSNPs tagging two different causal variants in the same interval (left panel). The RTC Score is right-skewed for simulated eQTLs and dSNPs tagging the same functional variant (middle panel). The RTC score is sensitive to associations tagging a common functional variant in non-simulated data, when the GWAS trait is gene expression (GenCord LCL samples – right panel).

The simulations show that the complexity and variability of the LD structure in the genome impede the simple use of correlation metrics to infer shared causal effects. The statistical correlation (r^2) between the eQTL and the dSNP is not on its own sufficient to predict whether they tag the same cSNP. The RTC outperforms r^2 since it is able to recover causal effects even for low correlated pairs (Figure 3.4).

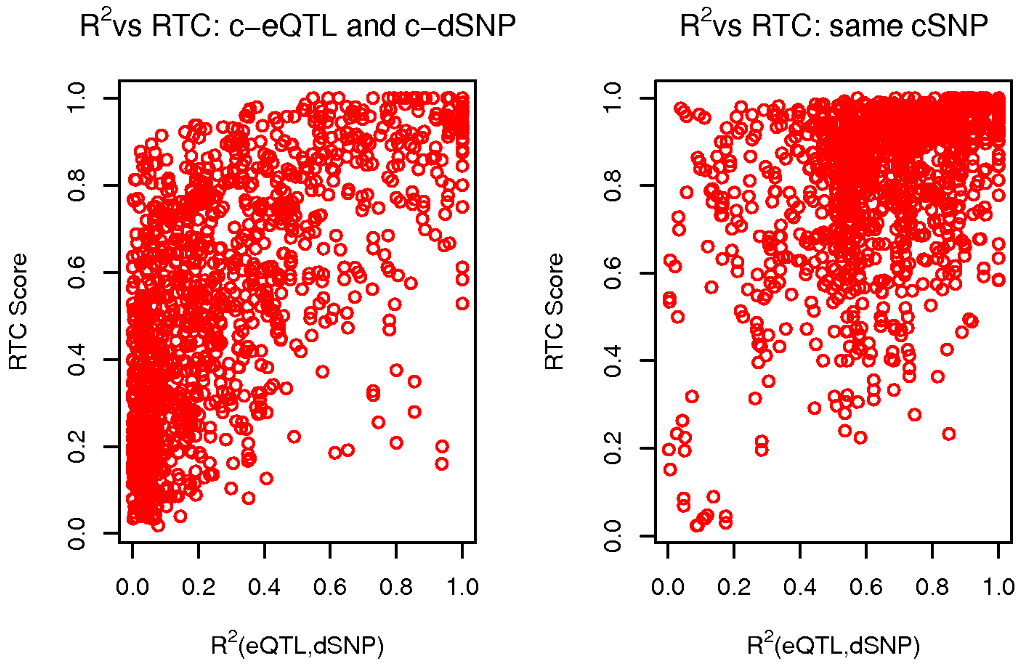


Figure 3.4. Properties of the RTC score when varying r^2. Simulation results depicting the relationship between the RTC score and the r^2 (eQTL, dSNP) when they tag different causal SNPs (H_0: left panel) versus one causal SNP (H_1: right panel). The RTC increases as expected with increased r^2 between the eQTL and the dSNP, but when tagging the same functional variant, various lower pairwise r^2 combinations can determine a high RTC. This makes r^2 on its own insufficient to detect shared causal effects.

The historical correlation metric between eQTLs and dSNPs (D') is also not fully predictive of high RTC scores (Figure 3.5). It can be observed from the H_0 simulation results that D' is not correlated with RTC, meaning that when the eQTL and dSNP tag different functional variants, the RTC score is not high just because D' is high. In addition, while high RTC scoring cases cluster much tighter around high D' values under the H_1 compared to r^2 previously, a high D' is not sufficient to predict causal effects. That is because it would be impossible to distinguish causal from coincidental effects given a perfect historical correlation scenario.

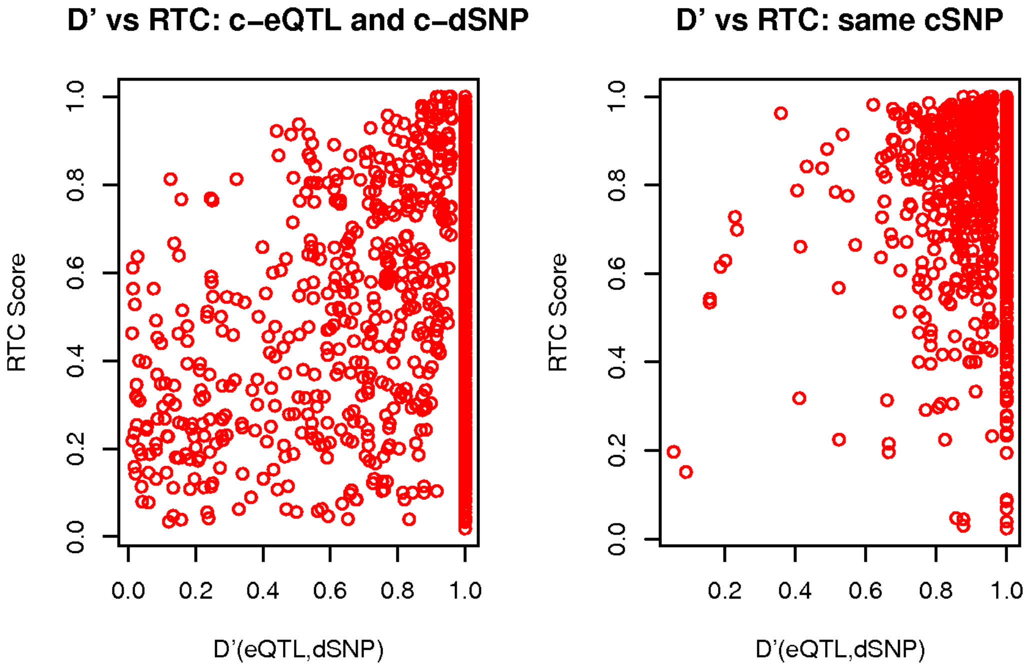


Figure 3.5. Properties of the RTC score when varying D'. Simulation results depicting the relationship between the RTC score and the D' (eQTL, dSNP) when they tag different causal SNPs (H_0: left panel) versus one causal SNP (H_1: right panel). D' is not correlated with RTC, therefore it will not determine high scores on its own in the absence of a common functional variant. Under the H_1, the majority of high RTC scoring pairs have high D', but in the case of a perfect historical correlation scenario, it's impossible to distinguish causal from coincidental effects with D' only.

Finally, the effect of the overall LD pattern in a region of interest on the estimation of the RTC score was investigated. For this purpose, I calculated the median r^2 of each hotspot interval (for all pairwise SNP combinations available per interval) and checked its relationship to the RTC score under the null and alternative hypothesis. It is expected that RTC will perform better in intervals with overall low LD, where the correlation between the eQTL and other non-disease SNPs will decay much faster, making the correction for the dSNP stand out. However, I confirm that the LD of the region does not determine high scores by itself. Intervals of low LD where different c-eQTLs and c-dSNPs reside have a uniform distribution of RTC scores (Figure 3.6, left panel). As expected, the H_1 simulations show that the RTC is most powerful in intervals with low median r^2 (Figure 3.6, right panel).

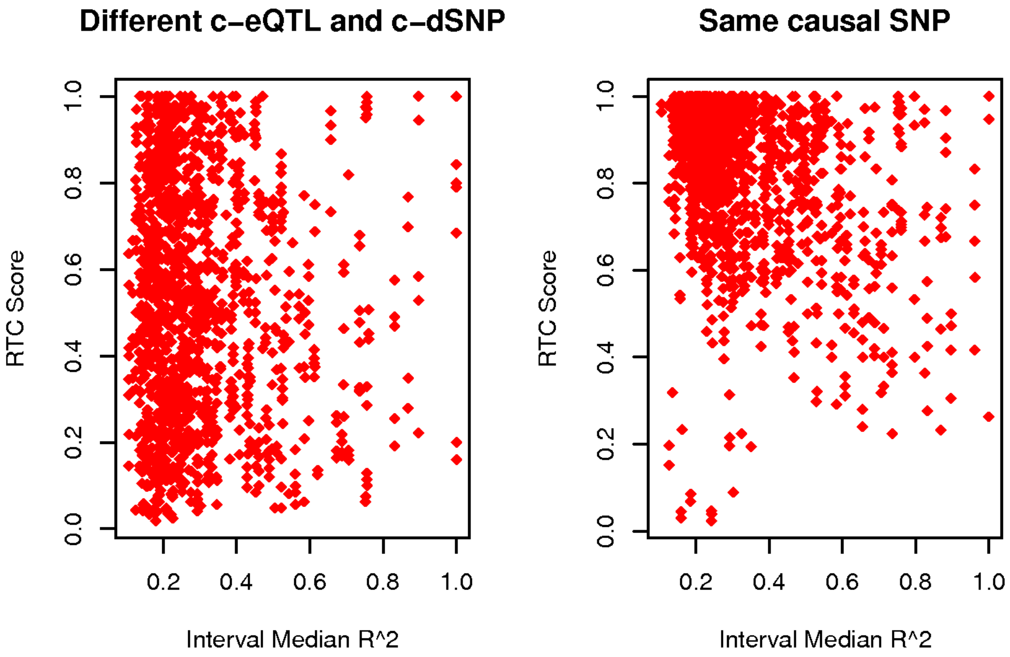


Figure 3.6. Properties of the RTC score when varying the median r^2 of the hotspot interval. Simulation results depicting the relationship between the RTC score and the local LD structure (median r^2) under the null (different causal SNPs - left panel) and alternative hypothesis (same causal SNP - right panel). Under H_0, the RTC score is evenly distributed, therefore intervals with overall low LD will not determine high RTC scores. Under H_1, the RTC performs best in intervals with overall low LD, where the correlation between the eQTL and other non-disease SNPs decays much faster, making the dSNP correction stand out.

3.4 RTC score when both traits are gene expression

In the first instance I tested the RTC method in a positive control experiment where intervals harbouring already identified regulatory associations were analyzed. I used published *cis* eQTLs (10^{-3} permutation threshold) discovered in the same tissue as the HapMap 3 CEU eQTLs (LCLs) but derived from an independent set of samples: 75 individuals of Western European origin from the GenCord resource (Dimas, Deutsch et al. 2009). In this experiment, I considered the GenCord eQTLs as the equivalent of GWAS SNPs and I limited the analysis to intervals with *cis* eQTLs in both datasets. Furthermore, I conditioned the associated genes for the same interval to be identical in the two expression datasets, expecting thus a common functional variant. As a result of this filtering, SNPs in 157 hotspot intervals were tested, associated with differential expression levels of 154 genes. As expected from the H_1 simulations, the RTC score distribution after correcting for the GenCord eQTLs is right-skewed (Figure 3.3, right panel), suggesting that the scoring method is sensitive to associations tagging the same functional variant. I detect 33 SNP-probe pairs with an RTC score of 1 out of the total 185 tested pairs. Given the marked difference in genotyping density between HapMap and GenCord (~1.2 million SNPs versus ~400,000 SNPs respectively) and the hypothesis that the 157 overlapping intervals share the same functional variant, approximately 3 times more perfect scoring cases (99 pairs with RTC score = 1) are expected than what we observe, had individuals from both datasets been equally densely genotyped. I use the degree of sharing between the eQTLs in the two datasets to derive a reasonable, yet conservative threshold: currently, 105 SNP-probe pairs pass the 0.9 RTC threshold, making it thus a suitable stringent cut-off for calling significant discoveries.

3.5 *Cis* results

Following the positive control analysis, I applied the scoring method in a disease GWAS setting using the NHGRI SNPs described in Section 3.1. The respective 976 common GWAS SNPs map to 784 hotspot intervals. Of these, I focused the *cis* analysis on GWAS intervals (N=130) where at least one significant *cis* eQTL at a 0.05 permutation P-value threshold also resides. For the *trans* analysis, I ordered all 784 GWAS intervals by their most significant *trans* eQTL and kept the topmost 50 intervals for further examination. Table 3.1 summarizes the most confident *cis* results ordered by RTC score. I detect SNP-gene combinations passing the 0.9 threshold for 28 intervals out of the 130, twice as many than expected by chance (13 expected top 10% scoring intervals under the uniform distribution). The RTC method confirms prior results in the literature suggestive of disease effects mediated through expression (*ORMDL3* for asthma risk (Moffatt, Kabesch et al. 2007), *C8orf13* locus for systemic lupus erythematosus risk (Hom, Graham et al. 2008), *SLC22A5* for Crohn’s disease (Peltekova, Wintle et al. 2004; Barrett, Hansoul et al. 2008). In addition, other yet unknown candidate genes for a variety of conditions are identified.

GWAS SNP	Complex Trait	Gene	RTC	Chr
rs2064689	Crohn's disease	WDR78	1	1
rs3129934	Multiple sclerosis	HLA-DRB1	1	6
rs2188962	Crohn's disease	SLC22A5	1	5
rs1015362	Burning and freckling	TRPC4AP	1	20
rs2735839	Prostate cancer	C19orf48	1	19
rs6830062	Height	LCORL	1	4
rs2242330	Parkinsons disease	TMPRSS11A	1	4
rs7498665	Body mass index,Weight	EIF3CL	1	16
rs2872507	Crohn's disease	ZPBP2	0.99	17
rs255052	HDL cholesterol	AGRP	0.99	16
rs4549631	Height	TRMT11	0.98	6
rs9469220	Crohn's disease	ILMN_29412	0.98	6
rs11083846	Chronic lymphocytic leukemia	SLC8A2	0.98	19
rs13277113	Systemic lupus erythematosus	C8orf13	0.97	8
rs9272346	Type 1 diabetes	HLA-DRB1	0.96	6
rs12324805	Body mass index	STARD5	0.96	15
rs3764261	HDL cholesterol	MT1H	0.96	16
rs3135388	Multiple sclerosis	HLA-DRB5	0.96	6
rs3814219	Endothelial function traits	FAM26B	0.95	10
rs12708716	Type 1 diabetes	ILMN_32084	0.95	16
rs2269426	Plasma eosinophil count	HLA-DRB1	0.95	6
rs10769908	Body mass index	C11orf17	0.94	11
rs4130590	Bipolar disorder	ILMN_17339	0.94	9
rs7216389	Asthma	ORMDL3	0.94	17
rs3796619	Recombination rate (males)	CRIPAK	0.93	4
rs1748195	Triglycerides	DOCK7	0.93	1
rs2903692	Type 1 diabetes	ILMN_32084	0.93	16
rs3197999	Crohn's disease	SLC38A3	0.92	3
rs9858542	Crohn's disease	SLC38A3	0.92	3
rs6441961	Celiac disease	LIMD1	0.92	3
rs660895	Rheumatoid arthritis	PSMB9	0.91	6
rs9652490	Essential tremor	ILMN_111363	0.91	15
rs1397048	Hemostatic factors	OR8H2	0.91	11
rs3825932	Type 1 diabetes	CTSH	0.91	15
rs2395185	Ulcerative colitis	ILMN_29412	0.9	6

Table 3.1. Candidate *cis* results. Candidate genes (RTC score ≥ 0.9) for *cis* regulatory mediated GWAS effects. RTC applied on 976 GWAS SNPs from NHGRI and HapMap 3 CEU expression data in LCLs. The higher the score, the more likely it is that the GWAS SNP and the eQTL for the gene shown are tagging the same functional variant.

An interesting example of a novel *cis* regulatory mediated effect is the one for Crohn's disease with gene *SLC38A3*, member 3 of the solute carrier family 38. Independent studies detected significant Crohn's associations of two SNPs in the same hotspot interval on chromosome 3: rs3197999 (Barrett, Hansoul et al. 2008), a non-synonymous SNP in gene *MST1* and rs9858542 (Parkes, Barrett et al. 2007; WTCCC 2007), a synonymous SNP in nearby gene *BSN*. Suggestive literature evidence supports the role of *MST1* in Crohn's pathogenesis: the protein encoded by *MST1* (macrophage-stimulating protein – MSP) and its receptor MST1R are reportedly involved in macrophage chemotaxis and activation (Leonard and Skeel 1976) and have a role also in regulating inflammatory responses following pro-inflammatory signals (Morrison, Wilson et al. 2004). These lines of evidence, in addition to the disease associated non-synonymous SNP made *MST1* the most attractive candidate gene out of the many present in that region (Goyette, Lefebvre et al. 2008). However, the data presented here supports an additional regulatory component underlying the susceptibility locus. For both GWAS SNPs, *SLC38A3* is the highest scoring candidate in the region (RTC score: 0.92). Interestingly, this is functionally similar to another Crohn's susceptibility gene *SLC22A5* confirmed with the RTC method (RTC score = 1) and also encoding a sodium dependent multi-pass membrane protein (solute carrier family protein). The observed direction of effect is the same for both genes (eQTLs associate with low expression levels) as in previous expression datasets (Barrett, Hansoul et al. 2008) and suggests a possible involvement of this gene family in the disease. This is in agreement with recent studies reporting that disease causative genes are functionally more closely related (Franke, van Bakel et al. 2006).

Overrepresentation of immunity-related results

The tissue under investigation is LCLs so it is expected that GWAS signals of immunity related traits (comprising here autoimmune disorders and diseases of the immune system e.g. AIDS progression) more likely show an overlap with eQTLs. In order to evaluate the relevance of the presented results, I analyzed the distributions of the best RTC scores per GWAS SNP stratified by the immunity relatedness of the complex trait they associate with (Figure 3.7).

I observe a significant overrepresentation of high-scoring genes (RTC ≥ 0.9) for immunity related traits compared to non-immunity related ones (Fisher's Exact Test, P-value = 0.0125) (Fraser and Xie 2009). This suggests that the scoring scheme predicts regulatory effects of the relevant phenotypes. In addition, we observed that for GWAS signals with RTC score ≥ 0.9, only 10% of the nearest gene to the GWAS SNP was also the eQTL gene. These however, correspond as expected to instances when the eQTL gene is also the nearest gene to the eQTL itself. If that is not the case, the inference of relevance of a gene simply based on its proximity to the GWAS SNP is not informative.

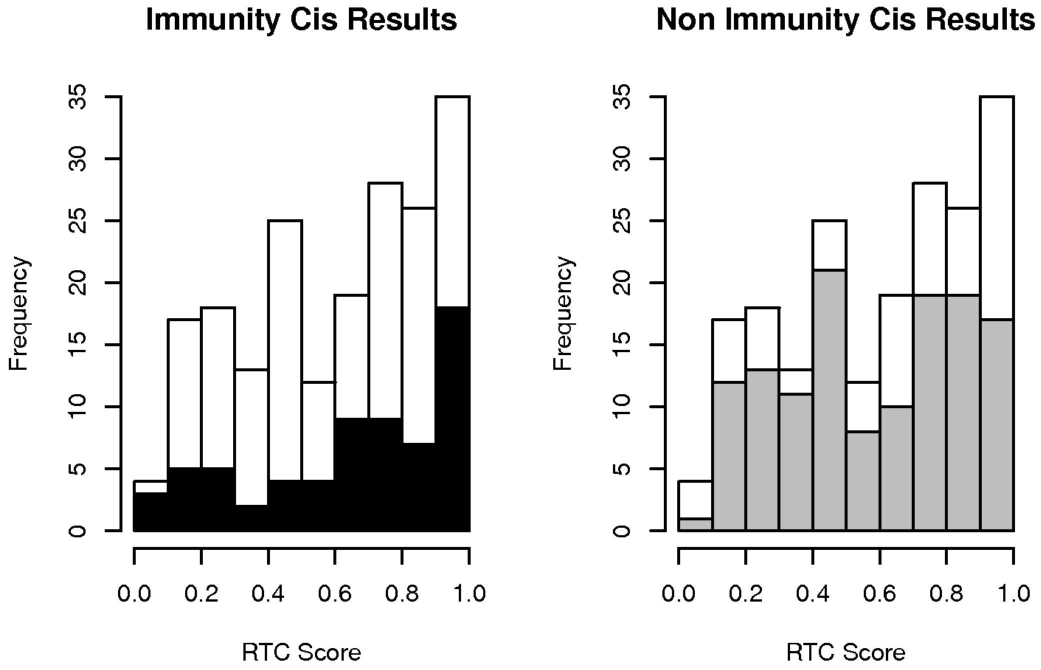


Figure 3.7. Overrepresentation of immunity-related high RTC scoring *cis* signals. Distribution of best RTC scores per GWAS SNP stratified by immunity relatedness. Histogram contains results from the analysis of 130 hotspot intervals with colocalizing disease SNPs and *cis* eQTLs. We observe a significant overrepresentation of high-scoring (RTC ≥ 0.9) candidate genes (black bars) for immunity related complex traits compared to non-immunity related ones (grey bars) (Fisher's Exact Test, P-value = 0.0125).

3.6 *Trans* results

Even if the causal SNP is not *cis*-regulatory, using gene expression to determine its downstream targets, coupled with information about the biological pathways these targets act in could help interpret the primary GWAS effect.

I investigate this hypothesis in the topmost 50 GWAS intervals ordered by their *trans* eQTL significance. For each interval, I apply the RTC scoring scheme on the subset of genes in the whole genome with a notable effect in *trans* (SRC nominal P-value < 10^{-5}). These signals amount to a total of 552 genes. I obtain SNP-gene combinations passing the 0.9 RTC score threshold for 24 of the 50 tested intervals (corresponding to a total of 85 genes). Six of these intervals contain GWAS SNPs associated with immunity related traits (Table 3.2).

While not statistically significant - unsurprisingly given that only a small subset of the total GWAS intervals is tested - these examples support the usefulness of the *trans* approach. As hypothesized, for the same complex trait associated SNP, several potential candidate genes in *trans* can be discovered throughout the genome. Some of these are biologically plausible results and merit further investigation. However, many *trans* candidates are hard to interpret at this stage given their incomplete annotation and further functional studies will need to be performed for validation.

Table 3.2. Candidate *trans* results. Candidate *trans* genes likely involved in the same biological pathways, relevant to the GWAS SNPs (GWAS SNP and the genes it affects in *trans* often reside on different chromosomes, as indicated in the SNP Chr and Genes Chr fields respectively). Signals related to the same hotspot interval separated by a horizontal line. Regulatory *trans* effects RTC applied in *trans* on 976 GWAS SNPs from NHGRI and HapMap 3 CEU expression data in LCLs. Table contains only the confident results (RTC Score ≥ 0.9) for the six immunity related intervals.

GWAS SNP	Complex Trait	Genes	RTC	SNP Chr	Genes Chr
rs2251746	Serum IgE levels	SLC25A18	0.99	1	22
rs983332	Response to TNF antagonists	RGS16, IGSF3	0.97	1	1
rs983332	Response to TNF antagonists	C17orf58	0.97	1	17
rs653178	Celiac disease	PAX8, DOK1	1	12	2
rs17696736	Type 1 diabetes	PAX8, DOK1	0.98	12	2
rs2542151	Crohn's,Type 1 diabetes	MMP12	1	18	11
rs2542151	Crohn's,Type 1 diabetes	SLC39A4, PSD3, AHNAK2, FAM108B1, CYP2S1, CLEC7A	0.97	18	8, 8, 14, 9, 19, 12
rs2542151	Crohn's,Type 1 diabetes	LENEP	0.91	18	1
rs3134792	Psoriasis	ADRA2C	1	6	4
rs3134792	Psoriasis	DPEP1, ARHGEF3	0.99	6	16, 3
rs1265181	Psoriasis	POU5F1P1	0.96	6	8
rs1265181	Psoriasis	DPEP1	0.95	6	16
rs1265181	Psoriasis	CYP4F8, ADRA2C	0.94	6	19, 4
rs1265181	Psoriasis	RGS9	0.92	6	17
rs2395185	Ulcerative colitis	B4GALT2, ASB5	0.97	6	1, 4
rs2395185	Ulcerative colitis	STK32A	0.94	6	5
rs2395185	Ulcerative colitis	OXT	0.93	6	20
rs2395185	Ulcerative colitis	CSRP3	0.92	6	11
rs2395185	Ulcerative colitis	LGALS4	0.91	6	19
rs3135388	Multiple sclerosis	LIMS1	0.95	6	2
rs477515	Inflammatory bowel disease	B4GALT2	1	6	1
rs477515	Inflammatory bowel disease	ASB5	0.99	6	4
rs477515	Inflammatory bowel disease	STK32A	0.95	6	5
rs477515	Inflammatory bowel disease	OXT	0.94	6	20
rs477515	Inflammatory bowel disease	CSRP3	0.93	6	11
rs477515	Inflammatory bowel disease	DCHS2	0.91	6	4
rs477515	Inflammatory bowel disease	LGALS4	0.9	6	19
rs615672	Rheumatoid arthritis	DCHS2	0.99	6	4
rs6457617	Rheumatoid arthritis	SMARCD3	0.95	6	7
rs6457620	Rheumatoid arthritis	SMARCD3	0.95	6	7
rs660895	Rheumatoid arthritis	RETSAT	0.99	6	2
rs660895	Rheumatoid arthritis	CALCR	0.98	6	7
rs9268877	Ulcerative colitis	LIMS1	0.97	6	2
rs9268877	Ulcerative colitis	B4GALT2	0.94	6	1
rs9268877	Ulcerative colitis	ASB5	0.91	6	4
rs9272346	Type 1 diabetes	LIMS1	0.97	6	2
rs9272346	Type 1 diabetes	WHDC1L1	0.94	6	15
rs9272346	Type 1 diabetes	ASB5	0.93	6	4
rs9272346	Type 1 diabetes	SEMA6D, OXT, B4GALT2	0.92	6	15, 20, 1

A subset (N=15) of the hotspot intervals containing GWAS SNPs and tested in this chapter harbour both *cis* and *trans* eQTLs (as defined in Methods). For two of the 15 intervals, I detect potential explanatory regulatory effects (genes with high RTC score) in both *cis* and *trans* (Table 3.3.). It is likely that changes in expression levels of all these genes are relevant to the single common GWAS signal. Interestingly for example, the *DOCK7* (dedicator of cytokinesis 7) locus has been implicated in coronary heart disease risk (Aulchenko, Ripatti et al. 2009) and SNP variants at the *SORCS2* (sortilin-related VPS10 domain containing receptor 2) locus have been associated with hemorrhagic stroke (Yoshida, Kato et al. 2010). Both genes score a high RTC with SNP rs1748195 associated with triglyceride levels, a quantitative trait highly relevant to heart disorders. Functional verification of similar gene connections might lead to the discovery of new disease-relevant pathways.

	GWAS SNP	Complex Trait	Gene	RTC	SNP Chr	Gene Chr	Interval
cis	rs1748195	Triglycerides	DOCK7	0.93	1	1	1:62673568-62974568
trans	rs1748195	Triglycerides	SORCS2	0.9	1	4	1:62673568-62974568
cis	rs1007738	Bone mineral density (hip)	ACP2	0.88	11	11	11:46234001-46861001
trans	rs1007738	Bone mineral density (hip)	CAPN12	0.98	11	19	11:46234001-46861001
trans	rs1007738	Bone mineral density (hip)	SYNGR3	0.87	11	16	11:46234001-46861001
trans	rs1007738	Bone mineral density (hip)	TMEM149	0.83	11	19	11:46234001-46861001
trans	rs1007738	Bone mineral density (hip)	PBXIP1	0.82	11	1	11:46234001-46861001

Table 3.3. Hotspot intervals with overlapping *cis* and *trans* effects as indicated by the high RTC score. Candidate regulatory effects explaining GWAS signals were detected for two of the 15 intervals tested for both *cis* and *trans* effects.

3.7 RTC outperforms alternative correlation metrics

The power to detect significant associations between genotyped SNP proxies and a phenotype depends on the correlation between those proxies and the functional variant (Pritchard and Przeworski 2001). Just like for the simulated data, I tested whether the correlation between a GWAS SNP and its co-localizing eQTL is sufficient for predicting a shared causal effect. For both the *cis* and the *trans* analysis, I observe that the r^2 between the eQTL and the disease SNP is not a direct predictor of the RTC score, and in several cases I predict that even pairs with low r^2 are likely tagging the same functional effect (Figure 3.8, top panel).

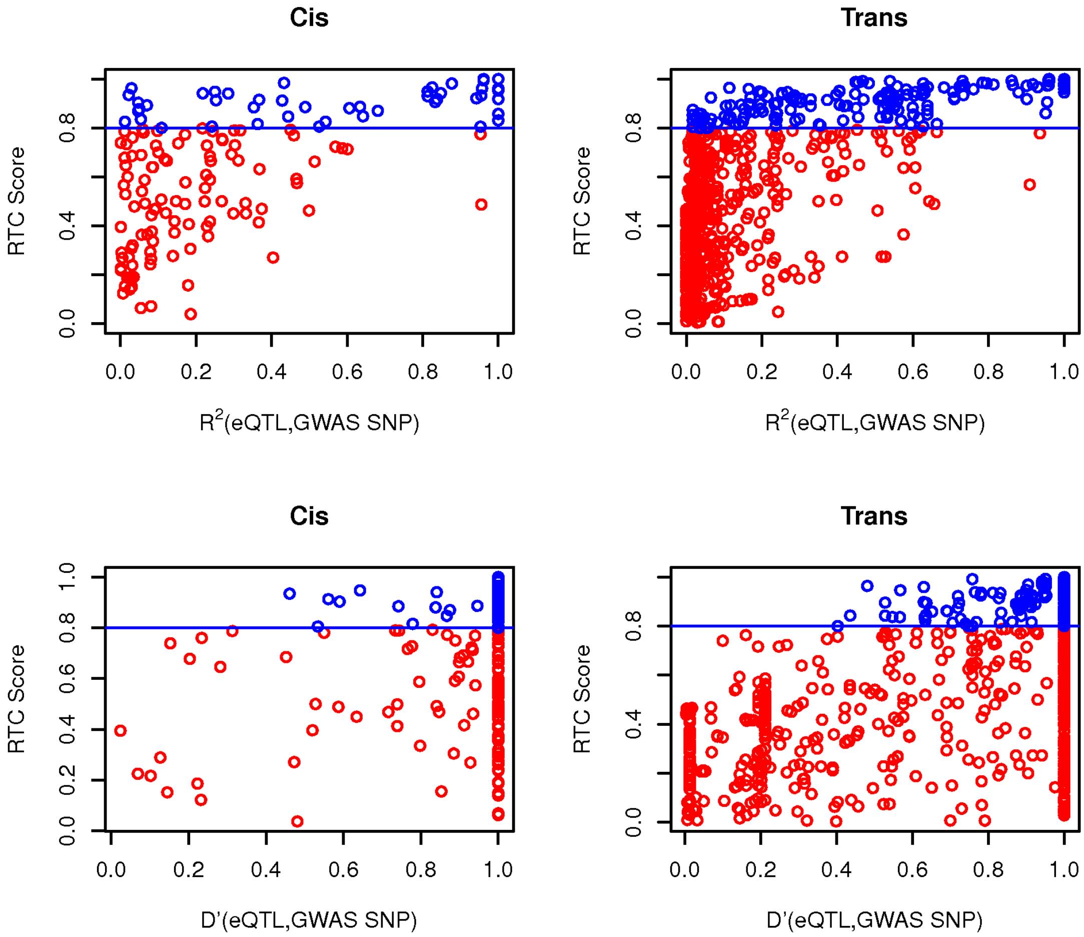


Figure 3.8. The RTC method compared to standard LD measurements in the observed data. Neither r^2 nor D' between the eQTL and the GWAS SNP are direct predictors of a high RTC score. Highlighted here are the results from the *cis* and *trans* analyses. I obtain high scoring results (RTC scores ≥ 0.8 in blue) for cases with a high correlation between the disease SNP and the eQTL as expected, but also for pairs with low statistical correlation (r^2 – top panel). As shown in the bottom panel, many of these high scoring pairs are historically correlated (D' = 1), but so are many more by chance. Additionally, high scoring pairs with low D' can be detected as well. Hence, no obvious combination of the two LD measures can predict a high RTC score.

The reason for this is that many of the high scoring pairs with poor statistical correlation (low r^2) are actually historically correlated (D'=1). Nevertheless, D' is not very informative either (Figure 3.8, bottom panel), the main problem here being that in regions with generally high D' among many SNPs, one cannot determine which of the pairs actually represents a common functional variant.

Another metric of potential predictive value is the fraction of eQTL variance explained by the dSNP. Figure 3.9 indicates the relationship between the RTC score and the fraction of explained variance at the eQTL left unexplained after the dSNP correction (ratio of linear regression adjusted R^2 after and before correction). As expected given the definition of the RTC, the highest density of good scoring results is registered for dSNPs that explain most of the eQTL variance. However, RTC outperforms the variance metric, scoring high even when less of the eQTL variance is explained by a dSNP. As such, setting a threshold on the explained variance would not be sufficiently informative either.

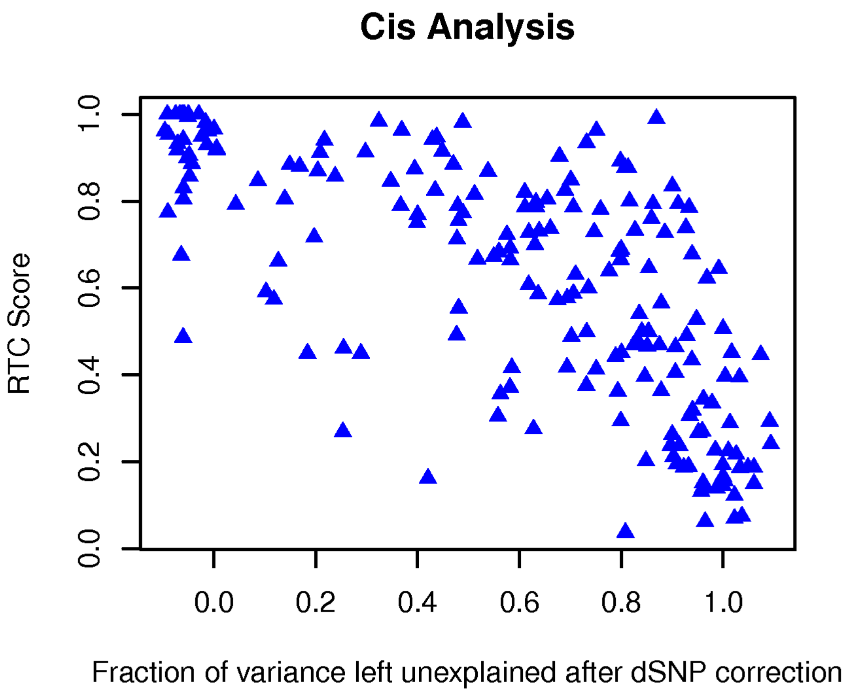


Figure 3.9. The fraction of eQTL variance explained away by the dSNP versus the RTC score. Contrasted are the LR adjusted R^2 at the eQTL after and before correction of the dSNP. It is observed that while most high scoring pairs correspond to cases of lowest variance left unexplained, solely using an arbitrary variance threshold would cause other interesting cases to be missed.

3.8 Conclusions

In this chapter, I described a newly developed empirical methodology, called Regulatory Trait Concordance (RTC). The purpose of this method is to account for local LD structure in the human genome and integrate eQTLs and GWAS results to reveal the subset of association signals that are due to *cis* eQTLs. This approach aims to help understand some of the biological mechanisms - should they be regulatory - behind the genetic associations with complex diseases. Candidate genes linked to the SNP variants reported so far as implicated in disease susceptibility are often chosen solely based on genomic proximity criteria. The RTC enables therefore a more informed choice of candidate disease genes, based on evidence in favour of common functional regulatory effects.

Genomic regions of various LD patterns were first simulated to explore the properties of the RTC score. Simulated intervals for both cases when a single or two different causal variants exist were analyzed. Consequently, I showed that the proposed scoring scheme outperforms SNP correlation metrics, be they statistical (r^2) or historical (D'). Following the observation of a significant abundance of regulatory signals among currently published GWAS loci, I applied the method on expression data in blood-derived LCLs extracted from HapMap 3 individuals of European descent. Relevant genes under regulatory control were prioritized for each of the respective complex traits. As such, I detected several potential disease causing regulatory effects, with a strong enrichment for immunity-related conditions, consistent with the nature of the cell line tested (LCLs). Furthermore, I presented an extension of the method in *trans*, where interrogating the whole genome for downstream effects of the disease variant can be informative regarding its unknown primary biological effect.

Overall, the RTC method supports the integration of cellular phenotype associations with organismal complex traits as a way to biologically interpret the genetic determinants of these traits.

4 Tissue-specificity of *cis* regulatory variants

Most of the knowledge on the genetic basis of regulatory variation has been gained so far by examining whole-genome expression patterns in blood-derived cells or cell lines (lymphocytes, LCLs) (Cheung, Spielman et al. 2005; Stranger, Forrest et al. 2005; Goring, Curran et al. 2007). Blood-derived cell-types continue to be the easiest accessible source for large-scale transcript level profiling, however incorporating information from a variety of other tissues is essential. Both during development as well as throughout the process of cellular differentiation, some genes are expressed ubiquitously while others display tissue-specific characteristics (Myers, Gibbs et al. 2007; Schadt, Molony et al. 2008). Additionally, many phenotypes manifest themselves only in certain tissues (Nowak and Davies 2004; Oksenberg and Baranzini 2010). Given the key role of regulatory variation in shaping complex phenotypes of medical importance, it is of special interest to assess the extent of expression differences between tissues that can be attributed to differential regulatory control.

Promising advancement towards this goal has been made recently by several studies identifying and comparing eQTLs in multiple human tissues. Myers etal. were the first to explore genetic variation influencing normal human cortical expression (Myers, Gibbs et al. 2007). The authors estimate using expression and genotypic data from 193 samples that 58% of the transcriptome is cortically expressed and of the expressed transcripts, 21% have significant eQTLs. Little overlap can be found between this eQTL set and results from previous analyses on blood-derived cells. While differences between the compared studies with respect to the samples and genotyping platforms used explain some of the modest overlap, it is very likely that variants discovered in the cortical samples underlie brain specific control of gene expression. In conjunction with results from GWAS, brain specific eQTLs could help uncover the genetic basis of some neurologic disorders.

In a study on 400 human liver samples, Schadt etal. identified more than 6000 SNP – gene associations (Schadt, Molony et al. 2008). Many of the genes detected in this experiment had already been linked to a variety of complex diseases, expectedly given the liver's essential role in many human metabolic processes. The same expression platform employed for the human liver cohort has also been used on a set of human blood and adipose tissue samples in another study (Emilsson, Thorleifsson et al. 2008). The authors evaluated the *cis* eQTL overlap in the three tissues, estimating ~30% sharing (Schadt, Molony et al. 2008). Efforts from our group have also contributed to the understanding of regulatory variation in a cell-type specific context. In a systematic study controlling for confounding associations due to different population samples or discrepant technological and statistical methods used, eQTLs were detected and compared across LCLs, fibroblasts and T-cells derived from the same 75 GenCord individuals (Dimas, Deutsch et al. 2009). The authors report that 69-80% of all discoveries (*cis* eQTLs) are cell-type specific, highlighting the need of sampling multiple tissue expression datasets in order to describe the full repertoire of regulatory variants.

Documenting cell-type specific regulatory variation is very important from the disease perspective. Integrating expression data with GWAS results can be informative for discovering genes and pathways whose disruption likely causes disease (Chen, Zhu et al. 2008; Nica and Dermitzakis 2008; Nica, Montgomery et al. 2010). However, this is only possible when the tissue of expression is relevant to the interrogated complex trait (Nica and Dermitzakis 2008). eQTLs discovered in LCLs have helped explain GWAS associations with childhood asthma (Moffatt, Kabesch et al. 2007) and Crohn's disease (Libioulle, Louis et al. 2007), two autoimmune inflammatory disorders. The adipose and blood cohorts analyzed by Emilsson etal. had been assessed for various phenotypes too, including obesity relevant traits. Notably, 50% of the *cis* signals were estimated as overlapping between the two cohorts, but a marked correlation with obesity-related traits was only observed for gene expression measured in adipose tissue (Emilsson, Thorleifsson et al. 2008). These observations certify the importance of integrating data from a relevant tissue when trying to interpret GWAS results using gene expression as an intermediate phenotype. Nevertheless, it is still unclear what the pattern of diminishing returns is across human tissues and what tissues could serve as highly informative in large cohorts. For example, LCLs have been useful in less expected cases enabling candidate gene discovery for associations with autism (Nishimura, Martin et al. 2007) or bipolar disorder (Iwamoto, Bundo et al. 2004).

In this chapter, I further explore the complexity of the human regulatory variation landscape in LCLs and two primary tissues (skin and fat) derived from the same subset of female twins from the UK Adult Twin registry (Spector and Williams 2006). In line with previous studies, I report extensive tissue-specificity of eQTLs using both a standard association method as well as a Bayesian factor analysis model. I describe the properties of eQTLs in each tissue and I propose that continuous estimates of statistical significance as well as the direct comparison of the magnitude of effect on the fold change in expression are essential properties that jointly provide a biologically realistic view of tissue-specificity.

4.1 Abundant eQTL discoveries per tissue

The pilot MuTHER samples were genotyped and profiled for gene expression in three tissues: LCLs, skin and fat. Normalization was performed separately in each tissue (Methods). The overlapping set of successfully genotyped samples with available expression data amounted to 156 individuals for LCL (30 MZ pairs, 37 DZ pairs, 22 singletons), 160 for skin (31 MZ pairs, 37 DZ pairs, 24 singletons) and 166 for fat (31 MZ pairs, 40 DZ pairs, 24 singletons). This final dataset was used for eQTL analysis (MZ and DZ pairs per tissue - Table 4.1).

The probes on the array were mapped to Ensembl gene IDs and only a confident subset was kept for analysis (27,499 probes mapping uniquely to 18,170 Ensembl genes). 865,544 SNPs passing quality check (Methods) were tested for associations with these probes.

		MZ pairs	DZ pairs
3 tissues	LCL-SKIN-FAT	28	30
2 tissues only	LCL-SKIN	1	2
	LCL-FAT	1	5
	SKIN-FAT	2	5
1 tissue only	LCL	0	0
	SKIN	0	0
	FAT	0	0
Total		32	42

Table 4.1. Successfully genotyped twin pairs (MZ and DZ) with available gene expression data. Number of twin pairs per tissue sharing both genotypic and expression information.

The eQTL analysis was performed separately in each tissue. I considered only unrelated individuals at a time by separating twins from the same pair and thus performing two independent eQTL analyses per tissue. This study design, hereafter named Matched Co-Twin Analysis (MCTA), permits immediate replication and validation of eQTL discoveries. This is important and unique with respect to previous eQTL studies which do not go beyond reporting the most significant findings (Morley, Molony et al. 2004; Cheung, Spielman et al. 2005). Given the known inter-individual variability in gene expression levels and the multiple sources of variation that can contribute to this, replicating the genetic determinants of expression differences (eQTLs) is essential, much like in any GWAS exercise. Furthermore, in a multiple-tissue expression design like here, where one of the main goals is to assess the extent of eQTL tissue-specificity, it is very useful to contrast between-tissue to within-tissue variability of expression changes for properly assessing the tissue-dependent level of regulatory control (section 4.4).

Spearman Rank Correlation (SRC) was used to detect associations and I restricted the search to *cis* effects located within 1Mb on either side of a gene's transcription start site (TSS). Statistical significance was assessed at different thresholds using permutations (10,000 per gene) (Methods). An abundance of *cis* eQTLs was detected in each tissue, at a comparable rate to other studies of similar sample size (Stranger, Nica et al. 2007; Dimas, Deutsch et al. 2009). At a permutation significance level of 10^{-3}, roughly 18 genes are expected to have at least one significant association by chance. At this threshold level, I detect significant associations with 509, 238 and 462 genes in LCL, skin and fat respectively for the first subset of the twin cohort (Twin 1) (Table 4.2). Unless otherwise stated, the 10^{-3} permutation cut-off corresponding to an FDR rate of 3.5% in LCL and fat and 7.5 % in skin was henceforth chosen when exploring eQTL properties.

	LCL		SKIN		FAT	
Permutation Threshold	Twin 1	Twin 2	Twin 1	Twin 2	Twin 1	Twin 2
10^{-4}	296	360	123	125	303	304
10^{-3}	509	556	238	231	462	488
10^{-2}	1014	1059	605	676	982	1068

Table 4.2. *Cis* eQTL associations detected with SRC analysis. Significant discoveries (number of genes with eQTLs) are shown at different permutation thresholds for each tissue. Within each tissue, two independent eQTL analyses were performed after separating related individuals in two subsets (Twin1, Twin2).

Compared to LCL and fat, proportionally less eQTLs were detected in skin, at all levels of significance. This is likely due to lower power in skin, which is a more heterogeneous tissue and consists of a variety of cell-types (Sorrell and Caplan 2004; Leek and Storey 2007).

The MCTA study design allows replication of eQTL discoveries in each tissue. Replication was assessed using the mean value of the proportion of true positives (π_1) (see Methods and (Storey and Tibshirani 2003)) estimated from the exploration of significant eQTLs in the reciprocal co-twin. Specifically, significant SNP - gene combinations discovered in the first co-twin are tested in the second co-twin and the nominal SRC p-value distribution of the same initial associations is analyzed. The reciprocal test (SNP - gene associations discovered in co-twin 2 tested in co-twin 1) is also performed. The enrichment of low p-values from the distribution described above is used to estimate π_1. For each tissue, the mean π_1 of the two reciprocal tests is reported (Table 4.3). The discovered eQTLs appear robust as they replicate well between individuals of the two co-twin groups per tissue, with a mean proportion of true positives from 0.93 for skin to 0.98 for LCL and fat. I also checked the proportion of true positives specifically among the subset of genes that do not replicate in the co-twin at the same threshold. This too is high (π_1 = 0.84 for skin and 0.94 for LCL and fat), suggesting that exact overlap of genes at a given permutation threshold (PT) is an underestimate of eQTL replication due to winner's curse i.e. I see eQTLs in the co-twin that clearly replicate the initial findings, but at higher p-value and thus marginally not meeting the initial discovery threshold.

	SRC analysis			SRC-FA analysis		
	Twin 1	Twin 2	Replication (Mean π_1)	Twin 1	Twin 2	Replication (Mean π_1)
LCL	509	556	0.98	1068	1226	0.97
SKIN	238	231	0.93	534	544	0.95
FAT	462	488	0.98	1054	1072	0.97

Table 4.3. Replication of *cis* eQTL discoveries (number of significant genes per tissue at 10^{-3} permutation threshold). Results from both the Spearman Rank Correlation (SRC) and Factor Analysis (SRC-FA) are presented. Proportion of replicating signals calculated as the mean co-twin π_1 estimates from the p-value distribution of same SNP-gene associations in the reciprocal twin set

4.2 Substantial increase in number of eQTLs per tissue by Factor analysis

The observed variation in gene expression is not entirely due to genetic effects. Experimental noise and environmental conditions also affect transcript levels in a global manner. Therefore, it is desirable to remove the effects of such random variables and thus increase the power to detect eQTLs. For this purpose, factor analysis (FA) was employed on each tissue separately (Stegle, Parts et al. 2010). We corrected for global latent effects on all individuals in each tissue and fitted various parameters such as number of learned factors and proportion of variance explained, in order to maximize for replication of eQTLs per tissue between twin sets (Methods).

After performing standard SRC eQTL analysis on the factor-corrected expression data (SRC-FA), a substantial improvement in eQTL discovery at each of the standard permutation thresholds used was obtained (Table 4.4). The MCTA design is useful as it permits the validation of the new eQTL discoveries in the replication co-twin for each tissue separately. This is essential in order to verify that FA performs as expected by modelling environmental factors and not correcting out a vast proportion of genetic effects. The improvement in eQTL discovery with SRC-FA is considerable (twice as many eQTLs at 10^{-3} PT) and consistent in all three tissues. The high eQTL replication between twin sets persists after FA, with an additional improvement of true positives detection in skin: π_1 = 0.95 (Table 4.3).

	LCL		SKIN		FAT	
Permutation Threshold	Twin 1	Twin 2	Twin 1	Twin 2	Twin 1	Twin 2
10^{-4}	721	828	329	344	690	720
10^{-3}	1064	1220	532	542	1052	1070
10^{-2}	1839	1967	1103	1080	1732	1812

Table 4.4. *Cis* eQTL associations detected with SRC-FA analysis. Number of genes with a significant eQTL is shown for each co-twin analysis per tissue at different permutation thresholds.

I validated the results of the FA correction by investigating the eQTLs resulting from the SRC-FA analysis. As expected, FA recovers the majority of eQTLs discovered with the initial analysis (roughly 90% of LCL and fat and 80% of skin results) and allows the discovery of additional signals (Table 4.5).

	Twin 1			Twin 2		
	Total Std	FA recovered (%)	Total FA	Total Std	FA recovered (%)	Total FA
LCL	509	460 (90.37%)	1064	556	494 (88.85%)	1220
SKIN	238	189 (79.41%)	532	231	188 (81.39%)	542
FAT	462	421 (91.13%)	1052	488	436 (89.34%)	1070

Table 4.5. Recovery of SRC eQTLs (10^{-3} PT gene associations) with factor analysis correction. In each tissue and for both co-twins, 80-90% of eQTLs detected before correction (standard analysis - Std) are recovered with SRC-FA.

The additional eQTLs likely represent real effects that could not be detected initially due to low power. To test this hypothesis, the eQTLs revealed only after FA correction were tested in the uncorrected expression dataset The p-value distribution of the exact same SNP – gene combinations showed a highly significant enrichment of low values (Figure 4.1). In each tissue and for each co-twin subset, the estimated enrichment corresponded to a π_1 value of 0.99. This confirms that the vast majority of new eQTLs are real and would be picked up using the standard SRC pipeline if a larger sample size would be available.

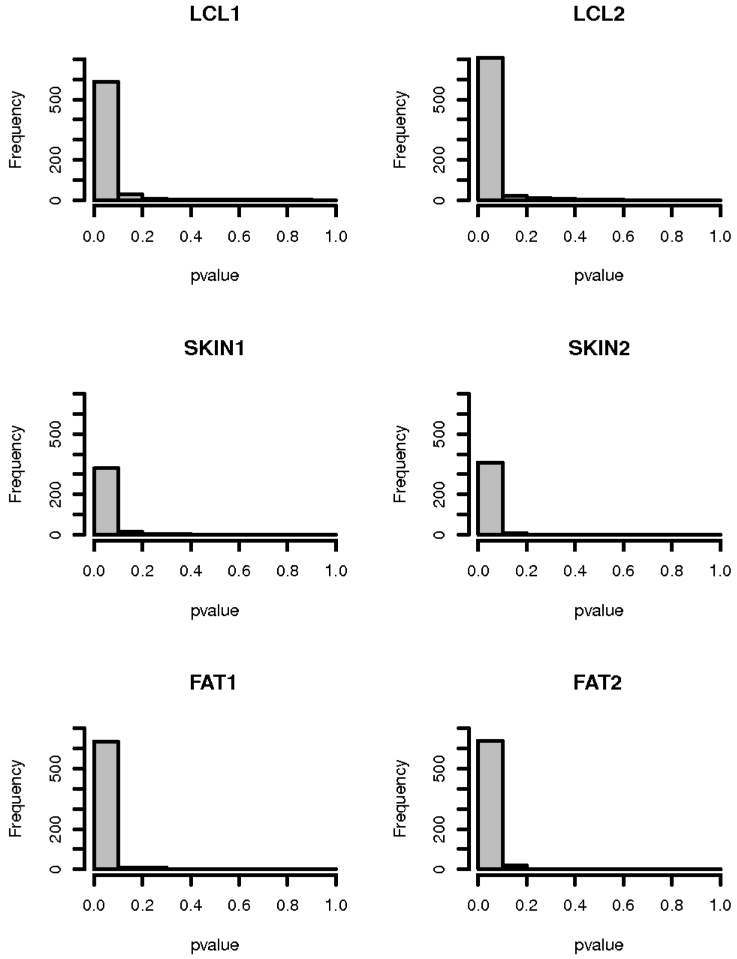


Figure 4.1. P-value distribution of *cis* eQTLs (10^{-3} PT) gained with FA correction in the uncorrected data. The significant overrepresentation of low p-values for the new eQTLs (π_1 = 0.99) shows that the signal existed in the uncorrected data but wasn't called significant due to low power. Result consistent in all tissues for both sets of co-twins (Twin 1 – left panel, Twin 2 – right panel).

4.3 eQTL properties across tissues

The eQTLs (10^{-3} permutation threshold) resulting from both SRC and SRC-FA analyses were compared across all three tissues. Initial direct tissue overlap of significant eQTLs supports an extensive level of tissue-specificity with very similar proportion in both detection methods employed.

A visual representation of the percentages of eQTLs found in only one tissue, shared in only two tissues and common in all three tissues for SRC and SRC-FA respectively can be seen in Figure 4.2.

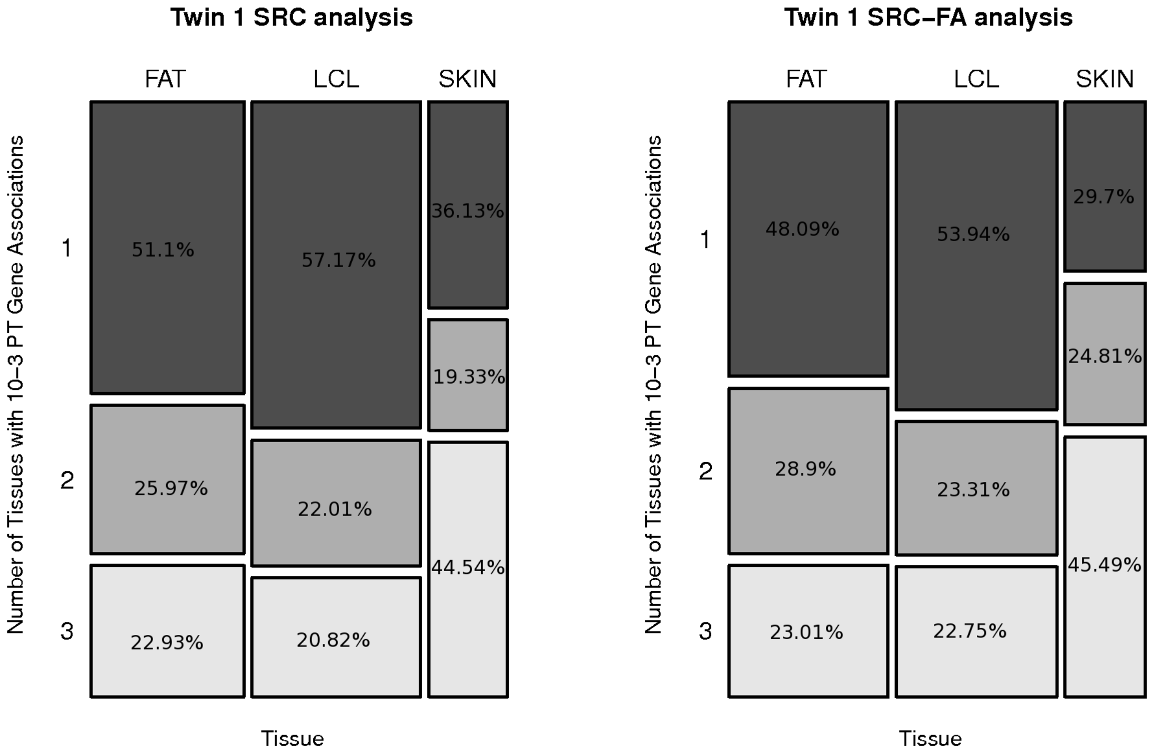


Figure 4.2. Percentage of eQTLs (10^{-3} PT) found only in one tissue, only in two tissues and in all three tissues with the SRC and SRC-FA analysis respectively. Both methods reveal similarly high extents of tissue-specificity. Skin specific eQTLs of smaller effects are harder to detect due to low power.

In the first co-twin set we discover 858 non-redundant eQTL genes at 10^{-3} PT in all three tissues (Table 4.6). Of these, 106 genes (12.35%) are shared across all tissues, 139 (16.2%) are shared in at least two tissues and 613 genes (71.44%) are detected in only one tissue. In skin, where we are least powered likely due to tissue heterogeneity and variety of cell-types, we detect proportionally fewer tissue-specific effects (10.02% of skin eQTLs are only present in skin at 10^{-3} PT).

		Twin 1			Twin 2	
		10^{-3} PT	% total	Overlap	10^{-3} PT	% total
3 tissues	LCL-SKIN-FAT	106	12.35	78	102	11.02
2 tissues only	LCL-SKIN	19	2.21	4	12	1.29
	LCL-FAT	93	10.84	52	107	11.56
	SKIN-FAT	27	3.15	11	26	2.81
1 tissue only	LCL	291	33.92	150	335	36.18
	SKIN	86	10.02	17	91	9.82
	FAT	236	27.5	103	253	27.32
Total significant	LCL	509		363	556	
	SKIN	238		132	231	
	FAT	462		304	488	
Union of total significant		858	100	563	926	100

Table 4.6. Tissue-shared and tissue-specific gene associations (10^{-3} PT), SRC analysis

SRC-FA results confirm the estimated ~30% of eQTLs to be shared in at least two tissues based on threshold eQTL discovery (Table 4.7).

		Twin 1			Twin 2	
		10^{-3} PT	% total	Overlap	10^{-3} PT	% total
3 tissues	LCL-SKIN-FAT	242	13.28	192	270	13.86
2 tissues only	LCL-SKIN	38	2.09	8	42	2.16
	LCL-FAT	210	11.53	84	232	11.91
	SKIN-FAT	94	5.16	28	70	3.59
1 tissue only	LCL	574	31.5	302	676	34.7
	SKIN	158	8.67	51	160	8.21
	FAT	506	27.77	221	498	25.56
Total significant	LCL	1064		781	1220	
	SKIN	532		338	542	
	FAT	1052		735	1070	
Union of total significant		1822	100	1312	1948	100

Table 4.7. Tissue-shared and tissue-specific gene associations (10^{-3} PT), SRC-FA analysis.

In the currently examined three tissues, shared eQTLs show the same allelic direction of effect (Figure 4.3), i.e. if one SNP allele predisposes to increased levels of expression of a gene, it will also tend to elevate the expression level of that gene in the other tissue. This is true for both eQTLs significant at 10^{-3} and 10^{-2} PT.

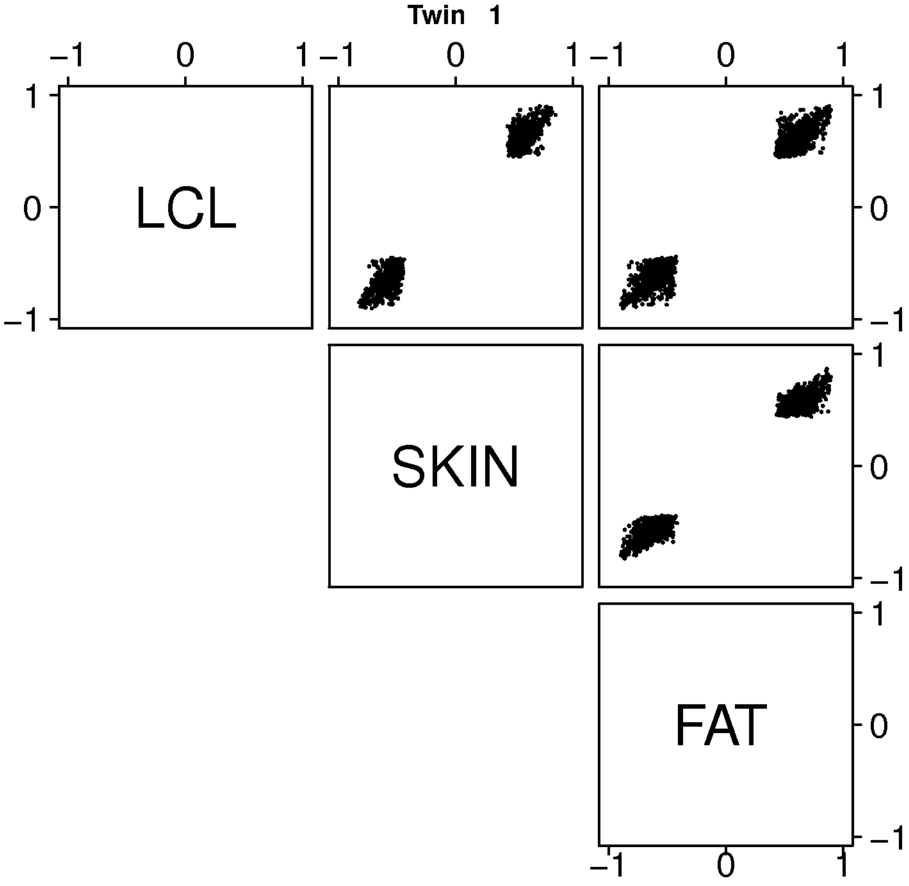


Figure 4.3. Shared eQTLs (10^{-2} PT, SRC) have the same direction of effect (SRC rho) across tissues

As reflected by the SRC correlation coefficient rho (Figure 4.4), eQTLs significant in one tissue explain a substantially higher fraction of gene expression variation in the tissue of discovery than in other tissues (same SNP-gene association), whereas shared effects at the same significance threshold (10^{-3} PT) have comparable variance explained by the SNP across tissues.

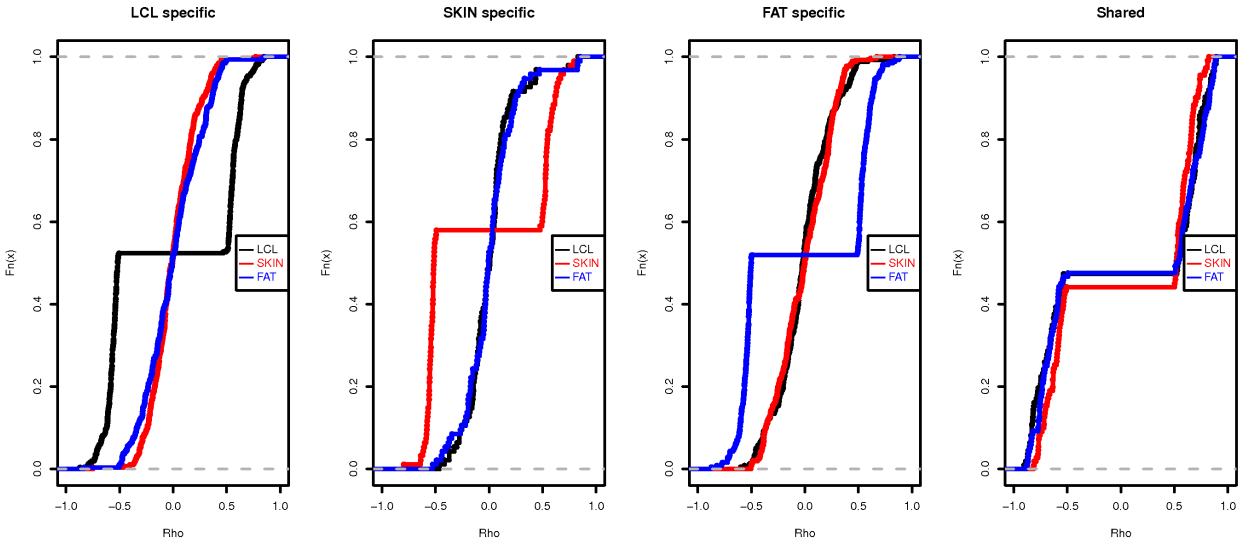


Figure 4.4. Cumulative SRC rho distribution across tissues for tissue-specific and shared eQTLs (10-3 PT, Twin 1). eQTLs discovered in one tissue only have distinctively higher variance in the tissue of discovery compared to shared effects.

In order to refine the expression association signals and describe independently acting eQTLs, I mapped them to recombination hotspot intervals and filtered subsequently by LD (Methods). I observe in all tissues that the majority of genes (90-95%) are controlled by single independent *cis* eQTLs with similar estimates from the standard and factor eQTL analysis. The finer comparison of eQTL effects requiring the sharing of both the gene and the genomic interval harboring the eQTL SNP yields similar counts of shared and specific effects (Table 4.8). The results are similar for SRC-FA. This suggests that the vast majority of shared genes also share regulatory variants across tissues.

		Twin 1			**Twin 2**	
		10^{-3} PT	**% total**	**Overlap**	**10^{-3} PT**	**% total**
3 tissues	LCL-SKIN-FAT	104	10.86	70	96	9.44
2 tissues only	LCL-SKIN	17	1.77	5	14	1.37
	LCL-FAT	90	9.39	49	103	10.13
	SKIN-FAT	30	3.13	12	26	2.56
1 tissue only	LCL	339	35.39	151	374	36.77
	SKIN	101	10.54	18	106	10.42
	FAT	277	28.91	100	298	29.3
Total significant	LCL	550		348	587	
	SKIN	252		128	242	
	FAT	501		302	523	
Union of total significant		958	100	565	1017	100

Table 4.8. Tissue-shared and tissue-specific interval-gene associations (10^{-3} PT), SRC analysis.

Furthermore, the genomic location of the independent eQTLs with respect to basic gene structure landmarks was investigated. Similar results to previous studies are observed (Dimas, Deutsch et al. 2009). As such, eQTLs cluster symmetrically around the TSS, with shared effects distributed more tightly compared to specific ones (Figure 4.5). The broader distribution of cell-type specific effects around the TSS suggests their role on tissue-specific enhancer elements. Independent eQTLs gained with FA correction were also investigated. It was found that they have the same pattern as the SRC eQTLs, supporting furthermore their likely biological role.

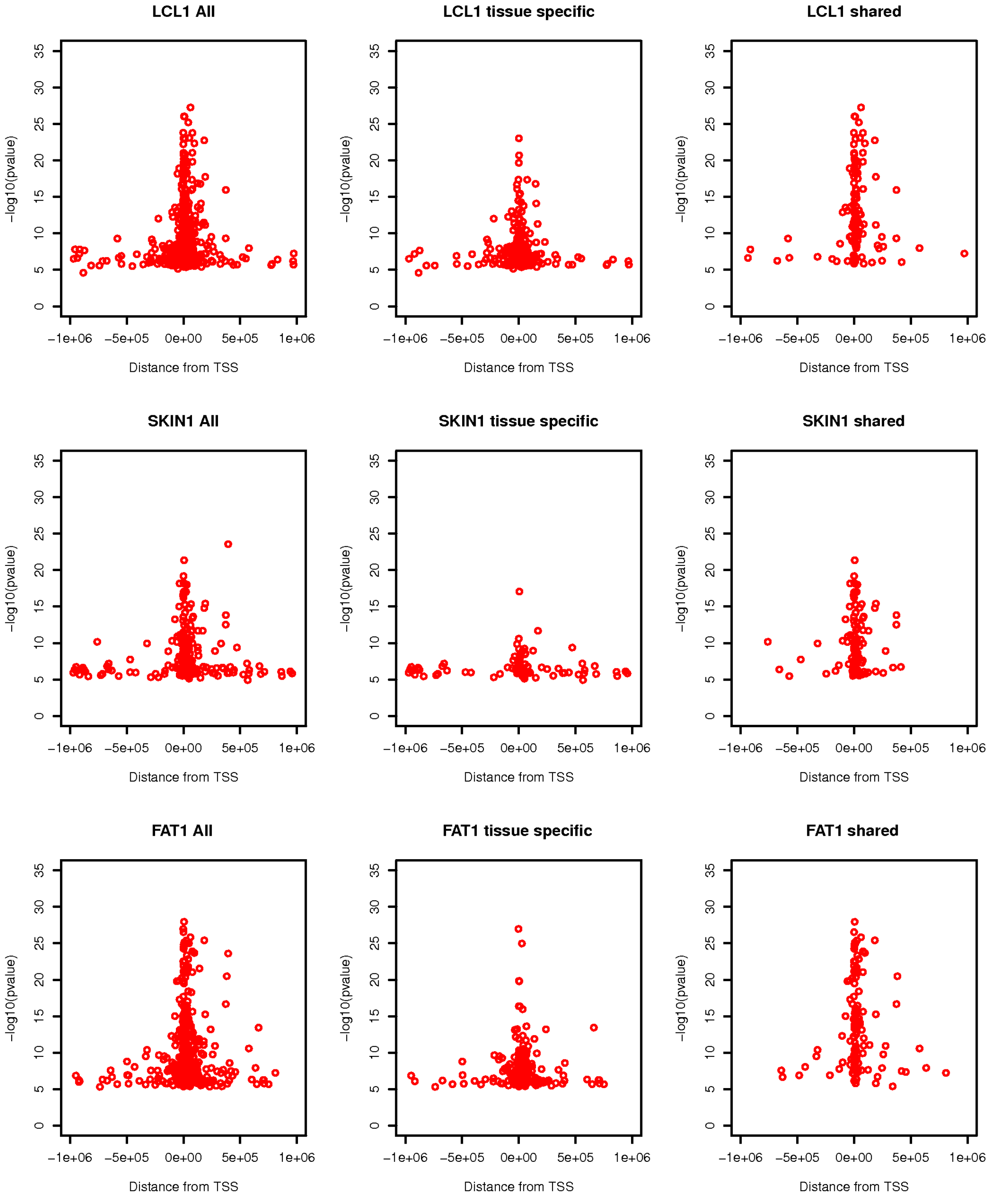


Figure 4.5. Distribution of independent *cis* eQTLs (10^{-3} PT, SRC) around the transcription start site (TSS). Data from co-twin 1 shown here; left panel displays all eQTLs, the middle panel includes only tissue-specific eQTLs while the right panel shows only eQTLs shared across all three tissues. Similar results are obtained for co-twin 2 and the independent eQTLs revealed by the SRC-FA analysis.

4.4 Alternative estimates of eQTL tissue-specificity

Thresholds are driven by statistical significance and overlaps at these levels are heavily dependent on power. In addition, eQTLs sharing statistical significance may still have notable effect differences on gene expression levels across tissues, with potentially different biological consequences. Given these caveats, I examined tissue-specificity in a continuous way using the estimate of significant low p-value enrichment (π_1). More specifically, I investigated the p-value distribution of significant SNP-gene pairs (10^{-3} PT) from a reference tissue in the other two tissues. The p-value distribution in the other two tissues suggests a high degree of tissue sharing (53 to 80%) both with the SRC and SRC-FA, varying slightly depending on the reference tissue in the comparison (Table 4.9). This indicates that we are still underpowered to detect eQTLs of smaller effects that would increase also the previous threshold-based estimates of tissue sharing. In any case, 29% of eQTLs (1-mean π_1) are expected to be exclusively tissue-specific.

#Twin 1			
Reference	**Secondary**	**SRC analysis π_1**	**SRC-FA analysis π_1**
LCL	SKIN	0.67	0.71
	FAT	0.73	0.77
SKIN	LCL	0.77	0.67
	FAT	0.72	0.84
FAT	LCL	0.63	0.72
	SKIN	0.73	0.78
#Twin 2			
Reference	**Secondary**	**SRC analysis π_1**	**SRC-FA analysis π_1**
LCL	SKIN	0.53	0.66
	FAT	0.73	0.75
SKIN	LCL	0.72	0.71
	FAT	0.8	0.84
FAT	LCL	0.69	0.58
	SKIN	0.81	0.76

Table 4.9. Continuous estimates of tissue sharing by enrichment of low p-values (π_1) of reference eQTLs (SNP-genes 10^{-3} PT) in the secondary tissues.

Tissue sharing should not just be the common presence of a statistically significant regulatory effect, but also the similar effect size (fold change in expression) of that variant across tissues. In this respect, I report the fold change as the difference between the gene expression means of the two homozygous genotypic classes. Within the same tissue, the two co-twin sets are only slightly different in their fold change estimates (0.94 Pearson's correlation of fold change between Twin 1 and Twin 2 in LCL, 0.80 in skin and 0.90 in fat – Figure 4.6). This difference in estimated effect size is much more apparent however between tissues (LCL eQTLs have a 0.65 and 0.72 fold change correlation with skin and fat eQTLs respectively). To a large extent, this is due to the tissue-specificity of eQTLs. However, shared eQTLs at the same threshold of significance don't always share the same effect size across tissues, suggesting additional possible hidden tissue-specific effects (LCL fold change correlation of 0.72 in skin and 0.77 in fat for shared eQTLs i.e. 20% difference in fold change magnitude between tissues compared to within tissue difference). This suggests that even statistically tissue shared eQTLs have additional dimensions of tissue-specificity and their mere discovery in multiple tissues does not guarantee similar magnitude of consequences.

The extent of these observations remains to be tested in *trans* in the better-powered full MuTHER dataset (N ~ 800 individuals). Here, an extension of the MCTA design will be most valuable. Building co-expression networks for each tissue will allow the discovery of tissue-specific modules, which combined with genotypic information could uncover further aspects of tissue-specific regulatory control. The topologies of the networks resulting from such approaches are however highly dependent on the methods and parameters used. Therefore, cross-validating the network predictions with the reciprocal co-twin will ensure that only genetically-relevant gene expression modules are compared.

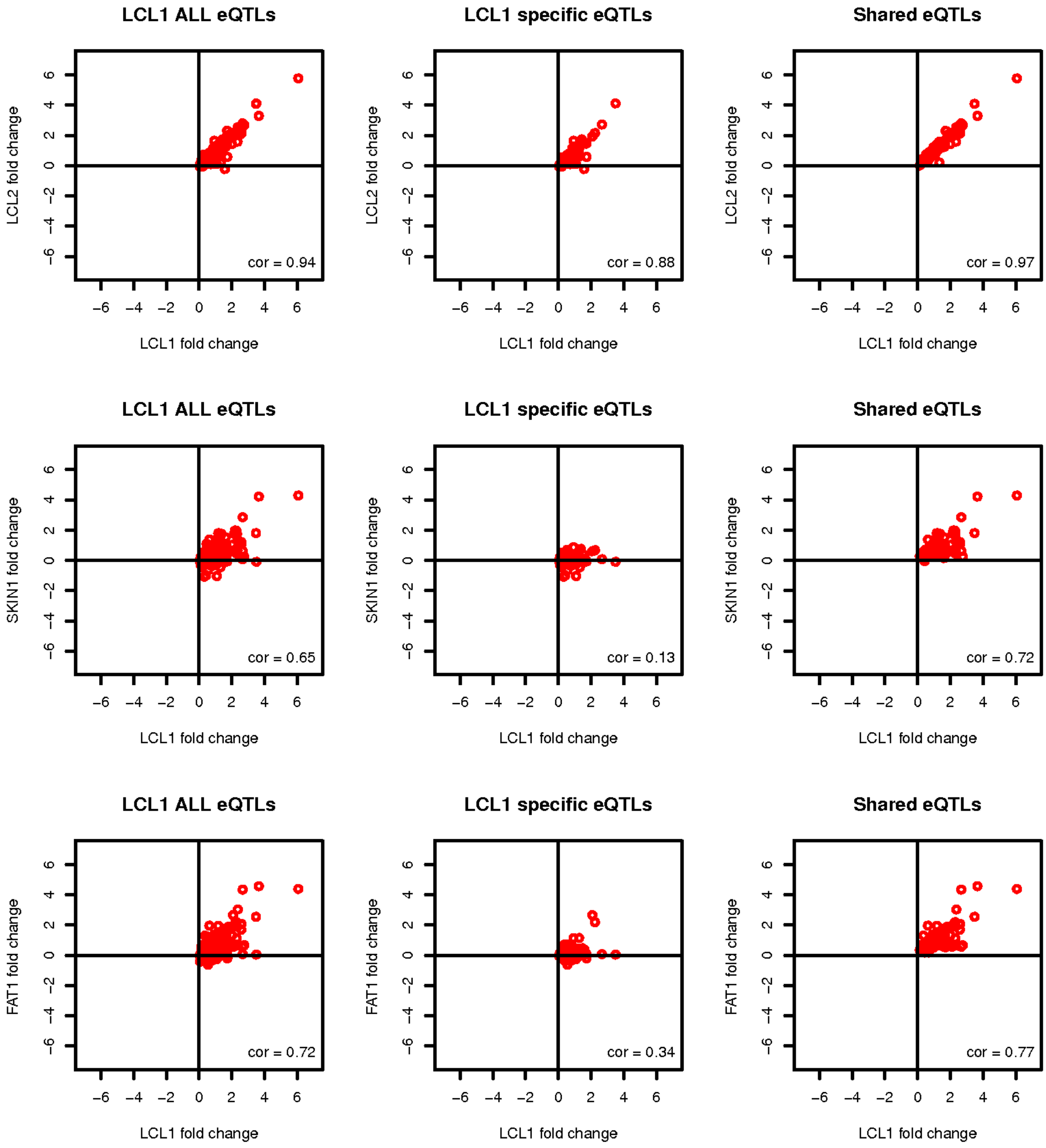


Figure 4.6. Fold change within twins and across tissues for LCL eQTLs (10^{-3} PT, SRC) discovered in Twin 1. The plotted fold change on the X and Y-axes was calculated as the difference in mean expression of homozygous genotypic classes. For each pairwise tissue comparison, the Pearson's correlation coefficient between fold changes is shown.

4.5 Conclusions

While there have been studies exploring regulatory variation in one or more tissues, the complexity of tissue-specificity in multiple primary tissues is not yet well understood. In this chapter, I explored in depth the role of regulatory variation in three human tissues: LCL, skin, and fat. The samples (156 LCL, 160 skin, 166 fat) were derived simultaneously from a subset of well-phenotyped healthy female twins of the MuTHER resource. An abundance of eQTLs in each tissue was discovered, similar to previous estimates (858 or 4.7% of genes). In addition, factor analysis (FA) was applied by removing effects of latent variables, increasing the power by at least 2-fold (1822 eQTL genes). The unique study design (Matched Co-Twin Analysis – MCTA) permits immediate replication of eQTLs with co-twins (93-98%) and validation of the considerable gain in eQTL discovery after FA correction. It was observed that the majority (>90%) of genes are regulated by single independent eQTLs with shared direction of effect across different tissues and their spatial distribution around basic gene structure landmarks was described. I highlight the challenges of comparing eQTLs between tissues and after verifying previous significance threshold-based estimates of extensive tissue-specificity, I show their limitations given their dependency on statistical power. Instead, I propose that continuous estimates of statistical significance and direct comparison of the magnitude of effect on the fold change in expression are essential properties that jointly provide a biologically realistic view of tissue-specificity. Under this framework, this study shows that 30% of eQTLs are shared among tissues, while another 29% are likely exclusively tissue-specific. However, even among the shared eQTLs a substantial proportion (10-20%) have significant differences in the magnitude of fold change between homozygote classes across tissues. These results underline the need to account for the complexity of eQTL tissue-specificity in an effort to assess consequences of such variants for complex traits.

5 Tissue-dependent causal regulatory effects

Gene expression studies performed so far on multiple human tissues support an extensive level of eQTL tissue-specificity. Comparisons across dissimilar cell-types documented first the considerable tissue dependency of *cis* regulatory variation. As such, overlaying LCL and cortical tissue eQTLs resulted in barely any overlap (Myers, Gibbs et al. 2007), the comparison of adipose and blood expression patterns in two Icelandic cohorts reported that 50% of the detected *cis* eQTLs were shared (Emilsson, Thorleifsson et al. 2008) and a study overlapping eQTLs from autopsy-derived cortical tissue and peripheral blood mononucleated cells revealed less than 50% sharing (Heinzen, Ge et al. 2008). In the previous chapter (Chapter 4), I further explored the complexity of eQTL tissue-specificity in LCLs, skin and fat, estimating a relatively low proportion of shared eQTL effects (~30%). Moreover, eQTLs do not display significant tissue-specific properties only among cell-types with substantially different cellular functions. The study from our lab overlapping regulatory variants in transformed B-cells (LCLs), fibroblasts and primary T-cells provided evidence that even cell-types as closely related as B-cells and T-cells share only a minority of *cis* eQTLs (<15%) (Dimas, Deutsch et al. 2009).

A similarly high emphasis on tissue-dependency has been put in the context of other complex traits, including disease phenotypes. Different diseases manifest themselves in different organs and have different tissues as primary targets of pathology (connective tissue diseases, muscle diseases, etc.). However, the tissues where diseases are manifested are not necessarily informative of the cell-type where the causal mechanism leading to disease progression occurs. Perhaps one of the best illustrative examples in this sense is the one of MC4R (melanocortin-4 receptor), a gene in the vicinity of which several common variants associated with fat mass, weight and obesity risk were discovered (Loos, Lindgren et al. 2008). Rare functional mutations in MC4R, known to cause monogenic severe childhood-onset obesity (Vaisse, Clement et al. 1998; Yeo, Farooqi et al. 1998) and further functional evidence from murine models (Huszar, Lynch et al. 1997) indicated that MC4R is also responsible for common obesity. Most likely, the susceptibility loci act by disrupting the expression of MC4R. However, this hypothesis remains hard to verify, given that MC4R is almost exclusively expressed in the brain and found at very low levels in most of the currently available expression datasets. The brain specific expression pattern of an obesity susceptibility gene is a clear example of the fact that predicting a tissue's relevance to disease is complicated by the discrepancy between the place of its manifestation and where the causal mechanism initiates.

In this chapter, I explore the role of tissue-dependency in predicting causal regulatory effects for GWAS loci. Specifically, I apply the RTC methodology described in Chapters 2 and 3 on the three cell-types (B cells, T cells and fibroblasts) available from the GenCord (because of time and availability constraints, this analysis was performed on the GenCord rather than the MuTHER resource, which is also of very high interest and constitutes a future project in the lab; see Methods). I show that finding regulatory variants and corresponding differentially expressed genes underlying complex disease associations is highly dependent on the nature of the cell-type tested. The results confirm previously suspected candidate loci and offer new functional insights into disease aetiology by revealing novel differentially regulated susceptibility genes.

5.1 RTC score distribution by tissue

To detect likely causal regulatory effects for GWAS signals across multiple tissues I used expression and genotypic data derived from the 75 GenCord European individuals. The ~400,000 genotyped SNPs were imputed first on HapMap 2 in order to increase power to detect associations with expression. After imputation and QC filtering (see Methods), 1,428,314 SNPs with MAF > 5% were available for analysis. Transcript level measurements in transformed B-cells, fibroblasts and primary T-cells for probes mapping uniquely to 15,596 Ensembl genes (Methods) were tested for association with SNP genotypes. I considered SNP–gene associations with a nominal SRC p-value < 10^{-4} as eQTLs. This nominal threshold corresponds roughly to a 0.05 permutation threshold (estimation from multiple eQTL analyses on different datasets run in our group). At this significance level, 1139 genes were detected to have at least one eQTL in B-cells, 1098 genes in T-cells and 1157 genes in fibroblasts. I overlapped these sets of eQTLs with GWAS results in each tissue separately. For this purpose, I revisited the NHGRI GWAS catalogue (Hindorff, Sethupathy et al. 2009) and downloaded the most recent list at that time of SNPs associated with complex traits (accessed 12.04.2010). 1750 SNPs were retrieved, associated with 248 traits. I further mapped both the GenCord eQTLs and NHGRI GWAS SNPs to recombination hotspot intervals (McVean, Myers et al. 2004) as described in Chapter 2. The 1750 GWAS SNPs mapped to 1400 intervals, suggesting that some intervals harbour multiple susceptibility loci often associated with multiple traits, as in the case of the well-known 8q24 gene desert region, where variant associations with breast, prostate and colorectal cancer have been reported (Ghoussaini, Song et al. 2008; Al Olama, Kote-Jarai et al. 2009). The overlap of recombination hotspot intervals where at least one GWAS SNP and one eQTL co-localize resulted in 106, 111 and 105 intervals in B-cells, T-cells and fibroblasts respectively. A subset of these intervals contained more than one disease GWAS locus. Each interval was tested in *cis* under the RTC framework (see Methods) for every disease association reported by NHGRI (in total 149, 150 and 144 interval-disease combinations were tested in B-cells, T-cells and fibroblasts – Table 5.1).

	#Intervals	#Interval-diseases
B-cells	106	149
T-cells	111	150
Fibro	105	144
Shared	26	53

Table 5.1. Number of nonredundant recombination hotspot intervals with co-localizing GenCord eQTLs and GWAS SNPs. Unique count of intervals (#Intervals) and interval-disease combinations (#Intervals with multiple disease GWAS loci - #Interval-diseases) per tissue tested for causal regulatory effects with the RTC. A subset of intervals (Shared) harbour eQTLs detected in all three tissues.

In each of the three tissues, an overrepresentation of GWAS regulatory candidates with high RTC scores is observed (Figure 5.1). I detect SNP-gene associations passing the 0.9 RTC threshold for 41 interval-disease combinations of the total 149 tested in B-cells, 36 high scoring interval-disease combinations of 150 tested in T-cells and 29 out of 144 tested in fibroblasts. The overall distribution of scores differs across tissues, with a more noticeable similarity between B-cells and T-cells. This similarity between the two cell-types in contrast to the different pattern of RTC scores in fibroblasts is even more apparent when focusing on the shared subset of 53 interval-disease combinations tested in all three tissues (Figure 5.2).

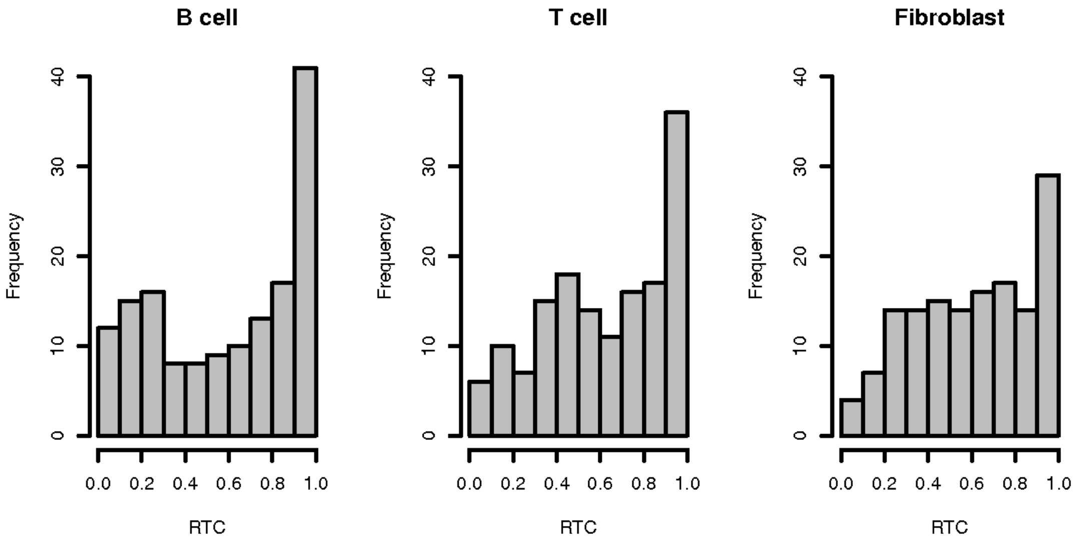


Figure 5.1. Distribution of RTC scores across tissues. The best score per GWAS SNP per interval-disease combination is plotted for all tested intervals. An enrichment of high RTC scoring candidates (≥ 0.9) is present in each tissue, corresponding to the subset of GWAS SNPs likely explained by regulatory effects.

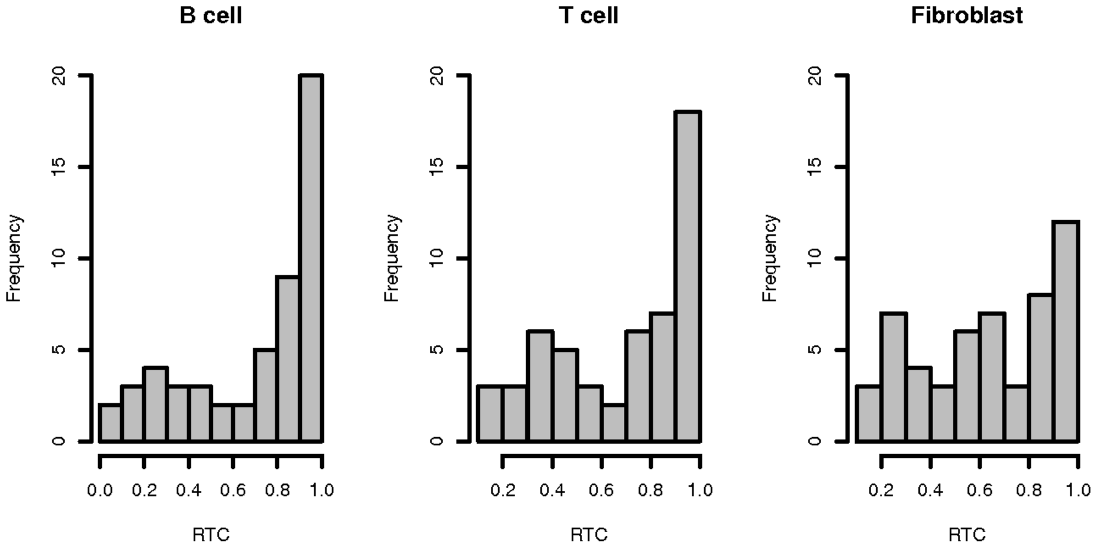


Figure 5.2. Distribution of RTC scores across tissues for shared intervals-disease combinations. The best score per GWAS SNP per interval-disease combination is plotted only for the 53 shared interval-disease combinations tested in all three tissues. A marked difference in distribution of scores can be observed, especially between fibroblasts and the more functionally similar B-cells and T-cells.

The differential RTC score distribution across the three tissues reflects the biological similarities of the tested cell-types. B-cells and T-cells are two classes of lymphocytes, the white blood cells involved in the body's adaptive immune response (Alberts 2002). Fibroblasts on the other hand, are a type of connective tissue cells secreting the extracellular matrix and collagen and playing an essential role in wound healing. Nevertheless, despite being much more functionally related, B-cells and T-cells carry out two distinct types of immune responses whose disruption may thus have different phenotypic consequences. B-cells participate in the antibody immune response whereby following their activation by foreign antigens, they secrete antibodies. These can then circulate in the bloodstream, bind the antigens that stimulated their production and thus inhibit the detrimental action of viruses or microbial toxins on the cell. In most mammals, B-cells are produced in the bone marrow (Alley 1987). T-cells, the other major class of lymphocytes, are produced in the thymus and are responsible for cell-mediated immune responses (Spits 2002). After their activation, T-cells react directly against a foreign antigen, for example by killing a virus-infected host cell displaying the respective antigens on its surface. T-cells also assist other cells in immunologic processes such as macrophage activation or differentiation of B-cells into plasma cells (McHeyzer-Williams, Pelletier et al. 2009) .

Our group documented substantial differences in regulation of gene expression in the three tissues, the authors finding that 69 to 80% of the genetic regulatory effects are cell-type specific (Dimas, Deutsch et al. 2009). RTC results on the same dataset reflect the high extent of *cis* eQTL tissue-specificity (Table 5.2). As such, of the total 78 nonredundant interval-disease combinations with confident evidence of causal regulatory effects (RTC score ≥ 0.9), only 5 (6.4%) were shared across all three tissues. The pairwise tissue overlap of RTC results mirrors as expected the biological properties of the cell-types compared. Specifically, B-cells and T-cells share more GWAS relevant regulatory effects (16.7% of the total) compared to any of the other pairwise combinations (2 shared intervals between B-cells and fibroblasts and 3 examples common to T-cells and fibroblasts). Most of the confident RTC results (around 70%) are cell-type specific. This shows that as predicted, detecting causal regulatory effects for complex trait associations is highly tissue-dependent.

		#Interval-Disease (RTC ≥ 0.9)	% Total
3 tissues	B cell - T cell - Fibro	5	6.4
2 tissues only	B cell - T cell	13	16.7
	B cell - Fibro	2	2.6
	T cell - Fibro	3	3.8
1 tissue only	B cell	21	26.9
	T cell	15	19.2
	Fibro	19	24.3
Total RTC ≥ 0.9	B cell	41	
	T cell	36	
	Fibro	29	
Union of total RTC ≥ 0.9		78	

Table 5.2. Tissue shared and tissue-specific interval-disease combinations with high RTC score. RTC results with a score ≥ 0.9 are overlapped and compared across tissues. The significant tissue-specific component (~70% of results are found in one tissue only) implies that predicting causal regulatory effects for GWAS loci is highly dependent on the tissue where expression is determined.

In the following sections, I present the best RTC results for each tissue and focus on some interesting biological examples.

5.2 B-cell results

Table 5.3 summarizes the most confident *cis* results in B-cells ordered by RTC score. I detect SNP-gene combinations passing the 0.9 RTC threshold for 41 interval-disease combinations of the 149 tested. Among the 41 confident signals, 21 (51%) are only found in B-cells at this score level. Candidate genes already suspected to have a role in disease susceptibility are confirmed by the RTC and additional candidates revealed. Within the same recombination hotspot interval on chromosome 8 (chr8:11374002-11504000), two significant genome-wide SNP associations exist, one with systemic lupus erythematosus (Hom, Graham et al. 2008)(rs13277113) and the other one with rheumatoid arthritis (rs2736340) (Gregersen, Amos et al. 2009). The C8orf13-BLK locus

GWAS SNP	Complex Trait	Gene	RTC	Chr
rs10903129	Cholesterol, total	TMEM50A	0.96	1
rs7512898	Electrocardiographic conduction measures	TNNT2	0.91	1
rs13160562	Alcohol dependence	ERAP1	1	5
rs7731657	Fasting plasma glucose	CDC42SE2	0.96	5
rs9272346	Type 1 diabetes	HLA-DQB1	1	6
rs3135388	Multiple sclerosis	HLA-DRB5	1	6
rs2517713	Nasopharyngeal carcinoma	HLA-A	0.99	6
rs9272219	Schizophrenia	HLA-DQA2	0.99	6
rs3129934	Multiple sclerosis	HLA-DRB5	0.98	6
rs8321	AIDS progression	HLA-G	0.98	6
rs2227139	Hematological parameters	HLA-DRB5	0.97	6
rs2523393	Multiple sclerosis	TRIM27	0.96	6
rs9268480	Ulcerative colitis	HLA-DQB1	0.95	6
rs9264942	HIV-1 control	NFKBIL1	0.95	6
rs2187668	Systemic lupus erythematosus	HLA-DQA2	0.94	6
rs2269426	Plasma eosinophil count	HLA-DQA2	0.93	6
rs13437082	Height	C6orf48	0.93	6
rs7743761	Ankylosing spondylitis	HLA-C	0.93	6
rs9461688	Protein quantitative trait loci	C6orf48	0.90	6
rs2237349	Attention deficit hyperactivity disorder	CREB5	0.99	7
rs17145738	Triglycerides	BCL7B	0.94	7
rs13277113	Systemic lupus erythematosus	C8orf13	1	8
rs2736340	Rheumatoid arthritis	C8orf13	1	8
rs216345	Bipolar disorder	NUDT2	1	9
rs10781500	Ulcerative colitis	CARD9	1	9
rs4130590	Bipolar disorder	SLC2A8	0.99	9
rs7871764	Height	NUDT2	0.94	9
rs1927702	Body mass index	C9orf52	0.92	9
rs4977574	Myocardial infarction (early onset)	CDKN2A	0.91	9
rs7481311	Weight	LIN7C	1	11
rs5215	Type 2 diabetes	C11orf58	0.96	11
rs11602954	Mean platelet volume	ATHL1	0.95	11
rs11171739	Type 1 diabetes	RPS26	0.98	12
rs8020441	Cognitive performance	ATP5S	1	14
rs10133111	Brain imaging in schizophrenia (interaction)	HSP90AA2	0.96	14
rs748404	Lung cancer	TGM5	0.93	15
rs2290400	Type 1 diabetes	GSDML	0.97	17
rs199533	Parkinsons disease	ILMN_25445	0.93	17
rs2014572	Hyperactive-impulsive symptoms	VN1R1	0.92	19
rs6060369	Height	UQCC	0.99	20
rs5751901	Protein quantitative trait loci	GGT4P	1	22

Table 5.3. Candidate B-cell results. Candidate genes (RTC score ≥ 0.9) for *cis* regulatory mediated GWAS effects. RTC applied on NHGRI GWAS SNPs and B-cell expression data from the GenCord.

has already been associated to lupus and confirmed in the initial RTC analysis on LCLs derived from HapMap 3 CEU individuals (Nica, Montgomery et al. 2010). The rheumatoid arthritis SNP identified more recently, scores equally high (RTC = 1) with the same gene of unknown function, *C8orf13*. Furthermore, the two SNPs are in very high LD ($r^2 = 0.95$, D' = 0.99) suggesting that most likely they are tagging the same functional variant, which plays an important role in the two autoimmune diseases.

The RTC applied on GenCord B-cells recovers another suspected autoimmune disease effect, namely the implication of *CARD9* (caspase recruitment domain family, member 9) in ulcerative colitis risk (Zhernakova, Festen et al. 2008). The GWAS SNP rs10781500 on chromosome 9 scores best (RTC = 1) with this gene, which is a very plausible candidate for inflammatory bowel diseases (including ulcerative colitis). *CARD9* was shown to be essential in the process of stimulating the innate immune signalling by intracellular and extracellular pathogens (Underhill and Shimada 2007). Studies in mice documented the role of *CARD9* in contributing to cytokine production via MAPK activation (Hsu, Zhang et al. 2007) or alternatively, leading to NF-κβ activation through the syk-CARD9 interaction (Hara, Ishihara et al. 2007). Given these, disrupting *CARD9* signalling by modifying the gene's expression levels appears to be a very probable mechanism leading to a deficient immune response predisposing to disease. Nevertheless, the authors of the GWAS study highlight that the extended 120 kb haplotype where the susceptibility SNP resides includes in addition to *CARD9* a few other genes that cannot be confidently excluded, namely *GPSM1*, *PSM1*, *LOC728489*, *SNAPC4*, *SDCCAG3*, *PMPCA*, *INPP5E* and *KIAA0310* (Zhernakova, Festen et al. 2008). Interestingly, for one of these genes, *INPP5E* (inositol polyphosphate-5-phosphatase), the RTC method provides supporting evidence in T-cells (RTC score = 0.9) and a modest effect in fibroblasts (RTC score = 0.76) for the same hotspot interval (chr9:138377986-138526984) where the GWAS SNP rs10781500 resides. *INPP5E* has also been shown to mediate cell responses to various stimulations (Kong, Speed et al. 2000). Both genes at the 9q34.3 locus would therefore merit further investigation for determining the causative roles behind the disease association.

The association of the *UQCC* (ubiquinol-cytochrome c reductase complex chaperone) locus on chromosome 20 with human height is one of the most robust signals for this trait, repeatedly replicated across multiple studies (Gudbjartsson, Walters et al. 2008;

Sanna, Jackson et al. 2008; Soranzo, Rivadeneira et al. 2009). The gene encodes a transmembrane protein and evidence from studies in mouse embryonic stem cells shows that the gene is down regulated in the presence of FGF2 (Vetter and Wurst 2001), which acts together with bone morphogenic proteins and *Hox* gene products to initiate and promote skeleton growth (Sanna, Jackson et al. 2008). The RTC successfully recovers the causal regulatory effect at this convincing gene locus with a score of 0.99 in B-cells only.

5.3 T-cell results

Using gene expression data from T-cells, I detect SNP-gene associations passing the 0.9 RTC score threshold for 36 interval-gene combinations of the 150 tested (discoveries sorted by RTC are presented in Table 5.4). Of these, results for 15 recombination hotspot intervals (41.7%) are restricted to T-cells, denoting once more the important role of the tissue-type in detecting disease relevant regulatory effects. As expected, the RTC method reveals candidate genes for a variety of autoimmune conditions, but also other interesting traits, some already flagged in the literature.

A recent GWAS identified significant associations at the 5q31 locus with osteoporosis risk (Guo, Tan et al. 2010). The strongest associated SNP, rs13182402 maps within the gene *ALDH7A1* (aldehyde dehydrogenase 7 family, member A1), that plays a major role in degrading and detoxifying acetaldehyde generated by alcohol metabolism and lipid peroxidation. Acetaldehyde was shown to inhibit osteoblast proliferation and result in decreased bone formation in murine and human bone marrow cultures (Giuliani, Girasole et al. 1999), making it thus a plausible candidate for explaining the disease association. We provide further evidence for this hypothesis by detecting a strong likely causal regulatory effect (RTC score = 1) for the same GWAS SNP with *ALDH7A1*. The effect is only detectable in T-cells and it suggests that rs13182402 predisposes to osteoporosis by substantially affecting the expression of *ALDH7A1*.

Similarly, the RTC recovers and confirms the implication of another suspected susceptibility gene, *GPR22* (G protein-coupled receptor 22) in osteoarthritis (Kerkhof, Lories et al. 2010). The most significant associated GWAS SNP, rs3815148 maps in an intronic region of a very large gene, *COG5* (component of oligomeric golgi complex 5) spanning ~3.6 Mb on the reverse strand of chromosome 7. However, the same SNP is

GWAS SNP	Complex Trait	Gene	RTC	Chr
rs3890745	Rheumatoid arthritis	MMEL1	0.98	1
rs6435862	Neuroblastoma (high-risk)	BARD1	0.97	2
rs10495928	Hemoglobin	SOCS5	0.91	2
rs3772130	Cognitive performance	IQCB1	1	3
rs13182402	Osteoporosis	ALDH7A1	1	5
rs7731657	Fasting plasma glucose	CDC42SE2	0.99	5
rs13160562	Alcohol dependence	ERAP1	0.91	5
rs1321311	Electrocardiographic traits	FGD2	1	6
rs3135388	Multiple sclerosis	HLA-DRB5	1	6
rs2517713	Nasopharyngeal carcinoma	HLA-G	1	6
rs9272346	Type 1 diabetes	HLA-DQB1	0.99	6
rs9264942	HIV-1 control	EHMT2	0.99	6
rs3129934	Multiple sclerosis	HLA-DRB5	0.98	6
rs13437082	Height	HLA-C	0.97	6
rs2523393	Multiple sclerosis	HLA-G	0.97	6
rs10484554	Psoriasis	HLA-C	0.96	6
rs13194053	Schizophrenia	BTN3A2	0.96	6
rs9461688	Protein quantitative trait loci	HLA-C	0.96	6
rs12216125	Serum markers of iron status	BTN3A2	0.95	6
rs9268480	Ulcerative colitis	HLA-DQB1	0.95	6
rs2227139	Hematological parameters	HLA-DRB5	0.94	6
rs8321	AIDS progression	HLA-A	0.92	6
rs741301	Diabetic nephropathy	GPR141	0.98	7
rs3815148	Osteoarthritis	GPR22	0.98	7
rs4130590	Bipolar disorder	SH2D3C	0.97	9
rs1927702	Body mass index	BNC2	0.93	9
rs10781500	Ulcerative colitis	INPP5E	0.90	9
rs703842	Multiple sclerosis	FAM119B	1	12
rs11171739	Type 1 diabetes	RPS26	0.98	12
rs1402279	Smoking behavior	OSBPL8	0.92	12
rs1378942	Diastolic blood pressure	C15orf39	0.91	15
rs4785763	Melanoma	CDK10	0.99	16
rs11648785	Tanning	CDK10	0.98	16
rs199533	Parkinsons disease	NSF	0.91	17
rs8099917	Response to Hepatitis C treatment	PSMC4	0.95	19
rs5751614	Height	BCR	0.97	22

Table 5.4. Candidate T-cell results. Candidate genes (RTC score ≥ 0.9) for *cis* regulatory mediated GWAS effects. RTC applied on NHGRI GWAS SNPs and T-cell expression data from the GenCord.

associated with differential expression levels of a nearby gene on the forward strand encoding a G protein-coupled receptor (*GPR22*). The RTC framework provides compelling evidence that the GWAS SNP acts by modulating *GPR22* expression, as revealed by the 0.98 RTC score on the interval tested (chr7: 106574287-107092285). Results from experimental work in mouse models strengthen the potential role of *GPR22* in osteoarthritis pathology. For example, immunohistochemistry experiments showed that the GPR22 protein was absent in normal mouse articular cartilage or synovium, but GPR22-positive chondrocytes (the only cell-type found in cartilage) were detected in osteophytes (bony projections usually formed along joints) in instability-induced osteoarthritis and in the upper layers of the articular cartilage of mouse knee joints challenged with papain or albumin treatment (Kerkhof, Lories et al. 2010). All these lines of evidence point towards *GPR22* as having a causative disease role rather than *COG5*, the gene where the GWAS SNP resides.

Finally, another interesting result of the RTC analysis is the one linking the *FAM119B* (family with sequence similarity 119, member B) gene to multiple sclerosis susceptibility. Multiple sclerosis is an immune-mediated disorder whereby the body's own immune system attacks and damages the myelin sheaths around the axons in the brain and spinal cord. This severe disease of the central nervous system is characterized by myelin loss, chronic inflammation, axonal and oligodendrocyte pathology, and progressive neurological dysfunction (Oksenberg and Baranzini 2010). Although mechanisms involved in the disease process are relatively well described, very little is known about what causes the disease. Recently, a handful of GWAS studies revealed a subset of well-replicating associations, which remain largely elusive with respect to the genes whose activity they disrupt. One significant risk associated locus has been found on chromosome 12q13–14 in a gene-dense region of very high LD (2009). The most significant SNP in this region (rs703842) maps to the 3' UTR of the *METTL1* (methyltransferase-like protein 1) gene, 1.76 kb upstream of the *CYP27B1* (cytochrome P450 family 27 subfamily B) gene. While not being able to confidently exclude the other genes in the 12q13-14 haplotype (17 genes in total), the authors propose *CYP27B1* as the strongest causative candidate based on current genetic, immunological and epidemiological evidence. *CYP27B1* encodes an enzyme which hydroxylates 25-hydroxyvitamin D into its bioactive form, 1,25$(OH)_2$D. This, along with the vitamin D endocrine system, have been shown to play an important role in the prevention of disease onset and progression of autoimmune conditions modelled in the mouse (Lemire and Archer 1991). Furthermore, the link between vitamin D deficiency and increased multiple sclerosis incidence (van der Mei, Ponsonby et al. 2007) as well as the association between common variants in *CYP27B1* and risk of Type 1 diabetes (Bailey, Cooper et al. 2007) - also an autoimmune condition - advocate for the functional role of this gene in disease. With the RTC method, an additional candidate susceptibility gene in the 12q13-14 haplotype is discovered. Having a maximum RTC score with the same risk SNP rs703842 (RTC = 1) both in T-cells and fibroblasts, *FAM119B* appears as a noteworthy likely causal gene with a regulatory effect. Unfortunately, little is known about the function of this gene. However, a recent study analyzing the whole blood mRNA transcriptome of 99 untreated multiple sclerosis patients supports the RTC discovery (Gandhi, McKay et al. 2010). The authors find evidence for specific dysregulation of T-cell pathways in the trait pathogenesis. Of the 17 genes at the 12q13-14 locus, they can quantify expression data in leukocytes for 13 genes and one of them, *FAM119B* is expressed at significantly lower levels in the susceptibility haplotype (P-value < 10^{-14}). This is yet another example of a non-intuitive disease susceptibility candidate, as the gene where the risk GWAS SNP maps to has no functional relevance to the trait. Instead, the GWAS SNP affects another proximal gene and the RTC methodology is helpful in discerning between such cases.

5.4 Fibroblast results

Finally, fibroblast expression data was also used to test for potential explanatory disease effects via regulatory mechanisms. Of the 144 recombination hotspot intervals harbouring nonredundant disease risk loci, SNP-gene associations passing the 0.9 RTC threshold were detected for 29 interval-disease combinations. Of these, 19 (65%) were confined to this tissue, reiterating the RTC tissue-specificity observed previously also in B-cells and T-cells. The most confident discoveries, sorted by RTC are listed in Table 5.5.

Notably, two signals for multiple sclerosis score a maximum RTC of 1 in this cell-type. One of the two (rs703842 associated with *FAM119B* on chromosome 12q13-14) has been described in the previous section as a plausible candidate supported additionally by recent evidence from the analysis of T-cell expression in untreated individuals with the disease. The common high score (RTC = 1) attributed to the same SNP-gene association in both T-cells and fibroblasts indicates the likelihood of potentially shared disease-relevant biological properties between the two tissues. Interestingly, both leukocytes and fibroblasts produce type I interferons, ubiquitous cytokines released following exposure to a stimulus (pathogen or tumour cell) to trigger an immune response (Meyer 2009). More specifically, fibroblasts produce interferon (INF) beta, a type I interferon that has been used extensively over the past decades as an effective first-line therapy against relapsing-remitting multiple sclerosis (Zhang and Markovic-Plese 2010). This common function of interferons - stimulating or inhibiting a variety of genes involved in immunity-related conditions - might partially explain the surprising usefulness of fibroblast expression in explaining associations with multiple sclerosis.

The SNP rs744166 on chromosome 17q21.1 maps to the first intron of the *STAT3* (signal transducer and activator of transcription 3) gene and has been associated to multiple sclerosis by studying a high-risk isolated Finnish population (Jakkula, Leppa et al. 2010). The role of *STAT3* in disease predisposition seems very likely, given its suspected implication in another autoimmune disorder (Crohn's disease (Barrett, Hansoul et al. 2008)) and the evidence from mouse studies where targeted deletion of the gene in CD4+ T-cells prevented the development of experimental autoimmune encephalomyelitis, the murine model of multiple sclerosis (Liu, Lee et al. 2008). However, another study in an independent Spanish population investigated the role of common variants in *STAT3* in multiple sclerosis and the two clinical subtypes of inflammatory bowel disease, ulcerative colitis and Crohn's disease (Cenit, Alcina et al. 2010). While *STAT3* polymorphisms confirmed the gene's implication in colitis and Crohn's, the authors found no evidence for a major role of this gene in multiple sclerosis. The RTC pinpoints the existence of a regulatory effect in another proximal gene, also a member of the STAT family of transcription factors - *STAT5* (signal transducer and activator of transcription 5A). rs744166 scores an RTC of 1 with this gene in fibroblasts only, making it also an interesting candidate. Functional studies in mouse have shown that *STAT5* mediates the antiapoptotic effects of methylprednisolone (a synthetic glucocorticoid agonist used widely for the clinical therapy of spinal cord injuries and multiple sclerosis) on oligodendrocytes (Xu, Chen et al. 2009). Overexpression of an activated form of *STAT5* prevents oligodendrocyte cell death whereas knocking down this gene leads to

GWAS SNP	Complex Trait	Gene	RTC	Chr
rs6537837	Major depressive disorder	UBL4B	1	1
rs1390401	Height	JMJD4	0.92	1
rs3197999	Crohns disease	WDR6	0.92	3
rs4380451	Bipolar disorder	CMTM7	0.91	3
rs4143832	Plasma eosinophil count	HSPA4	0.97	5
rs2517713	Nasopharyngeal carcinoma	HLA-A	1	6
rs2523393	Multiple sclerosis	HLA-F	0.99	6
rs7743761	Ankylosing spondylitis	IER3	0.99	6
rs3130340	Bone mineral density (spine)	HLA-DMB	0.97	6
rs3131379	Systemic lupus erythematosus	HLA-DMB	0.96	6
rs9461688	Protein quantitative trait loci	IER3	0.96	6
rs742132	Serum uric acid	HIST1H4C	0.95	6
rs9264942	HIV-1 control	IER3	0.95	6
rs9469220	Crohns disease	HLA-DMA	0.94	6
rs12191877	Psoriasis	HLA-C	0.93	6
rs3131296	Schizophrenia	HLA-DMB	0.92	6
rs198846	Hemoglobin	HIST1H2BH	0.92	6
rs703842	Multiple sclerosis	FAM119B	1	12
rs11171739	Type 1 diabetes	RPS26	0.98	12
rs1994090	Parkinsons disease	C12orf4	0.98	12
rs10444502	Biochemical measures	RFC5	0.92	12
rs8020441	Cognitive performance	CDKL1	0.92	14
rs3825932	Type 1 diabetes	CTSH	0.95	15
rs744166	Multiple sclerosis	STAT5A	1	17
rs758642	Smoking behavior	OR1A1	1	17
rs8073783	Conduct disorder (interaction)	KIF2B	0.98	17
rs2191566	Acute lymphoblastic leukemia (childhood)	ZNF155	0.91	19
rs1555322	Attention deficit hyperactivity disorder	TRPC4AP	0.94	20
rs5751614	Height	BCR	0.97	22

Table 5.5. Candidate fibroblast results. Candidate genes (RTC score ≥ 0.9) for *cis* regulatory mediated GWAS effects. RTC applied on NHGRI GWAS SNPs and fibroblast expression data from the GenCord.

the loss of the protective effect. Thus, both genes (*STAT3* and *STAT5*) could be further considered as potential key determinants of disease risk. Additional studies of the interactions between each of the two gene products with other transcription factors might provide a better insight into the disease mechanisms characteristic for multiple sclerosis or other autoimmune traits.

Finally, for a subset of the GWAS SNPs tested in this chapter (N = 9) the RTC reveals identical gene candidates independently in at least two of the GenCord tissues (Table 5.6). Such consistent examples deserve special attention in future functional studies focusing on the respective genomic regions.

GWAS SNP	Complex Trait	Gene	RTC	Chr	Tissue
rs13160562	Alcohol dependence	ERAP1	1	5	B-cells
rs13160562	Alcohol dependence	ERAP1	0.91	5	T-cells
rs7731657	Fasting plasma glucose	CDC42SE2	0.99	5	T-cells
rs7731657	Fasting plasma glucose	CDC42SE2	0.96	5	B-cells
rs2227139	Hematological parameters	HLA-DRB5	0.97	6	B-cells
rs2227139	Hematological parameters	HLA-DRB5	0.94	6	T-cells
rs3129934	Multiple sclerosis	HLA-DRB5	0.98	6	B-cells
rs3129934	Multiple sclerosis	HLA-DRB5	0.98	6	T-cells
rs3135388	Multiple sclerosis	HLA-DRB5	1	6	B-cells
rs3135388	Multiple sclerosis	HLA-DRB5	1	6	T-cells
rs9268480	Ulcerative colitis	HLA-DQB1	0.95	6	B-cells
rs9268480	Ulcerative colitis	HLA-DQB1	0.95	6	T-cells
rs9272346	Type 1 diabetes	HLA-DQB1	1	6	B-cells
rs9272346	Type 1 diabetes	HLA-DQB1	0.99	6	T-cells
rs11171739	Type 1 diabetes	RPS26	0.98	12	B-cells
rs11171739	Type 1 diabetes	RPS26	0.98	12	T-cells
rs11171739	Type 1 diabetes	RPS26	0.98	12	Fibro
rs5751614	Height	BCR	0.97	22	T-cells
rs5751614	Height	BCR	0.97	22	Fibro

Table 5.6. RTC signals consistent across at least two tissues. Candidate genes (RTC score ≥ 0.9) for *cis* regulatory mediated GWAS effects consistent in at least two GenCord tissues.

5.5 Conclusions

In this chapter I explored the tissue-dependent value of gene expression variation in predicting candidate disease genes. The RTC methodology was applied to expression data in B-cells, T-cells and fibroblasts derived from 75 Swiss individuals and GWAS data from the NHGRI catalogue. As expected given the demonstrated high extent of *cis* eQTL tissue-specificity and the tissue-restricted manifestation of diseases, the results of the RTC analysis are overall highly cell-type specific. Of the total number of confident discoveries passing the 0.9 RTC score threshold, roughly 70% of the predicted effects are found only in one cell-type. In each of the three tissues, this corresponds to approximately 50% of all confident effects being specific per cell-type. The distribution of the scores and the pairwise comparisons of RTC discoveries mimic the biological properties of the tissues tested, whereby B-cells and T-cells share as expected proportionally more causal regulatory effects, many for immunity-related conditions. Each of the three cell-types permits the discovery of candidate disease genes whose differentiated regulation is affected by GWAS SNPs. The RTC confirms previously suspected expression mediated disease effects but also facilitates the informative prioritization of novel candidate causal genes, some already having plausible functional justification from experimental studies.

I highlight the risks of misinformed candidate gene prediction by relying solely on genetic distance criteria and give examples where the RTC can help distinguish likely causal effects from genes coincidentally residing closest to the GWAS SNP loci. Finally, the data suggests that establishing relevance of a cell-type to a complex trait is not trivial. The current knowledge about disease biology is generally limited and thus, predicting candidate disease tissues could be as unsuccessful as candidate gene approaches have proved to be for complex diseases. Additionally, the estimated fair amount (~30%) of tissue shared regulatory variation should encourage the interrogation of any available cell-type for potential regulatory disease effects. In this sense, I conclude by presenting a few unexpected associations revealed by the RTC, which could offer new insights into disease aetiology.

6 Discussion

Throughout my PhD I have been exploring further aspects of the genetics of human gene expression in an attempt to understand its role in the biology of complex disorders. Pioneering work studying gene expression variation documented its fundamental role in shaping phenotypic differences among cell-types (Schadt, Molony et al. 2008; Dimas, Deutsch et al. 2009), individuals (Cheung, Spielman et al. 2005; Stranger, Forrest et al. 2005) and populations (Stranger, Nica et al. 2007). This development has been concomitant with the progress in discovering genetic associations with complex traits by genome-wide association studies (GWAS) (McCarthy, Abecasis et al. 2008). However, the GWAS signals are hard to interpret in the absence of additional information (Dermitzakis 2008), as they often map to either non-genic regions or genes of no apparent functional relevance to the associated trait. Transcript abundance (mRNA levels) is a very proximal endophenotype immediately affected by DNA sequence variation. Thus, it provides a link between genotype and organismal phenotypes, which can be used to explain some of the genotype-phenotype associations revealed by GWAS. In this thesis, I developed a novel empirical methodology to explore the role of gene expression as an informative intermediate phenotype between DNA variation and disease and offered also new insights into the complexity of regulatory variation across multiple tissues. In the following sections, I summarize the main results of my study and discuss other relevant advancements and current pressing issues in the field.

6.1 eQTL and GWAS integration – RTC score

To aid the functional interpretation of complex trait association signals, I describe in Chapter 3 an empirical methodology (Regulatory Trait Concordance - RTC) that directly integrates eQTL and GWAS data while correcting for the local correlation structure in the human genome (linkage disequilibrium - LD). The RTC methodology addresses the issue of coincidental eQTL-GWAS SNP overlaps due to the pervasiveness of regulatory variants and prioritizes candidate disease genes based on their differential regulation.

Investigating the explanatory potential of regulatory variation is appropriate, as confirmed by the significant overrepresentation of eQTLs observed among currently published GWAS SNPs.

As a proof of principle I applied the RTC method initially on expression profiles quantified in LCLs. In line with the biological expectation, immunity-related traits were overrepresented among the significant results. It is clear that the tissue of expression has a decisive impact on the results of the method, as further exemplified in Chapter 5. Therefore, the RTC is unlikely to yield meaningful results for traits such as obesity or type 2 diabetes, unless expression data from the hypothalamus and β–cells respectively becomes available for analysis. Like many other experiments relying on genotyping assays, the method is limited by the SNP coverage in each region of interest. While the calculation of the RTC score accounts for the number of tested SNPs so that the metric is comparable across regions of variable sizes, for the same hotspot interval tested, the denser the SNP coverage, the more informative the score with respect to the relationship between the eQTL and the disease SNP. Imputation helps alleviate this constraint by inferring additional informative genetic variation. It should be noted however, that unlike other methods using whole-genome transcriptome data to discover disease candidates (e.g. network-based approaches), the RTC is a gene prioritization method relying on the validity and existence of prior GWAS results. The method requires prior information about the identity of disease susceptibility variants and helps direct functional studies towards the potential candidates affected by the disease SNPs.

With the limitations of tissue type, SNP coverage and prior GWAS information required, the RTC helps nonetheless discover likely causal *cis* regulatory effects for a variety of traits, confirming some already suspected as well as identifying a multitude of novel candidates. Long-range *trans* effects are also present but harder to identify due to lower power to test for such associations. Applying RTC in *trans* for intervals where a significant *cis* effect has been highlighted would be a useful next step in understanding the regulatory interactions underlying the respective GWAS signals. Ultimately, proving causality will demand the individual functional examination of each candidate proposed with the RTC approach, but in absence of such prioritization directions, the biological interpretation of the ever-increasing list of GWAS signals would be unattainable.

Finally, the RTC method is not limited to gene expression but could be generalized to any other endophenotype. As new methods are developed and larger cohorts become available, various intermediate cellular phenotypes are interrogated via association studies with the hope to find explanatory links between genotypic variation and complex trait predisposition. The biological interpretation of these discoveries will also be hardened by the presence of tight LD. It is therefore necessary to evaluate them in a conservative manner, correcting for the local correlation structure in each genomic interval with overlapping association signals. The integration of more intermediate cellular phenotypes will enhance our understanding of the biology of complex traits.

6.2 *Cis* eQTL tissue-specificity

Gene expression (mRNA transcript abundance) has already facilitated the identification of candidate susceptibility genes for a variety of conditions such as metabolic disease traits (Chen, Zhu et al. 2008), asthma (Moffatt, Kabesch et al. 2007) or Crohn's disease (McCarroll, Huett et al. 2008). Using the RTC methodology, further evidence has been acquired in favour of the overall GWAS explanatory potential of regulatory variation and new differentially expressed genes with potential disease causing role were revealed (Nica, Montgomery et al. 2010). However, some phenotypes manifest themselves only in certain tissues (Emilsson, Thorleifsson et al. 2008) and our guess of tissue relevance is yet far from satisfactory. Given this, the value of measuring expression in multiple cell-types, including primary tissues reflecting in vivo patterns, is incontestable. Transcriptional regulatory networks are expected to dictate tissue-specificity of regulatory effects (Ravasi, Suzuki et al. 2010) but the extent of this is still under debate.

In Chapter 4, I investigated further aspects of tissue-specificity in three human tissues: one cell-line (LCL) and two primary tissues of clinical importance (skin – previously uncharacterized and fat). An abundance of *cis* eQTLs was detected in all three tissues, at a comparable rate to other studies of similar sample size (Stranger, Nica et al. 2007). The eQTLs appear robust, replicating in a very high proportion (93-98%) in independent co-twin samples of identical (monozygotic twins) or 50% similar (dizygotic twins) genetic background. Using recombination hotspot coordinates and stringent LD filters, the detected signals were refined to likely independently acting *cis* eQTLs. Most genes were observed to have single associated regulatory variants, which, if shared across tissues, share the same direction of effect and map to the same recombination hotspot interval.

This suggests that largely, shared differentially regulated genes also share regulatory functional variants across tissues. Additionally, factor analysis (FA) was employed, accounting for global variance components in the data, which can be also of non-genetic nature (e.g. experimental noise or environmental conditions). FA further increased the power to detect eQTLs of smaller genetic effects, implying that future expression studies on larger sample sizes are expected to reveal a plethora of additional regulatory variants in each tissue.

The three tissues analyzed here support a large degree of tissue-specificity of eQTLs and emphasize the importance of accounting not only for statistical significance but also for continuous biological properties such as effect size. Most notably, significant eQTLs at the same threshold were observed to exhibit differential fold changes in expression between genotypes across tissues. Despite sharing statistical significance, these are also tissue-specific effects since they are likely to have different biological consequences. Given this, the biological interpretation of eQTLs - much like in the case of complex traits – is tissue-dependent and requires collecting multiple tissue expression datasets. Studying regulation of expression during different developmental stages as well as regulatory changes following exposure to various stimuli are essential future steps towards understanding gene regulation in more detail. Furthermore, *trans* effects and their tissue-specific properties are still largely unknown and remain to be discovered in better-powered eQTL studies. Understanding the genetic architecture of gene expression with its complexities and context-dependent effects is fundamental, especially if employed in explaining the biological properties of disease causing variants.

6.3 Tissue-dependent prediction of disease regulatory effects

The extensive tissue-specific component of regulatory variation is tested specifically in a disease context in Chapter 5. Here, I apply the RTC methodology on a multiple tissue dataset (GenCord) in order to prioritize disease relevant genes based on their potential causal regulatory effects. Each of the three tissues is informative with respect to a subset of GWAS signals, allowing the discovery of several regulatory effects with potential implications in disease aetiology.

The results support the decisive role of the tissue of origin where transcript abundance is quantified, for predicting trait-relevant candidate genes. Specifically, I observe that of the total amount of confident results, the majority (~70%) are restricted to one tissue only and when considering these discoveries in each tissue separately, 50% of the RTC results per tissue appear tissue-specific. The distribution of RTC scores in each of the three tissues reflects their distinct biological properties. As such, while expression data in each tissue contributes to the discovery of candidates undetectable in the other two tissues, the two immunity-related cell-types (B-cells and T-cells) share, as expected, more causal regulatory effects than any other pairwise tissue comparison. Nevertheless, establishing which tissue is relevant for which trait is not trivial. In addition to anticipated autoimmune signals revealed in B-cells and T-cells, a series of other biologically interesting and less expected candidates are detected. Upon further careful validation, some of these unexpected results may provide new clues about shared biological mechanisms involved in the pathology of different diseases, a hypothesis supported by the current overlap in GWAS results between apparently dissimilar complex traits. For the moment, the currently scarce knowledge about disease biology as well as the reasonable proportion of regulatory effects shared across tissues, justify the informative value of investigating any available expression dataset for potential RTC signals. The current results suggest that the more tissues we sample, the more likely we are to detect regulatory effects of special relevance to complex diseases. It would be ideal to screen a wide range of human tissues in the future and by combining it with GWAS data to create a “tissue map” of natural variation, whereby one could determine the most biologically relevant expression changes for a variant of interest and estimate how distant this prediction is compared to the case when one would access the tissue where the first molecular change relevant to the disease occurs.

6.4 Next-generation genomics

The development of high-throughput microarray and genotyping technologies enabled the current progress in understanding the genetics of gene expression variation and complex disease risk. While this has been a great achievement, several limitations still exist and need to be addressed in the near future.

Firstly, most of the association studies performed so far rely on human DNA sequence representing the common genetic variation in any region of interest. This means that the susceptibility variants reported are most probably only tagging the real functional variants and are not causal themselves. Initial discoveries should ideally be followed by fine mapping the regions harbouring the significant statistical signals. However, this was not thoroughly attempted so far, primarily because in the absence of other prior biological information, such tasks were financially unaffordable. The drop in sequencing costs is gradually reducing this impediment, but the perfect correlation (LD) between variants precludes the identification of functional SNPs even in narrower susceptibility regions. Most likely, the smaller set of susceptibility variants revealed by targeted resequencing will need to be further analyzed in functional assays to establish causation beyond doubt. Traditional microarray experiments also suffer from capturing only a subset of the overall transcriptome diversity. Typically, only few probes are presently designed per gene making it impossible to resolve issues like alternative splicing. Measurements of transcript abundance are also problematic in cases of genes expressed at low levels, which are hard to distinguish from background noise or in cases when genes are expressed at very high levels, as microarrays reach saturation.

The development of protocols for next-generation sequencing (Margulies, Egholm et al. 2005; Shendure, Porreca et al. 2005) marked the start of a revolutionary direction for genetic studies, addressing the above-mentioned limitations. Next-generation sequencing has already made efforts like the 1000 Genomes Project possible (http://www.1000genomes.org/), a resource set up to generate a human genetic variation map at unprecedented resolution. The initial goal of the project was to sequence more than 1000 individuals and catalogue almost all variants found at minor allele frequency > 1% in different human populations (European, African and East Asian). Within genes, sequencing goes even deeper, down to 0.5% frequency. After the completion of the pilot tests, the project is currently being extended towards a full set of genomes coming from 2,500 individuals from 27 populations around the world. Clearly, such detailed sequence information will allow the discovery of additional disease susceptibility variants through GWAS (limited by technology, current GWAS studies have typically surveyed only common DNA variants with frequency greater than 5-10%). Furthermore, the 1000 Genomes Project will significantly enhance our knowledge by surveying other forms of genetic variation in addition to the traditionally typed single base polymorphisms (SNPs).

Small insertions or deletions (indels) as well as larger changes in the structure and copy number of certain genomic regions (CNVs) will also be documented. These additional forms of genetic variation together with previously undetected rare SNP variants will lead to the discovery of potentially new disease risk factors.

Next-generation sequencing technology has also been recently applied to profile in depth the transcriptome (Wang, Gerstein et al. 2009). RNA sequencing (RNA-seq) has several important advantages compared to gene expression measurements using microarrays: a much more accurate quantification of transcript levels, assessment of alternative splicing and the ability to detect novel gene structures (Montgomery and Dermitzakis 2009). Two recent landmark papers demonstrated the value of RNA-seq in linking genetic sequence variation to transcript abundance at an unparalleled resolution (Montgomery, Sammeth et al. 2010; Pickrell, Marioni et al. 2010). In the two studies, RNA from LCLs derived from ~60 European (CEU) and African (YRI) HapMap individuals respectively, was deep-sequenced. The transcript information thus generated was used in conjunction with genotypic data available from the HapMap project in order to detect genome-wide associations (eQTLs). Both papers reveal a greater number of eQTLs than previously reported by studies using microarray technologies. The eQTL overlap between the two studies, as well as their overlap with prior discoveries validate them as real genetic effects. RNA-seq allows a better quantification of transcript isoforms and facilitates the discovery of a considerable number of variants responsible for alternative splicing. Furthermore, allele-specific expression was assayed in the same experiment, permitting also the identification of rare eQTLs and allelic differences in transcript structure (Montgomery, Sammeth et al. 2010). Finally, new putative coding-exons were discovered, as well as a multitude of unannotated exons and new polyadenylation sites, highlighting the current lack of completeness of gene annotation (Pickrell, Marioni et al. 2010).

These new important aspects of the complexity in the transcriptional landscape will offer new insights into the genetic control of gene expression and in turn, its intermediate role in determining other complex traits. Next-generation genomics will soon be able to combine detailed genetic variation maps (e.g. 1000 Genomes Project) with high-resolution transcriptional information sampled over multiple tissues and enable thus a more accurate description of the tissue-specific features of regulatory variation.

Next-generation sequencing is being also used to produce genome-scale epigenomic and interactome data (Hawkins, Hon et al. 2010). Epigenetic modifications play an essential role in transcriptional control and substantial variation in chromatin states has been recently observed, along with evidence that chromatin differences are heritable (Martienssen and Colot 2001; Eckhardt, Lewin et al. 2006; Vaughn, Tanurdzic et al. 2007). So far, the best characterized examples of epigenetic heritability come from plant studies (e.g. segregation of parental alleles with different epigenetic signatures has been implicated in variation of height and flowering time of *Arabidopsis thaliana* (Johannes, Porcher et al. 2009)). These results motivate documenting epigenetic variation at a large scale and investigating its consequences on variation in human complex traits. It is now possible to perform nucleotide resolution mapping of methylated DNA sites at genome-wide scale, by coupling next-generation sequencing with bisulphite treatment of DNA (MethylC–seq) (Lister, Pelizzola et al. 2009) or with immunoprecipitation of methylated DNA using antibodies (MeDIP-seq) (Li, Ye et al. 2010). Determining physical and functional interactions across the genome (interactome) is yet another crucial development facilitated by next-generation sequencing. ChIP-seq (Robertson, Hirst et al. 2007) and more recently CLIP-seq (Chi, Zang et al. 2009) methods combine chromatin immunoprecipitation (ChIP) techniques with deep sequencing to determine DNA-protein and RNA-protein interactions respectively. Long-range DNA interactions mediated potentially also through protein interactions are being investigated too, using chromosome confirmation capture (3C) technologies (Dekker, Rippe et al. 2002). These, combined with high-throughput paired-end sequencing have demonstrated the feasibility of detecting genomic interactions at genome-wide scale (Lieberman-Aiden, van Berkum et al. 2009).

Together, all these comprehensive datasets will greatly improve the functional annotation of the human genome. The emerging era of next-generation genomics will be dominated by attempts to integrate these different sources of information. Their success will be crucial for our ability to explain the biology behind the presently known genetic associations with complex traits.

6.5 The missing heritability of complex diseases

The value of GWAS studies in advancing the knowledge on the genetics of complex diseases is indisputable. The results so far offer new insights into disease biology by revealing previously unsuspected susceptibility pathways and highlighting unanticipated overlaps between loci associated with different conditions. For example, the pathogenesis of type 2 diabetes is now confidently linked to disruptions of the function of insulin-producing β-cells and multiple studies on Crohn's disease point now to autophagy - the process by which cells digest themselves via the lysosome - and innate immunity mechanisms as being implicated in disease aetiology (Barrett, Hansoul et al. 2008). Surprising GWAS overlaps have been observed, including the 8q24 gene desert region harbouring several independent susceptibility loci for prostate cancer, colon cancer, as well as one breast cancer variant. Weather these loci share a common mechanism leading to cancer onset is unknown, as well as the genes whose function they might disrupt. However, the *MYC* oncogene is a plausible nearby candidate and its interaction with tissue-specific enhancers within 8q24 is one recently proposed mechanism explaining the statistical associations overlap (Ahmadiyeh, Pomerantz et al. 2010). Further functional studies will better characterise these intricate disease links, otherwise undiscovered in the absence of GWAS studies. More interesting lessons about disease biology will surely be learned from the other >500 independent strong SNP associations (P-value < 10^{-8}) reported so far with various complex traits (Hindorff, Sethupathy et al. 2009).

GWAS studies started revealing the genetic landscape of many common diseases, yet most of the variants identified (typically common SNPs with MAF > 5%) have very small effect sizes and explain only a very small proportion of the heritability of their associated traits. The proportion of phenotypic variation attributable to genetic variation (heritability) is very modest for most of the common traits investigated, even when the traits themselves have an estimated high level of heritability (Cirulli and Goldstein 2010). For example, the heritability of height has been estimated at ~ 0.8 (Silventoinen, Sammalisto et al. 2003; Visscher, Hill et al. 2008), yet the 50 associated common variants identified so far account only for ~5% of the phenotypic variance in the population (Visscher 2008; Weedon, Lango et al. 2008). Similarly, schizophrenia has an estimated heritability of 0.8-0.85 and a GWAS meta-analysis including over 8,000 cases and 19,000 controls identified only 7 significant SNPs, each with an odds ratio below 1.3 (Shi, Levinson et al. 2009). Finally, the 18 common variants significantly associated with type 2 diabetes only explain 6% of the increased disease risk among relatives (Zeggini, Scott et al. 2008; Manolio, Collins et al. 2009). These observations bring up the important issue of finding out where the rest of the 'missing heritability' is and how can it be explained.

Several possible hypotheses have been formulated in order to elucidate the missing heritability problem (Eichler, Flint et al. 2010). First, the incomplete assessment of the spectrum of human genetic variation has been criticized. Compared to single nucleotide changes (SNPs), larger structural variants like deletions, duplications or inversions have been understudied. Although individually rare, this type of variation is collectively common in the human population (Redon, Ishikawa et al. 2006) and can offer new insights into disease genetics. In fact, in a few instances common CNVs have been shown to play key disease susceptibility roles. A 20 kb deletion polymorphism upstream of *IRGM* (immunity-related GTPase family, M) and in perfect LD (r^2 = 1.0) with the most significant Crohn's disease SNP in that region has been causally implicated in the disorder through a distinctly altered expression pattern affecting autophagy efficiency (McCarroll, Huett et al. 2008). Another deletion (45-kb long) is a strong candidate for explaining the BMI association signal at the *NEGR1* (neuronal growth regulator 1) locus (Willer, Speliotes et al. 2009). Here too, the structural variant was in perfect LD with the most significant SNPs detected by the GWAS analysis. Recent studies report similar observations on a large scale. The WTCCC analyzed eight complex diseases with 3,432 common CNVs in 17,000 individuals and concluded that common copy number polymorphisms contributing to phenotypic variation are already largely accounted for by GWAS (Conrad, Pinto et al. 2010; Craddock, Hurles et al. 2010). It is possible that rare CNVs (e.g. rare recurrent variants of larger effect size (Bochukova, Huang et al. 2010)) or those of a more complex nature and currently not detectable with existing technology would have a higher impact on disease risk. Common CNVs however are unlikely to account for much of the missing heritability.

Another relevant heritability aspect, largely overlooked due to the difficulty in detecting and accounting for this type of effect, is the parent of origin dependent disease risk. Recently, a few susceptibility variants for cancer and type 2 diabetes were reported as conferring disease risk only when inherited from a certain parent (Kong, Steinthorsdottir et al. 2009). Heritability values of such variants are underestimated if parental origin is not taken into account. However, the overall proportion of these effects and the likely number of diseases where they might play a role remains unknown and hard to approximate due to low power.

Assessing the contribution of rare variants to common disease predisposition is perhaps one of the most immediate questions of disease genetics and the most promising explanation for the current missing heritability. Extremely rare (private, MAF<0.5%) or intermediately rare variants (0.5%<MAF<5%) are currently out of the scope of genotyping arrays employed in GWAS and have been underexplored. Low frequency variants are suspected to have greater effect sizes, increasing the disease risk by two or threefold compared to the typically modest (1.1-1.5-fold) risk conferred by common variants. Few examples, mostly from lipids studies, already exist in the literature supporting the hypothesis that genes harbouring common disease risk variants can also contain rare variants with larger effects. 11 out of 30 genes containing common susceptibility variants influencing plasma lipid concentrations have been shown to also harbour rare variants of large effects identified previously in Mendelian dyslipidemias (abnormal lipids amount in the blood) (Kathiresan, Willer et al. 2009). Johansen et al. further explored the extent to which rare variants affect lipid phenotypes (Johansen, Wang et al. 2010). The authors report an excess of rare variants in GWAS-identified susceptibility genes for hypertriglyceridemia, the polygenic condition characterized by high fasting plasma triglycerides levels. Resequencing of four genes (*APOA5*, *GCKR*, *LPL* and *APOB*) containing common GWAS variants uncovered a significant burden of 154 rare missense or nonsense SNPs in 438 cases, compared to only 53 variants in 327 controls. Considering the rare variants in these genes alongside the common susceptibility SNPs increases the proportion of explained heritability of the trait.

Next-generation sequencing will enable the comprehensive detection of similar rare genetic changes in susceptibility genes for other complex traits. However, the genotype-phenotype relationship is of a complex nature and most likely distinct across different common traits. As such, it is possible that for other human traits, a more realistic biological view would be one involving rare combinations of common variants (Eichler, Flint et al. 2010). This hypothesis has been tested very recently in a study on human height, providing supporting evidence for its soundness (Yang, Benyamin et al. 2010).

The authors show that the missing heritability problem is overstated for this trait, evaluating that a large proportion of the heritability is in fact hidden by current estimates, and not missing. Yang et al. argue that a large proportion of the height heritability can already be explained by common variants, provided that all SNPs are considered simultaneously. Traditional GWAS approaches test for strong independent genetic effects and require evidence of replication in independent cohorts. Such a stringent approach is bound to miss many causal SNPs that do not pass these significance cut-offs. Therefore, the authors use a linear model where they regress at the same time all GWAS SNPs against an adjusted measure of height. With this model they estimate that 45% of the 80% height heritability can actually be explained, an almost ten-fold increase from the typical 5% height variance accounted for in the literature. By accounting for incomplete LD between the tagging and causal variants, the authors increase their explained heritability estimate of stature to at least 67%. The difference in LD between the common genotyped SNPs and the actual causal variants is explained by the fact that causal variants, being likely deleterious are kept at lower MAF than the tagging SNPs surveyed by GWAS. Therefore, most of the heritability for height can actually already be captured by common variants. Weather this will be the case for other complex traits, especially common diseases, remains to be tested. Rare causal SNPs of larger effects can have a marked genetic contribution to the risk of particular diseases and their discovery remains necessary. The ultimate goal of translating genetic knowledge into clinical practice can only be attained through a thorough understanding of trait-specific genetic architecture and next-generation sequencing will play an essential role towards this end.

References

(2009). "Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20." Nat Genet **41**(7): 824-828.

Abzhanov, A., M. Protas, et al. (2004). "Bmp4 and morphological variation of beaks in Darwin's finches." Science **305**(5689): 1462-1465.

Adams, M. D., A. R. Kerlavage, et al. (1995). "Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence." Nature **377**(6547 Suppl): 3-174.

Ahmadiyeh, N., M. M. Pomerantz, et al. (2010). "8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC." Proc Natl Acad Sci U S A **107**(21): 9742-9746.

Al Olama, A. A., Z. Kote-Jarai, et al. (2009). "Multiple loci on 8q24 associated with prostate cancer susceptibility." Nat Genet **41**(10): 1058-1060.

Alberts, B. (2002). Molecular biology of the cell. New York, Garland Science.

Alley, C. D. (1987). "Human bone marrow-derived IgA is produced by IgA-committed B cells in vitro." J Clin Immunol **7**(2): 151-158.

Aulchenko, Y. S., S. Ripatti, et al. (2009). "Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts." Nat Genet **41**(1): 47-55.

Bailey, R., J. D. Cooper, et al. (2007). "Association of the vitamin D metabolism gene CYP27B1 with type 1 diabetes." Diabetes **56**(10): 2616-2621.

Barrett, J. C., S. Hansoul, et al. (2008). "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease." Nat Genet **40**(8): 955-962.

Bentires-Alj, M., M. I. Kontaridis, et al. (2006). "Stops along the RAS pathway in human genetic disease." Nat Med **12**(3): 283-285.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Bochukova, E. G., N. Huang, et al. (2010). "Large, rare chromosomal deletions associated with severe early-onset obesity." Nature **463**(7281): 666-670.

Bolstad, B. M., R. A. Irizarry, et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.

Boxer, L. M. and C. V. Dang (2001). "Translocations involving c-myc and c-myc function." Oncogene **20**(40): 5595-5610.

Boyle, A. P., S. Davis, et al. (2008). "High-resolution mapping and characterization of open chromatin across the genome." Cell **132**(2): 311-322.

Braun, T., G. Buschhausen-Denker, et al. (1989). "A novel human muscle factor related to but distinct from MyoD1 induces myogenic conversion in 10T1/2 fibroblasts." EMBO J **8**(3): 701-709.

Brem, R. B., J. D. Storey, et al. (2005). "Genetic interactions between polymorphisms that affect gene expression in yeast." Nature **436**(7051): 701-703.

Brem, R. B., G. Yvert, et al. (2002). "Genetic dissection of transcriptional regulation in budding yeast." Science **296**(5568): 752-755.

Browning, S. R. and B. L. Browning (2007). "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering." Am J Hum Genet **81**(5): 1084-1097.

Cenit, M. C., A. Alcina, et al. (2010). "STAT3 locus in inflammatory bowel disease and multiple sclerosis susceptibility." Genes Immun **11**(3): 264-268.

Chen, K. and N. Rajewsky (2007). "The evolution of gene regulation by transcription factors and microRNAs." Nat Rev Genet **8**(2): 93-103.

Chen, Y., J. Zhu, et al. (2008). "Variations in DNA elucidate molecular networks that cause disease." Nature **452**(7186): 429-435.

Cheung, V. G., L. K. Conlin, et al. (2003). "Natural variation in human gene expression assessed in lymphoblastoid cells." Nat Genet **33**(3): 422-425.

Cheung, V. G. and R. S. Spielman (2009). "Genetics of human gene expression: mapping DNA variants that influence gene expression." Nat Rev Genet **10**(9): 595-604.

Cheung, V. G., R. S. Spielman, et al. (2005). "Mapping determinants of human gene expression by regional and genome-wide association." Nature **437**(7063): 1365-1369.

Chi, S. W., J. B. Zang, et al. (2009). "Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps." Nature **460**(7254): 479-486.

Choy, E., R. Yelensky, et al. (2008). "Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines." PLoS Genet **4**(11): e1000287.

Churchill, G. A. and R. W. Doerge (1994). "Empirical threshold values for quantitative trait mapping." Genetics **138**(3): 963-971.

Cirulli, E. T. and D. B. Goldstein (2010). "Uncovering the roles of rare variants in common disease through whole-genome sequencing." Nat Rev Genet **11**(6): 415-425.

Conrad, D. F., D. Pinto, et al. (2010). "Origins and functional impact of copy number variation in the human genome." Nature **464**(7289): 704-712.

Consortium, I. H. (2003). "The International HapMap Project." Nature **426**(6968): 789-796.

Consortium, I. H. (2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-1320.

Cookson, W., L. Liang, et al. (2009). "Mapping complex disease traits with global gene expression." Nat Rev Genet **10**(3): 184-194.

Cowles, C. R., J. N. Hirschhorn, et al. (2002). "Detection of regulatory variation in mouse genes." Nat Genet **32**(3): 432-437.

Craddock, N., M. E. Hurles, et al. (2010). "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls." Nature **464**(7289): 713-720.

Crick, F. H. (1958). "On protein synthesis." Symp Soc Exp Biol **12**: 138-163.

Dawson, E., G. R. Abecasis, et al. (2002). "A first-generation linkage disequilibrium map of human chromosome 22." Nature **418**(6897): 544-548.

De Gobbi, M., V. Viprakasit, et al. (2006). "A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter." Science **312**(5777): 1215-1217.

Dekker, J., K. Rippe, et al. (2002). "Capturing chromosome conformation." Science **295**(5558): 1306-1311.

Dermitzakis, E. T. (2008). "From gene expression to disease risk." Nat Genet **40**(5): 492-493.

Dimas, A. S., S. Deutsch, et al. (2009). "Common regulatory variation impacts gene expression in a cell type-dependent manner." Science **325**(5945): 1246-1250.

Dimas, A. S., B. E. Stranger, et al. (2008). "Modifier effects between regulatory and protein-coding variation." PLoS Genet **4**(10): e1000244.

Dixon, A. L., L. Liang, et al. (2007). "A genome-wide association study of global gene expression." Nat Genet **39**(10): 1202-1207.

Doerge, R. W. and G. A. Churchill (1996). "Permutation tests for multiple loci affecting a quantitative character." Genetics **142**(1): 285-294.

Dupuis, J., C. Langenberg, et al. (2010). "New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk." Nat Genet **42**(2): 105-116.

Eckhardt, F., J. Lewin, et al. (2006). "DNA methylation profiling of human chromosomes 6, 20 and 22." Nat Genet **38**(12): 1378-1385.

Eeles, R. A., Z. Kote-Jarai, et al. (2008). "Multiple newly identified loci associated with prostate cancer susceptibility." Nat Genet **40**(3): 316-321.

Eichler, E. E., J. Flint, et al. (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." Nat Rev Genet **11**(6): 446-450.

Emilsson, V., G. Thorleifsson, et al. (2008). "Genetics of gene expression and its effect on disease." Nature **452**(7186): 423-428.

Evans, W. E. and M. V. Relling (1999). "Pharmacogenomics: translating functional genomics into rational therapeutics." Science **286**(5439): 487-491.

Franke, L., H. van Bakel, et al. (2006). "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes." Am J Hum Genet **78**(6): 1011-1025.

Fraser, H. B. and X. Xie (2009). "Common polymorphic transcript variation in human disease." Genome Res **19**(4): 567-575.

Frayling, T. M., N. J. Timpson, et al. (2007). "A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity." Science **316**(5826): 889-894.

Frazer, K. A., D. G. Ballinger, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-861.

Gandhi, K. S., F. C. McKay, et al. (2010). "The multiple sclerosis whole blood mRNA transcriptome and genetic associations indicate dysregulation of specific T cell pathways in pathogenesis." Hum Mol Genet **19**(11): 2134-2143.

Geyer, P. K. and V. G. Corces (1992). "DNA position-specific repression of transcription by a Drosophila zinc finger protein." Genes Dev **6**(10): 1865-1873.

Ghoussaini, M., H. Song, et al. (2008). "Multiple loci with different cancer specificities within the 8q24 gene desert." J Natl Cancer Inst **100**(13): 962-966.

Gibson, G. (2008). "The environmental contribution to gene expression profiles." Nat Rev Genet **9**(8): 575-581.

Gilad, Y., A. Oshlack, et al. (2006). "Natural selection on gene expression." Trends Genet **22**(8): 456-461.

Gilad, Y., S. A. Rifkin, et al. (2008). "Revealing the architecture of gene regulation: the promise of eQTL studies." Trends Genet **24**(8): 408-415.

Giuliani, N., G. Girasole, et al. (1999). "Ethanol and acetaldehyde inhibit the formation of early osteoblast progenitors in murine and human bone marrow cultures." Alcohol Clin Exp Res **23**(2): 381-385.

Gompel, N., B. Prud'homme, et al. (2005). "Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila." Nature **433**(7025): 481-487.

Goring, H. H., J. E. Curran, et al. (2007). "Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes." Nat Genet **39**(10): 1208-1216.

Goyette, P., C. Lefebvre, et al. (2008). "Gene-centric association mapping of chromosome 3p implicates MST1 in IBD pathogenesis." Mucosal Immunol **1**(2): 131-138.

Grange, T., J. Roux, et al. (1991). "Cell-type specific activity of two glucocorticoid responsive units of rat tyrosine aminotransferase gene is associated with multiple binding sites for C/EBP and a novel liver-specific nuclear factor." Nucleic Acids Res **19**(1): 131-139.

Gregersen, P. K., C. I. Amos, et al. (2009). "REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis." Nat Genet **41**(7): 820-823.

Guan, Y. and M. Stephens (2008). "Practical issues in imputation-based association mapping." PLoS Genet **4**(12): e1000279.

Gudbjartsson, D. F., G. B. Walters, et al. (2008). "Many sequence variants affecting diversity of adult human height." Nat Genet **40**(5): 609-615.

Guenther, M. G., S. S. Levine, et al. (2007). "A chromatin landmark and transcription initiation at most promoters in human cells." Cell **130**(1): 77-88.

Guo, Y., L. J. Tan, et al. (2010). "Genome-wide association study identifies ALDH7A1 as a novel susceptibility gene for osteoporosis." PLoS Genet **6**(1): e1000806.

Hakonarson, H., S. F. Grant, et al. (2007). "A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene." Nature **448**(7153): 591-594.

Hara, H., C. Ishihara, et al. (2007). "The adaptor protein CARD9 is essential for the activation of myeloid cells through ITAM-associated and Toll-like receptors." Nat Immunol **8**(6): 619-629.

Harris, M. B., J. Mostecki, et al. (2005). "Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function." J Biol Chem **280**(13): 13114-13121.

Hawkins, R. D., G. C. Hon, et al. (2010). "Next-generation genomics: an integrative approach." Nat Rev Genet **11**(7): 476-486.

Heinzen, E. L., D. Ge, et al. (2008). "Tissue-specific genetic control of splicing: implications for the study of complex traits." PLoS Biol **6**(12): e1.

Hindorff, L. A., P. Sethupathy, et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proc Natl Acad Sci U S A **106**(23): 9362-9367.

Hirschhorn, J. N. and M. J. Daly (2005). "Genome-wide association studies for common diseases and complex traits." Nat Rev Genet **6**(2): 95-108.

Holden, C. and R. Mace (1997). "Phylogenetic analysis of the evolution of lactose digestion in adults." Hum Biol **69**(5): 605-628.

Hom, G., R. R. Graham, et al. (2008). "Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX." N Engl J Med **358**(9): 900-909.

Hoogendoorn, B., S. L. Coleman, et al. (2003). "Functional analysis of human promoter polymorphisms." Hum Mol Genet **12**(18): 2249-2254.

Hsu, Y. M., Y. Zhang, et al. (2007). "The adaptor protein CARD9 is required for innate immune responses to intracellular pathogens." Nat Immunol **8**(2): 198-205.

Hugot, J. P., M. Chamaillard, et al. (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." Nature **411**(6837): 599-603.

Huszar, D., C. A. Lynch, et al. (1997). "Targeted disruption of the melanocortin-4 receptor results in obesity in mice." Cell **88**(1): 131-141.

Iafrate, A. J., L. Feuk, et al. (2004). "Detection of large-scale variation in the human genome." Nat Genet **36**(9): 949-951.

Idaghdour, Y., J. D. Storey, et al. (2008). "A genome-wide gene expression signature of environmental geography in leukocytes of Moroccan Amazighs." PLoS Genet **4**(4): e1000052.

Ingram, C. J., C. A. Mulcare, et al. (2009). "Lactose digestion and the evolutionary genetics of lactase persistence." Hum Genet **124**(6): 579-591.

Ioannidis, J. P., E. E. Ntzani, et al. (2001). "Replication validity of genetic association studies." Nat Genet **29**(3): 306-309.

Iwamoto, K., M. Bundo, et al. (2004). "Expression of HSPF1 and LIM in the lymphoblastoid cells derived from patients with bipolar disorder and schizophrenia." J Hum Genet **49**(5): 227-231.

Jakkula, E., V. Leppa, et al. (2010). "Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene." Am J Hum Genet **86**(2): 285-291.

Jansen, R. C. and J. P. Nap (2001). "Genetical genomics: the added value from segregation." Trends Genet **17**(7): 388-391.

Jimenez-Sanchez, G., B. Childs, et al. (2001). "Human disease genes." Nature **409**(6822): 853-855.

Johannes, F., E. Porcher, et al. (2009). "Assessing the impact of transgenerational epigenetic variation on complex traits." PLoS Genet **5**(6): e1000530.

Johansen, C. T., J. Wang, et al. (2010). "Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia." Nat Genet **42**(8): 684-687.

Jones, S., X. Zhang, et al. (2008). "Core signaling pathways in human pancreatic cancers revealed by global genomic analyses." Science **321**(5897): 1801-1806.

Kathiresan, S., C. J. Willer, et al. (2009). "Common variants at 30 loci contribute to polygenic dyslipidemia." Nat Genet **41**(1): 56-65.

Kellum, R. and P. Schedl (1992). "A group of scs elements function as domain boundaries in an enhancer-blocking assay." Mol Cell Biol **12**(5): 2424-2431.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-664.

Kerkhof, H. J., R. J. Lories, et al. (2010). "A genome-wide association study identifies an osteoarthritis susceptibility locus on chromosome 7q22." Arthritis Rheum **62**(2): 499-510.

King, M. C. and A. C. Wilson (1975). "Evolution at two levels in humans and chimpanzees." Science **188**(4184): 107-116.

Kleinjan, D. A. and V. van Heyningen (2005). "Long-range control of gene expression: emerging mechanisms and disruption in disease." Am J Hum Genet **76**(1): 8-32.

Koch, C. M., R. M. Andrews, et al. (2007). "The landscape of histone modifications across 1% of the human genome in five human cell lines." Genome Res **17**(6): 691-707.

Kong, A., V. Steinthorsdottir, et al. (2009). "Parental origin of sequence variants associated with complex diseases." Nature **462**(7275): 868-874.

Kong, A. M., C. J. Speed, et al. (2000). "Cloning and characterization of a 72-kDa inositol-polyphosphate 5-phosphatase localized to the Golgi network." J Biol Chem **275**(31): 24052-24064.

Lage, K., N. T. Hansen, et al. (2008). "A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes." Proc Natl Acad Sci U S A **105**(52): 20870-20875.

Lander, E. S. and N. J. Schork (1994). "Genetic dissection of complex traits." Science **265**(5181): 2037-2048.

Leask, A., M. Rosenberg, et al. (1990). "Regulation of a human epidermal keratin gene: sequences and nuclear factors involved in keratinocyte-specific transcription." Genes Dev **4**(11): 1985-1998.

Leek, J. T. and J. D. Storey (2007). "Capturing heterogeneity in gene expression studies by surrogate variable analysis." PLoS Genet **3**(9): 1724-1735.

Lemire, J. M. and D. C. Archer (1991). "1,25-dihydroxyvitamin D3 prevents the in vivo induction of murine experimental autoimmune encephalomyelitis." J Clin Invest **87**(3): 1103-1107.

Leonard, E. J. and A. Skeel (1976). "A serum protein that stimulates macrophage movement, chemotaxis and spreading." Exp Cell Res **102**(2): 434-438.

Lettre, G., A. U. Jackson, et al. (2008). "Identification of ten loci associated with height highlights new biological pathways in human growth." Nat Genet **40**(5): 584-591.

Levine, M. and R. Tjian (2003). "Transcription regulation and animal diversity." Nature **424**(6945): 147-151.

Lewontin, R. C. and L. C. Dunn (1960). "The Evolutionary Dynamics of a Polymorphism in the House Mouse." Genetics **45**(6): 705-722.

Li, L., S. He, et al. (2004). "Gene regulation by Sp1 and Sp3." Biochem Cell Biol **82**(4): 460-471.

Li, N., M. Ye, et al. (2010). "Whole genome DNA methylation analysis based on high throughput sequencing technology." Methods.

Li, Q., K. R. Peterson, et al. (2002). "Locus control regions." Blood **100**(9): 3077-3086.

Libioulle, C., E. Louis, et al. (2007). "Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4." PLoS Genet **3**(4): e58.

Lieberman-Aiden, E., N. L. van Berkum, et al. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." Science **326**(5950): 289-293.

Lister, R., M. Pelizzola, et al. (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences." Nature **462**(7271): 315-322.

Liu, X., Y. S. Lee, et al. (2008). "Loss of STAT3 in CD4+ T cells prevents development of experimental autoimmune diseases." J Immunol **180**(9): 6070-6076.

Loos, R. J., C. M. Lindgren, et al. (2008). "Common variants near MC4R are associated with fat mass, weight and risk of obesity." Nat Genet **40**(6): 768-775.

Lowrey, C. H., D. M. Bodine, et al. (1992). "Mechanism of DNase I hypersensitive site formation within the human globin locus control region." Proc Natl Acad Sci U S A **89**(3): 1143-1147.

Mann, M. and O. N. Jensen (2003). "Proteomic analysis of post-translational modifications." Nat Biotechnol **21**(3): 255-261.

Manolio, T. A., F. S. Collins, et al. (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Martienssen, R. A. and V. Colot (2001). "DNA methylation and epigenetic inheritance in plants and filamentous fungi." Science **293**(5532): 1070-1074.

Maston, G. A., S. K. Evans, et al. (2006). "Transcriptional regulatory elements in the human genome." Annu Rev Genomics Hum Genet **7**: 29-59.

McCarroll, S. A., A. Huett, et al. (2008). "Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease." Nat Genet **40**(9): 1107-1112.

McCarroll, S. A., A. Huett, et al. (2008). "Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease." Nat Genet.

McCarroll, S. A., F. G. Kuruvilla, et al. (2008). "Integrated detection and population-genetic analysis of SNPs and copy number variation." Nat Genet **40**(10): 1166-1174.

McCarthy, M. I., G. R. Abecasis, et al. (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nat Rev Genet **9**(5): 356-369.

McCracken, S., N. Fong, et al. (1997). "The C-terminal domain of RNA polymerase II couples mRNA processing to transcription." Nature **385**(6614): 357-361.

McHeyzer-Williams, L. J., N. Pelletier, et al. (2009). "Follicular helper T cells as cognate regulators of B cell immunity." Curr Opin Immunol **21**(3): 266-273.

McVean, G. A., S. R. Myers, et al. (2004). "The fine-scale structure of recombination rate variation in the human genome." Science **304**(5670): 581-584.

Meyer, O. (2009). "Interferons and autoimmune disorders." Joint Bone Spine **76**(5): 464-473.

Miller, R. G. (1981). Simultaneous statistical inference. New York, Springer-Verlag.

Modrek, B. and C. J. Lee (2003). "Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss." Nat Genet **34**(2): 177-180.

Modrek, B., A. Resch, et al. (2001). "Genome-wide detection of alternative splicing in expressed sequences of human genes." Nucleic Acids Res **29**(13): 2850-2859.

Moffatt, M. F., M. Kabesch, et al. (2007). "Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma." Nature **448**(7152): 470-473.

Monks, S. A., A. Leonardson, et al. (2004). "Genetic inheritance of gene expression in human cell lines." Am J Hum Genet **75**(6): 1094-1105.

Montgomery, S. B. and E. T. Dermitzakis (2009). "The resolution of the genetics of gene expression." Hum Mol Genet **18**(R2): R211-215.

Montgomery, S. B., M. Sammeth, et al. (2010). "Transcriptome genetics using second generation sequencing in a Caucasian population." Nature **464**(7289): 773-777.

Morley, M., C. M. Molony, et al. (2004). "Genetic analysis of genome-wide variation in human gene expression." Nature **430**(7001): 743-747.

Morrison, A. C., C. B. Wilson, et al. (2004). "Macrophage-stimulating protein, the ligand for the stem cell-derived tyrosine kinase/RON receptor tyrosine kinase, inhibits IL-12 production by primary peritoneal macrophages stimulated with IFN-gamma and lipopolysaccharide." J Immunol **172**(3): 1825-1832.

Myers, A. J., J. R. Gibbs, et al. (2007). "A survey of genetic human cortical gene expression." Nat Genet **39**(12): 1494-1499.

Myers, S., L. Bottolo, et al. (2005). "A fine-scale map of recombination rates and hotspots across the human genome." Science **310**(5746): 321-324.

Naukkarinen, J., I. Surakka, et al. (2010). "Use of genome-wide expression data to mine the "Gray Zone" of GWA studies leads to novel candidate obesity genes." PLoS Genet **6**(6): e1000976.

Nica, A. C. and E. T. Dermitzakis (2008). "Using gene expression to investigate the genetic basis of complex disorders." Hum Mol Genet **17**(R2): R129-134.

Nica, A. C., S. B. Montgomery, et al. (2010). "Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations." PLoS Genet **6**(4): e1000895.

Ning, Z., A. J. Cox, et al. (2001). "SSAHA: a fast search method for large DNA databases." Genome Res **11**(10): 1725-1729.

Nishimura, Y., C. L. Martin, et al. (2007). "Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways." Hum Mol Genet **16**(14): 1682-1698.

Nowak, K. J. and K. E. Davies (2004). "Duchenne muscular dystrophy and dystrophin: pathogenesis and opportunities for treatment." EMBO Rep **5**(9): 872-876.

Ogura, Y., D. K. Bonen, et al. (2001). "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease." Nature **411**(6837): 603-606.

Oksenberg, J. R. and S. E. Baranzini (2010). "Multiple sclerosis genetics-is the glass half full, or half empty?" Nat Rev Neurol.

Oleksiak, M. F., G. A. Churchill, et al. (2002). "Variation in gene expression within and among natural populations." Nat Genet **32**(2): 261-266.

Paigen, K. and P. Petkov (2010). "Mammalian recombination hot spots: properties, control and evolution." Nat Rev Genet **11**(3): 221-233.

Parkes, M., J. C. Barrett, et al. (2007). "Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility." Nat Genet **39**(7): 830-832.

Parsons, D. W., S. Jones, et al. (2008). "An integrated genomic analysis of human glioblastoma multiforme." Science **321**(5897): 1807-1812.

Peltekova, V. D., R. F. Wintle, et al. (2004). "Functional variants of OCTN cation transporter genes are associated with Crohn disease." Nat Genet **36**(5): 471-475.

Pickrell, J. K., J. C. Marioni, et al. (2010). "Understanding mechanisms underlying human gene expression variation with RNA sequencing." Nature **464**(7289): 768-772.

Pikaart, M. J., F. Recillas-Targa, et al. (1998). "Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators." Genes Dev **12**(18): 2852-2862.

Plagnol, V., D. J. Smyth, et al. (2009). "Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13." Biostatistics **10**(2): 327-334.

Plomin, R., C. M. Haworth, et al. (2009). "Common disorders are quantitative traits." Nat Rev Genet **10**(12): 872-878.

Price, A. L., N. Patterson, et al. (2008). "Effects of cis and trans genetic ancestry on gene expression in African Americans." PLoS Genet **4**(12): e1000294.

Price, A. L., N. J. Patterson, et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat Genet **38**(8): 904-909.

Pritchard, J. K. and M. Przeworski (2001). "Linkage disequilibrium in humans: models and data." Am J Hum Genet **69**(1): 1-14.

Prokopenko, I., C. Langenberg, et al. (2009). "Variants in MTNR1B influence fasting glucose levels." Nat Genet **41**(1): 77-81.

Ravasi, T., H. Suzuki, et al. (2010). "An atlas of combinatorial transcriptional regulation in mouse and man." Cell **140**(5): 744-752.

Redon, R., S. Ishikawa, et al. (2006). "Global variation in copy number in the human genome." Nature **444**(7118): 444-454.

Ren, D., M. Li, et al. (2005). "Identification of SH2-B as a key regulator of leptin sensitivity, energy balance, and body weight in mice." Cell Metab **2**(2): 95-104.

Rice, T. K., N. J. Schork, et al. (2008). "Methods for handling multiple testing." Adv Genet **60**: 293-308.

Rioux, J. D., R. J. Xavier, et al. (2007). "Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis." Nat Genet **39**(5): 596-604.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-1517.

Robertson, G., M. Hirst, et al. (2007). "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." Nat Methods **4**(8): 651-657.

Rockman, M. V. and L. Kruglyak (2006). "Genetics of global gene expression." Nat Rev Genet **7**(11): 862-872.

Sanna, S., A. U. Jackson, et al. (2008). "Common variants in the GDF5-UQCC region are associated with variation in human height." Nat Genet **40**(2): 198-203.

Saxena, R., B. F. Voight, et al. (2007). "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels." Science **316**(5829): 1331-1336.

Schadt, E. E. (2009). "Molecular networks as sensors and drivers of common human diseases." Nature **461**(7261): 218-223.

Schadt, E. E., C. Molony, et al. (2008). "Mapping the genetic architecture of gene expression in human liver." PLoS Biol **6**(5): e107.

Schadt, E. E., S. A. Monks, et al. (2003). "Genetics of gene expression surveyed in maize, mouse and man." Nature **422**(6929): 297-302.

Scott, L. J., K. L. Mohlke, et al. (2007). "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants." Science **316**(5829): 1341-1345.

Sebat, J., B. Lakshmi, et al. (2004). "Large-scale copy number polymorphism in the human genome." Science **305**(5683): 525-528.

Selbach, M., B. Schwanhausser, et al. (2008). "Widespread changes in protein synthesis induced by microRNAs." Nature **455**(7209): 58-63.

Shendure, J., G. J. Porreca, et al. (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." Science **309**(5741): 1728-1732.

Shi, J., D. F. Levinson, et al. (2009). "Common variants on chromosome 6p22.1 are associated with schizophrenia." Nature **460**(7256): 753-757.

Silventoinen, K., S. Sammalisto, et al. (2003). "Heritability of adult body height: a comparative study of twin cohorts in eight countries." Twin Res **6**(5): 399-408.

Sladek, R., G. Rocheleau, et al. (2007). "A genome-wide association study identifies novel risk loci for type 2 diabetes." Nature **445**(7130): 881-885.

Smirnov, D. A., M. Morley, et al. (2009). "Genetic analysis of radiation-induced changes in human gene expression." Nature **459**(7246): 587-591.

Soranzo, N., F. Rivadeneira, et al. (2009). "Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size." PLoS Genet **5**(4): e1000445.

Sorrell, J. M. and A. I. Caplan (2004). "Fibroblast heterogeneity: more than skin deep." J Cell Sci **117**(Pt 5): 667-675.

Spector, T. D. and F. M. Williams (2006). "The UK Adult Twin Registry (TwinsUK)." Twin Res Hum Genet **9**(6): 899-906.

Spielman, R. S., L. A. Bastone, et al. (2007). "Common genetic variants account for differences in gene expression among ethnic groups." Nat Genet **39**(2): 226-231.

Spits, H. (2002). "Development of alphabeta T cells in the human thymus." Nat Rev Immunol **2**(10): 760-772.

Srinivasan, L. and M. L. Atchison (2004). "YY1 DNA binding and PcG recruitment requires CtBP." Genes Dev **18**(21): 2596-2601.

Stegle, O., L. Parts, et al. (2010). "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies." PLoS Comput Biol **6**(5): e1000770.

Stern, D. L. (1998). "A role of Ultrabithorax in morphological differences between Drosophila species." Nature **396**(6710): 463-466.

Storey, J. D., J. Madeoy, et al. (2007). "Gene-expression variation within and among human populations." Am J Hum Genet **80**(3): 502-509.

Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proc Natl Acad Sci U S A **100**(16): 9440-9445.

Strachan, T. and A. P. Read (2004). Human molecular genetics 3. London ; New York
Independence, KY, Garland Science ;
Distributed in the USA by Taylor & Francis.

Stranger, B. E., M. S. Forrest, et al. (2005). "Genome-wide associations of gene expression variation in humans." PLoS Genet **1**(6): e78.

Stranger, B. E., M. S. Forrest, et al. (2007). "Relative impact of nucleotide and copy number variation on gene expression phenotypes." Science **315**(5813): 848-853.

Stranger, B. E., A. C. Nica, et al. (2007). "Population genomics of human gene expression." Nat Genet **39**(10): 1217-1224.

Stratton, M. R. and N. Rahman (2008). "The emerging landscape of breast cancer susceptibility." Nat Genet **40**(1): 17-22.

Su, A. I., M. P. Cooke, et al. (2002). "Large-scale analysis of the human and mouse transcriptomes." Proc Natl Acad Sci U S A **99**(7): 4465-4470.

Swanberg, M., O. Lidman, et al. (2005). "MHC2TA is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction." Nat Genet **37**(5): 486-494.

Teo, Y. Y., M. Inouye, et al. (2007). "A genotype calling algorithm for the Illumina BeadArray platform." Bioinformatics **23**(20): 2741-2746.

Teslovich, T. M., K. Musunuru, et al. (2010). "Biological, clinical and population relevance of 95 loci for blood lipids." Nature **466**(7307): 707-713.

Thomas, G., K. B. Jacobs, et al. (2008). "Multiple loci identified in a genome-wide association study of prostate cancer." Nat Genet **40**(3): 310-315.

Todd, J. A., N. M. Walker, et al. (2007). "Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes." Nat Genet **39**(7): 857-864.

Torkamani, A., E. J. Topol, et al. (2008). "Pathway analysis of seven common diseases assessed by genome-wide association." Genomics **92**(5): 265-272.

Ueda, H., J. M. Howson, et al. (2003). "Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease." Nature **423**(6939): 506-511.

Underhill, D. M. and T. Shimada (2007). "A pair of 9s: it's in the CARDs." Nat Immunol **8**(2): 122-124.

Vaisse, C., K. Clement, et al. (1998). "A frameshift mutation in human MC4R is associated with a dominant form of obesity." Nat Genet **20**(2): 113-114.

van der Mei, I. A., A. L. Ponsonby, et al. (2007). "Vitamin D levels in people with multiple sclerosis and community controls in Tasmania, Australia." J Neurol **254**(5): 581-590.

Vaughn, M. W., M. Tanurdzic, et al. (2007). "Epigenetic natural variation in Arabidopsis thaliana." PLoS Biol **5**(7): e174.

Vetter, K. and W. Wurst (2001). "Expression of a novel mouse gene 'mbFZb' in distinct regions of the developing nervous system and the adult brain." Mech Dev **100**(1): 123-125.

Visel, A., M. J. Blow, et al. (2009). "ChIP-seq accurately predicts tissue-specific activity of enhancers." Nature **457**(7231): 854-858.

Visscher, P. M. (2008). "Sizing up human height variation." Nat Genet **40**(5): 489-490.

Visscher, P. M., W. G. Hill, et al. (2008). "Heritability in the genomics era--concepts and misconceptions." Nat Rev Genet **9**(4): 255-266.

Wang, W. Y., B. J. Barratt, et al. (2005). "Genome-wide association studies: theoretical and practical concerns." Nat Rev Genet **6**(2): 109-118.

Wang, Y., C. B. Harvey, et al. (1995). "The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element." Hum Mol Genet **4**(4): 657-662.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.

Weatherall, D. J. (1998). "Pathophysiology of thalassaemia." Baillieres Clin Haematol **11**(1): 127-146.

Weedon, M. N., H. Lango, et al. (2008). "Genome-wide association analysis identifies 20 loci that influence adult height." Nat Genet **40**(5): 575-583.

Weedon, M. N., G. Lettre, et al. (2007). "A common variant of HMGA2 is associated with adult and childhood height in the general population." Nat Genet **39**(10): 1245-1250.

Weintraub, H., R. Davis, et al. (1991). "The myoD gene family: nodal point during specification of the muscle cell lineage." Science **251**(4995): 761-766.

Willer, C. J., S. Sanna, et al. (2008). "Newly identified loci that influence lipid concentrations and risk of coronary artery disease." Nat Genet **40**(2): 161-169.

Willer, C. J., E. K. Speliotes, et al. (2009). "Six new loci associated with body mass index highlight a neuronal influence on body weight regulation." Nat Genet **41**(1): 25-34.

Wittkopp, P. J., B. K. Haerum, et al. (2008). "Regulatory changes underlying expression differences within and between Drosophila species." Nat Genet **40**(3): 346-350.

Workman, J. L. and R. G. Roeder (1987). "Binding of transcription factor TFIID to the major late promoter during in vitro nucleosome assembly potentiates subsequent initiation by RNA polymerase II." Cell **51**(4): 613-622.

Wray, G. A. (2007). "The evolutionary significance of cis-regulatory mutations." Nat Rev Genet **8**(3): 206-216.

WTCCC (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-678.

Xu, J., S. Chen, et al. (2009). "STAT5 mediates antiapoptotic effects of methylprednisolone on oligodendrocytes." J Neurosci **29**(7): 2022-2026.

Yan, H., Z. Dobbie, et al. (2002). "Small changes in expression affect predisposition to tumorigenesis." Nat Genet **30**(1): 25-26.

Yan, H., W. Yuan, et al. (2002). "Allelic variation in human gene expression." Science **297**(5584): 1143.

Yang, J., B. Benyamin, et al. (2010). "Common SNPs explain a large proportion of the heritability for human height." Nat Genet **42**(7): 565-569.

Ye, S., P. Eriksson, et al. (1996). "Progression of coronary atherosclerosis is associated with a common genetic variant of the human stromelysin-1 promoter which results in reduced gene expression." J Biol Chem **271**(22): 13055-13060.

Yeo, G. S., I. S. Farooqi, et al. (1998). "A frameshift mutation in MC4R associated with dominantly inherited human obesity." Nat Genet **20**(2): 111-112.

Yoshida, T., K. Kato, et al. (2010). "Association of genetic variants with hemorrhagic stroke in Japanese individuals." Int J Mol Med **25**(4): 649-656.

Zeggini, E., L. J. Scott, et al. (2008). "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." Nat Genet **40**(5): 638-645.

Zeggini, E., M. N. Weedon, et al. (2007). "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes." Science **316**(5829): 1336-1341.

Zhang, X. and S. Markovic-Plese (2010). "Interferon beta inhibits the Th17 cell-mediated autoimmune response in patients with relapsing-remitting multiple sclerosis." Clin Neurol Neurosurg **112**(7): 641-645.

Zhang, Z. D., A. Paccanaro, et al. (2007). "Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions." Genome Res **17**(6): 787-797.

Zhernakova, A., E. M. Festen, et al. (2008). "Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP." Am J Hum Genet **82**(5): 1202-1210.

Abbreviations

ASE	allele-specific expression
ASW	African ancestry in Southwest USA
BMI	body mass index
bp	base pairs
c-dSNP	causal disease SNP
c-eQTL	causal eQTL SNP
CD	Crohn's 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
cM	centimorgan
CNV	copy number variant
cRNA	copy RNA
cSNP	causal SNP
dSNP	disease SNP
DZ	dizygotic
EBV	Epstein-Barr virus
ENCODE	encyclopedia of DNA elements
eQTL	expression quantitative trait locus
EST	expressed sequence tag
FA	factor analysis
FDR	false discovery rate
GIH	Gujarati Indians in Houston, Texas, USA
GWAS	genome-wide association study/studies
IVT	in vitro transcription
JPT	Japanese in Tokyo, Japan
kb	kilobase
KCL	King's College London
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
MCTA	matched co-twin analysis
MEMN	macrophage enriched metabolic network
MEX	Mexican ancestry in Los Angeles, California, USA

miRNA	microRNA
MKK	Maasai in Kinyawa, Kenya
MS	multiple sclerosis
MuTHER	multiple tissue human expression resource
MZ	monozygotic
NHGRI	National Human Genome Research Institute
PCA	principal component analysis
PT	permutation threshold
QC	quality control
QTL	quantitative trait locus
RNA-seq	RNA sequencing
RTC	regulatory trait concordance
SAM	sentrix array matrix
SNP	single nucleotide polymorphism
SRC	Spearman rank correlation
SSAHA	sequence search and alignment by hashing algorithm
T2D	type 2 diabetes
TF	transcription factor
TP	true positives
TSI	Toscans in Italy
TSS	transcription start site
UGMS	University of Geneva Medical School
UTR	untranslated region
WTCCC	Wellcome Trust Case Control Consortium
WTCHG	Wellcome Trust Centre for Human Genetics
WTSI	Wellcome Trust Sanger Institute
YRI	Yoruban in Ibadan, Nigeria

List of Figures

List of Tables

Appendix

1. **Biopsy technique protocol**

1. The lower abdominal biopsy site is cleaned. Local anesthetic with adrenaline is infiltrated into the pre-inked skin.

2. Stretching the skin perpendicular to the relaxed skin tension lines between thumb and finger either side of the area to be sampled, the punch blade is placed on the skin and rotated under gentle pressure by rolling it between the thumb and finger using a twisting drilling action.

3. One should penetrate to the level of the fat layer to achieve a full thickness skin biopsy specimen. The specimen should be weighed, cut in half, and stored immediately in liquid nitrogen

4. The specimen will either float up on the fat layer or can be gently lifted using a skin hook or gently applied forceps to allow specimen collection by cutting through the fat layer using a scalpel or sharp scissors.

5. Further fat can be obtained by careful dissection of the fat layer using forceps and scalpel. The fat sample should be weighed and immediately stored in liquid nitrogen.

6. Haemostasis can be achieved with direct pressure and/or interrupted sutures. Both absorbent and non-absorbent sutures can be used in a layered closure for larger punch defects.

7. The resultant defect can be allowed to heal by secondary intention as an alternative method although optimal haemostasis and cosmesis as well as reduced healing time are usually seen with sutured wounds.