6 DISCUSSION

6.1 GENETIC VARIATION IN GENE REGULATION

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

6.1.1 Genetic interactions with an impact on gene expression

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

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

6.1.2 Fine-scale architecture of the cis regulatory landscape

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

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

6.1.3 Cell type specificity of regulatory variation

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

6.2 Overlap of GenCord eQTLs with disease and complex trait SNPs

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

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

6.2.1 Crohn disease

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

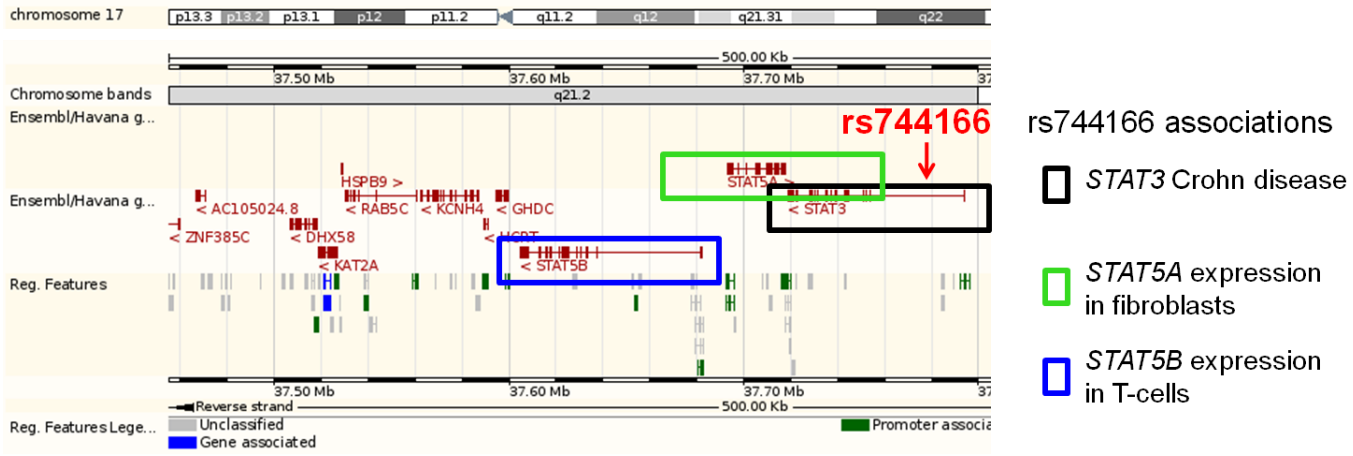

Figure 41. Overlap between disease-associated SNPs and eQTLs in the case of Crohn disease (CD). rs744166 is associated with CD and is an eQTL in fibroblasts and T-cells. *STAT3* (black box) is the GWAS-reported gene, *STAT5A* and *STAT5B* are the expression-associated genes in fibroblasts (green box)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

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

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

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

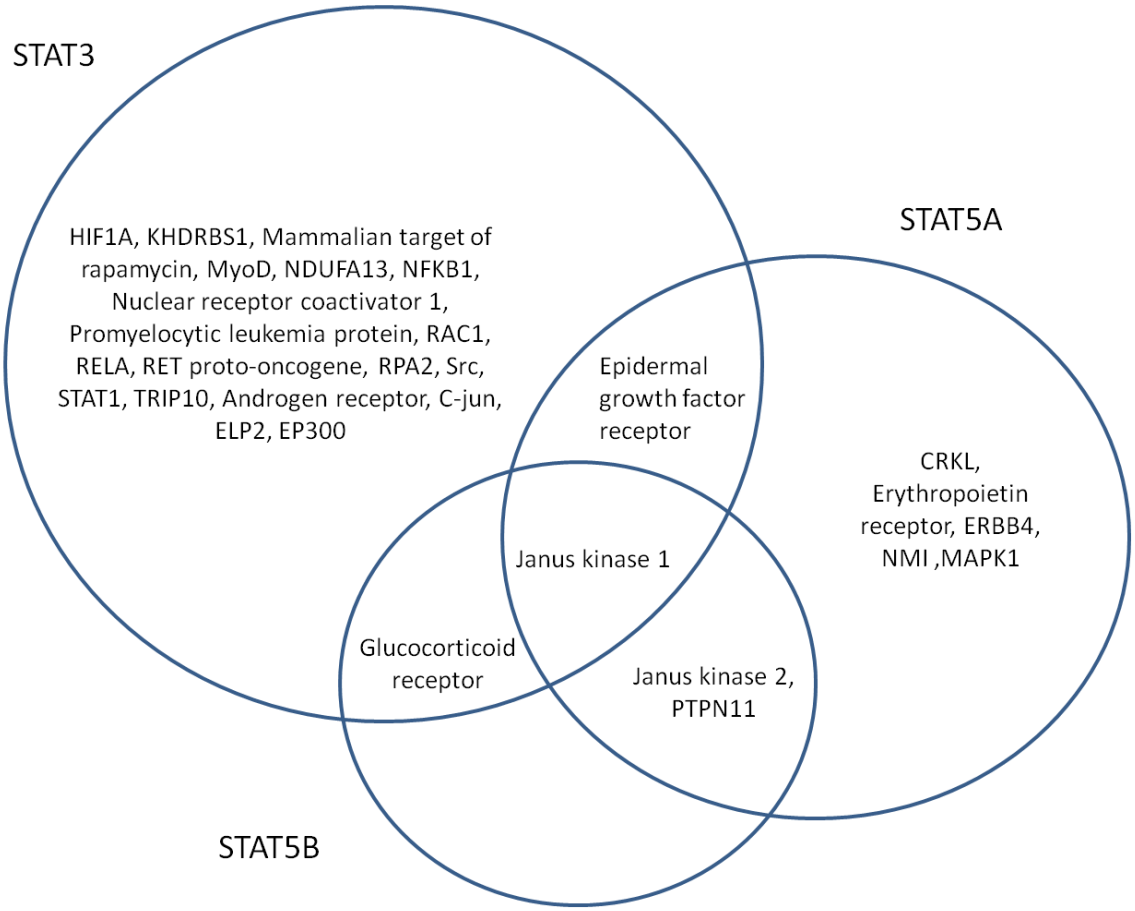

Figure 42. STAT3, STAT5A and STAT5B interacting proteins. All gene products interact with Janus kinase 1, a component of the JAK-STAT pathway. STAT3 and STAT5A interact with Epidermal growth factor receptor. STAT3 and STA5B interact with Glucocorticoid receptor. STAT5A and STAT5B interact with Janus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

6.2.2 Bipolar disorder

An association to bipolar disorder was detected for rs4130590 (Ferreira, O'Donovan et al. 2008), but no gene has been reported for this variant. In LCLs rs4130590 is an eQTL (0.01 permutation threshold) associated with the expression of genes *SLC2A8* (*GLUT8*) an insulin-regulated facilitative glucose transporter and ZNF79 a zinc finger protein (Figure 43). GLUT8 catalyzes transport of sugars or sugar derivatives through intracellular membranes. *GLUT8* knockout mice showed strong evidence for hyperactivity and increased arousal. Additionally mild alterations were observed in brain (neuronal cell increased proliferation in the hippocampus, hyperactivity), heart (impaired transmission of electrical wave through the atrium), and sperm cells (reduced number of motile sperm cells) (Schmidt, Joost et al. 2009). It has been suggested that the behavioural alterations observed stem from dysfunctions in neuronal processes arising as a consequence of defects in glucose metabolism (Schmidt, Gawlik et al. 2008). A study of closely-related gene *GLUT3* showed that its deficiency in mice led to abnormal spatial learning and working memory, abnormal cognitive flexibility with intact gross motor ability, electroencephalographic seizures and perturbed social behaviour. The authors state that this phenotypic expression is unique, as it combines neurobehavioral and epileptiform characteristics of autism spectrum disorders. Furthermore, they point out that metabolic adaptations were observed in GLUT3-deficient mice, including an increase in neuronal GLUT8 levels, microvascular/glial GLUT1, and monocarboxylate transporter isoform 2 concentrations. They suggest that this deficiency forms a possible novel genetic mechanism for developmental disorders, such as the neuropsychiatric autism spectrum disorders (Zhao, Fung et al. 2009).

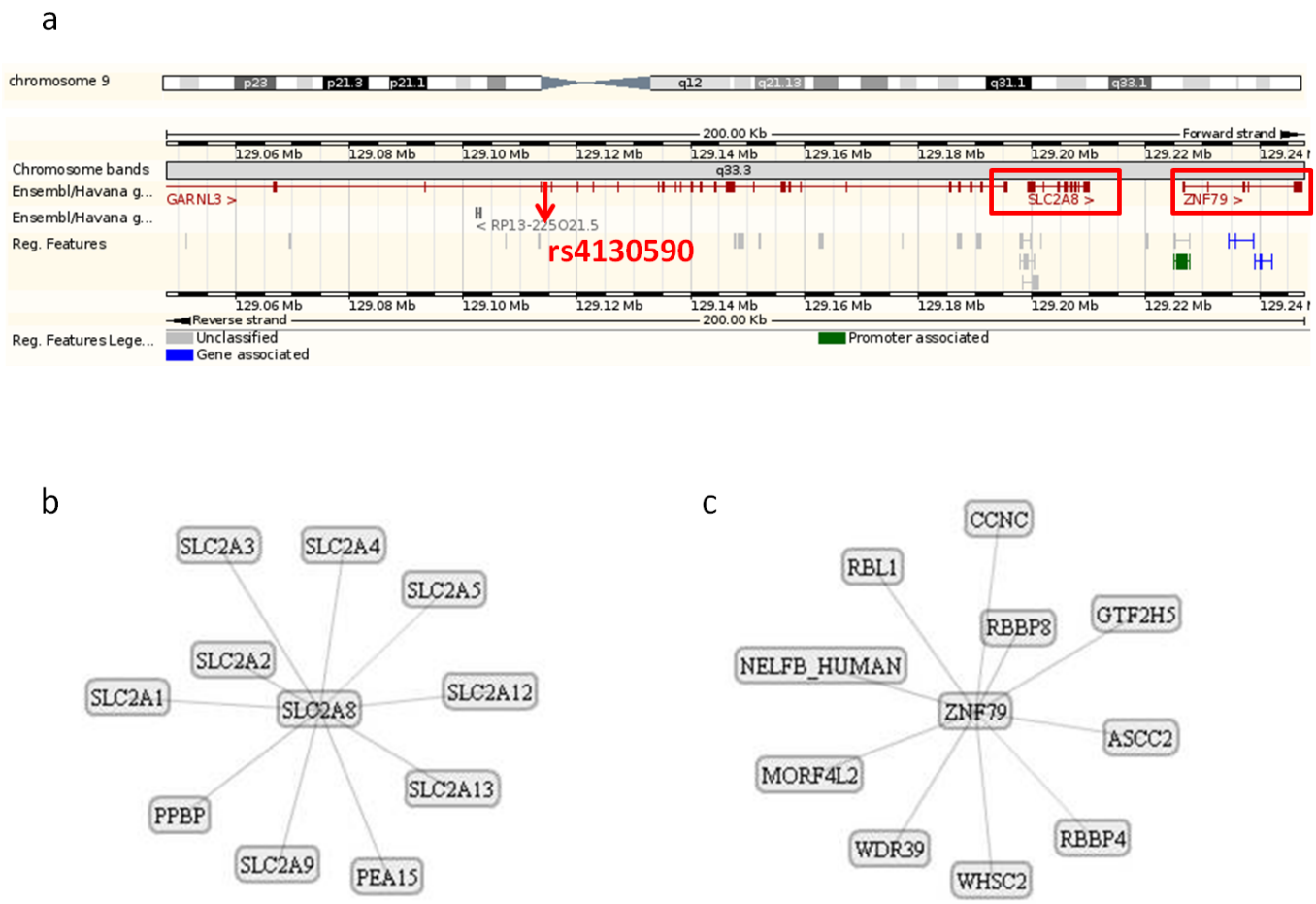

Figure 43. Disease GWAS revealed association of rs4130590 with bipolar disorder, but no genes have been reported for this variant to date. a) In such cases integrating results from disease and expression association studies may prove very useful: rs4130590 is an eQTL (0.01 permutation threshold) associated with expression of *SLC2A8 (GLUT8)* and *ZNF79* genes in LCLs (red boxes). Figure created using the Ensembl genome browser. Using this as a starting point, a possible next step is to explore gene interactions (in this case using GENETWORK (http://www.genenetwork.nl/) (Franke, van Bakel et al. 2006)) to uncover further candidate genes interacting with **b)** *SLC2A8* and **c)** *ZNF79*.

6.2.3 Weight and body mass index

Another example of GWAS SNP-eQTL overlap is rs7481311, a GWAS-reported variant associated with weight and body mass index (BMI) (Thorleifsson, Walters et al. 2009). The GWAS-reported gene is *BDNF*, a member of the nerve growth factor family,

induced by cortical neurons and necessary for survival of striatal neurons in the brain. Its expression is reduced in Alzheimer and Huntington disease patients and it may also play a role in regulation of stress response and in the biology of mood disorders (Zuccato and Cattaneo 2009). In LCLs, rs7481311 is an eQTL (0.01 permutation threshold) associated with the expression of gene *LIN7C*, a homologue of *C.elegans* lin-7 PDZ domain protein (Figure 44). PDZ domains are common structural domains of 80-90 amino-acids, which help anchor transmembrane proteins to the cytoskeleton and hold together signalling complexes. Further investigation of the role of LIN7C may prove informative for the traits in question.

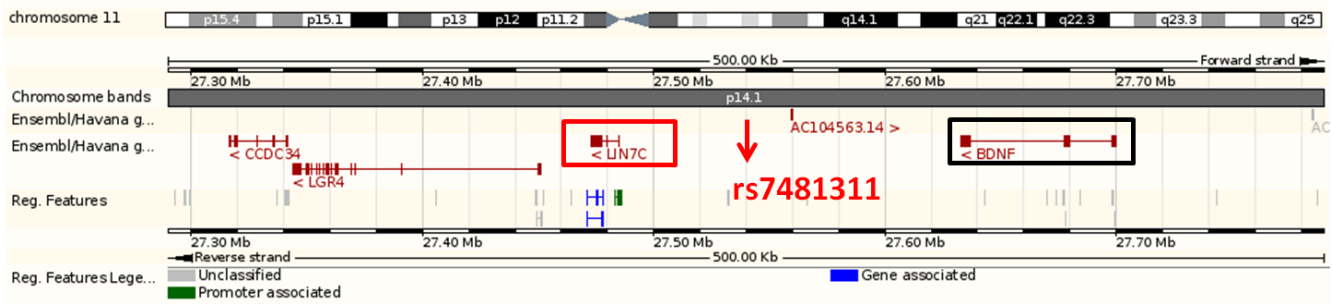

Figure 44. rs7481311 is a GWAS SNP associated with weight and body mass index, and an eQTL (0.01 permutation threshold) associated with the expression of the *LIN7C* gene in LCLs. The GWAS-reported gene is *BDNF* (black box), a member of the nerve growth factor family, whose expression is reduced in Alzheimer and Huntington disease patients. *LIN7C* (red box) is a PDZ domain containing gene and shows an expression association with rs7481311. Figure created using the Ensembl genome browser.

6.2.4 HDL cholesterol and triglycerides

GWAS have identified rs1800775 as a variant associated with HDL cholesterol (Kathiresan, Melander et al. 2008) and triglycerides (Saxena, Voight et al. 2007). The GWAS-reported gene, *CETP*, has a role in the transfer of cholesteryl esters between lipoproteins. The identical SNP is an eQTL in fibroblasts (0.01 permutation threshold) associated with expression levels of *PLLP*, a gene encoding a tetraspan plasma

membrane proteolipid (plasmolipin) which is thought to participate in ion transport and addition of plasmolipin to lipid bilayers (Figure 45).

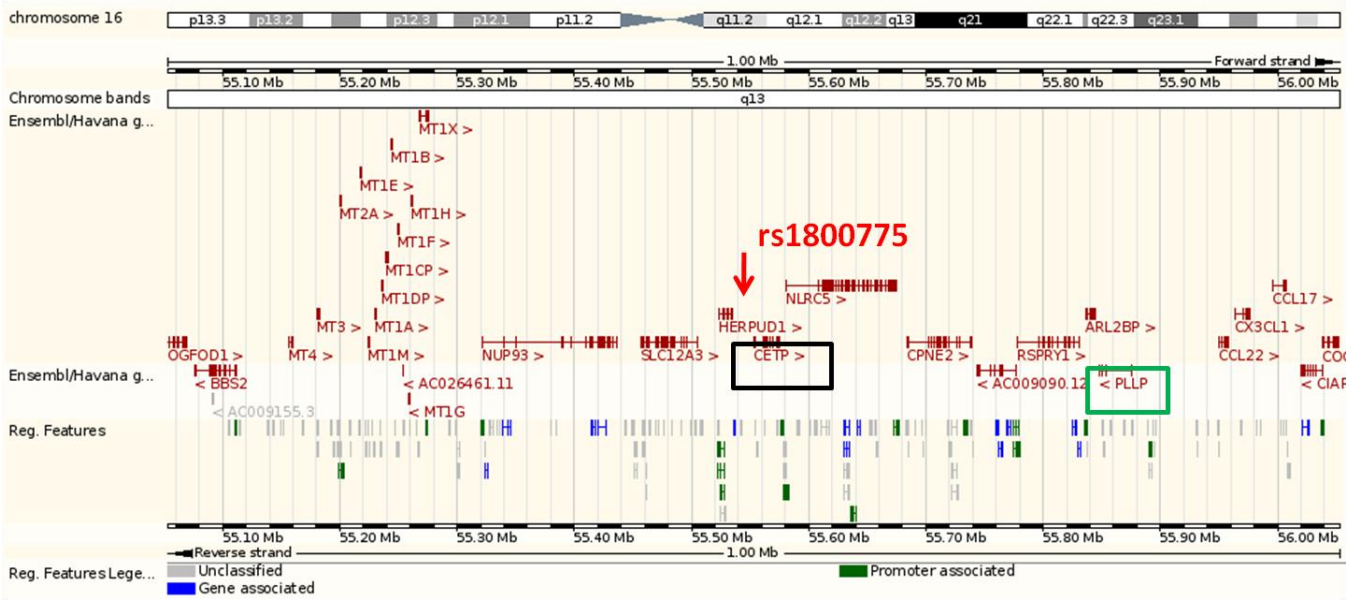

Figure 45. rs1800775 is associated with HDL cholesterol and triglycerides and is an eQTL (0.01 permutation threshold) in fibroblasts. The GWAS-reported gene *CETP* (black box) has a role in the transfer of cholesteryl esters between lipoproteins. *PLLP* (green box), the expression-associated gene, has a role in ion transport and addition of plasmolipin to lipid bilayers. Figure created using the Ensembl genome browser.

6.2.5 The value of integrating disease and expression association data

The examples outlined above illustrate how integrating information from disease and expression GWAS can serve as a first step in complementing disease and trait association studies with functional information. At the time of writing, an instance of eQTL and unpublished GWAS SNP overlap was identified and has enabled identification of a candidate gene in a study interrogating migraine headaches. rs1835740 is the top well-replicating SNP in four European cohorts (V.A. personal communication). This variant is also an eQTL associated with expression levels of the

MTDH gene in LCLs (0.001 permutation threshold). MTDH1 downregulates the excitatory amino acid transporter EAAT2 (GLT1) (Lee, Jeon et al. 2009), which is responsible for clearing excess glutamate from synapses. Given this overlap, exploring the function of this gene has become a priority for this study.

To date, GWAS rely chiefly on proximity criteria and prior biological knowledge to identify possible candidate genes. Combining prior knowledge with clues from gene expression studies is likely to reveal additional components of complex biological pathways and adds a quantitative dimension to biological pathways. Assimilating growing evidence on cis and trans gene regulation with clues from disease GWAS will therefore put us in a much better position to understand the mechanisms governing pathogenesis and will enable better prognosis and treatment.

6.3 Future Directions

Although this thesis has dealt with cis gene regulation, trans effects are thought to make an important contribution to control of gene expression. It is likely that such effects have been largely underestimated mostly because studies to date have been underpowered. Ongoing analyses of HapMap Phase 3 and GenCord data carried out in our group have shown substantial trans regulatory effects, with a fraction of trans eQTLs being shared across populations, but also across cell types.

In this thesis I have highlighted the biological significance of interrogating gene expression in multiple cell types. Similarly, it is also expected that studying expression during different developmental stages, but also following a range of environmental stimuli, will also yield important information about gene regulation. Quantifying the contribution of environmental factors is crucial and it is very likely that joint estimation of genetic and environmental factors will put genetic contributions in context and help dissect the genetic component of gene expression. In his 2008 review, Gibson (2008, pg

575) states that "environmental components are prevalent and ignoring them misses more than half the point if the objective is to understand the origin of variation for complex traits". The environment plays a major role in determining plasticity of genomic responses and addressing gene-environment interactions will not only contribute to elucidating phenotypic variation, but will also help unravel mechanisms of genetic adaptation and evolution.

Cis regulatory sequences are considered an important component of the genetic basis for adaptation (Wray 2007) and evolutionary biologists have argued that a number of phenotypic traits (morphological, physiological and behavioural) are more likely to result from these differences rather than from differences in coding regions. In many instances, natural selection has been shown to operate more efficiently on cis regulatory variants (Wray 2007) and for example mutations in cis regulatory regions of *PDYN*, a gene whose decreased expression in humans is associated with schizophrenia and bipolar disorder, show signatures of positive selection (Rockman, Hahn et al. 2005).

A better grasp of gene regulation pathways is also likely to shed light on the pleiotropic action of genes (i.e. a single gene affecting multiple phenotypes). There is growing evidence that pleiotropy is common and occurs between traits that have never been considered as functionally related (Flint and Mackay 2009). Elucidating common regulatory pathways is likely to help us understand pleiotropic effects that influence phenotypes in different directions and as a consequence do not result in significant correlations between the traits. Furthermore, considering the joint effects of pleiotropy and epistasis is important, as pleiotropic effects can themselves be genetically variable when differences in epistatic interactions occur between loci with an impact on multiple traits (Mackay, Stone et al. 2009). Similarly, elucidating regulatory pathways is expected to be informative for understanding the structure of biological pathways and networks.

Ultimately, integrating all this information will help us understand how genetic variation impacts on developmental processes and shapes natural range phenotypic variation, as well as disease.