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.