Chapter 8

Discussion

8.1 Past

The work I describe in this thesis relied on a candidate gene approach to attempt to identify genetic variation involved in syndromes of insulin resistance and common complex type 2 diabetes and related traits. When I started my PhD in April 2005, genome-wide association studies (GWAS) were not feasible because of the prohibitive costs of large-scale genotyping and the absence of genome-wide genotyping arrays based on completed HapMap data. Instead two methods, linkage and candidate gene association, had already been used to identify disease genes. As described in Chapter 1, linkage studies in type 2 diabetes were disappointing as positive results did not generally replicate in other studies. This was also true across many other complex diseases and traits. For example, genomic loci identified through candidate gene or genome-wide linkage scans for coronary artery disease and myocardial infarction were largely non-overlapping and only one robustly replicated gene, *ALOX5AP*, has come out of these analyses (Hamsten and Eriksson 2008; Helgadottir et al. 2004). In type 1 diabetes the human leukocyte antigen (HLA) class II genes (thought to explain up to half of the heritability) were reproducibly linked to risk of disease, but linkage analysis was not successful in identifying loci with more modest effects (Smyth et al. 2006). These studies demonstrate that linkage analysis tended to overlook the very small effect sizes expected of complex disease genes.

By contrast, candidate gene association studies were suggested to have more statistical power to detect modest effects of genetic loci on complex disease (Risch and Merikangas 1996). Still, many reported susceptibility loci showed inconsistent evidence for association between studies, casting doubt on whether there were in fact many common causal alleles to be found. However, at the start of my PhD a growing number

of research groups were realising the importance of collecting large well-phenotyped sample sizes and of careful study design (such as appropriate matching of cases and controls to avoid the effects of population substructure) for replicating results and identifying novel loci. Also, awareness of the pitfalls of multiple hypothesis testing yielded more conservative interpretations of nominal significant findings. In type 2 diabetes, candidate gene association approaches succeeded in identifying and verifying several risk loci including *PPARγ* and *KCNJ11* (Altshuler et al. 2000; Gloyn et al. 2001). Association studies also helped to identify those type 1 diabetes susceptibility genes with small impacts on disease risk relative to the HLA locus. For example, the *INS*, *PTPN22*, and *CTLA4* genes were found to be associated with type 1 diabetes (Bell et al. 1984; Bottini et al. 2004; Nistico et al. 1996). Candidate gene studies also proved successful at identifying genes underlying more extreme phenotypes that demonstrated Mendelian patterns of inheritance, such as severe insulin resistance (Barroso et al. 1999; George et al. 2004; Savage et al. 2002).

Given the effectiveness of candidate gene studies for detecting genetic loci causing Mendelian disease and predisposing to common complex disease, and given the prohibitive costs of genome-wide approaches, I adopted a hypothesis-driven rather than a hypothesis-free study design to identify genes involved in severe insulin resistance and type 2 diabetes traits. I selected candidate genes based on their known or putative role in insulin action and/or secretion in animal models and human phenotypes.

In Chapter 3 I describe investigation of the lipin gene family, so chosen because Lpin1 is responsible for two independent mouse models of lipodystrophy and insulin resistance (Peterfy et al. 2001), and its expression levels correlate with insulin sensitivity and adiposity in mice and humans (Yao-Borengasser et al. 2006). I screened *LPIN1* in 158 patients with syndromes of severe insulin resistance, including 23 cases of

lipodystrophy, but detected no fully penetrant pathogenic mutations (Fawcett et al. 2008). *LPIN1* common variation was not statistically associated with insulin sensitivity in a population-based cohort but SNPs were nominally associated with BMI, blood pressure, cholesterol levels and risk of hypertension (Fawcett et al. 2008). These associations will need to be confirmed in further cohorts.

As presented in Chapter 4, I also screened members of the mTORC1 and mTORC2 complexes and *AS160*, which are important downstream components of the insulin signalling cascade, in insulin resistant patients. A nonsense mutation in *AS160* was shown to impair insulin-stimulated GLUT4 translocation and segregated with high peak-to-fasting insulin ratios in a pedigree of six genotyped individuals with five affected members. I recommend that *AS160* should be screened in families with similar syndromes of insulin action.

My interest in the *PARL* gene, which was identified during a screen for genes differentially expressed in obese, type 2 diabetic Israeli sand rats, was stimulated by reports of an association between a nonsynonymous SNP in *PARL* and plasma insulin levels in a US cohort (Walder et al. 2005). As described in Chapter 5, I did not replicate this association in UK populations (Fawcett et al. 2006). This demonstrates the importance of replication of association results to distinguish between true associations and statistical artefacts.

Finally, in Chapter 6 I described a large scale candidate gene association study of genes involved in pancreatic ß-cell function. This strategy lead to the discovery of another robustly replicated type 2 diabetes susceptibility gene, *WFS1* (Sandhu et al. 2007). Common variation within a linkage disequilibrium block encompassing most of the gene was associated with type 2 diabetes in different populations including UK, Ashkenazi and Swedish case-control studies (Franks et al. 2008; Sandhu et al. 2007).

8.2 Present

Four years later the field of complex disease genetics has undergone a revolution due to the availability of completed HapMap phase I and II data, relatively cheap genotyping technology, and the formation of large consortia able to pool samples to generate larger study sizes. Such advances made GWAS not only feasible but successful in identifying new complex disease loci. For example, an early GWAS identified a strong association between risk of age-related macular degeneration and an intronic SNP in the complement factor H (*CFH*) gene (Klein et al. 2005). Resequencing and fine-mapping lead to the identification of a non-synonymous SNP (Y402H) which was corroborated by two independent groups published in the same issue of Science (Edwards et al. 2005; Haines et al. 2005). Another early GWAS also identified a novel type 1 diabetes locus in the interferon-induced helicase (*IFIH1*) region (Smyth et al. 2006). As examined in Chapter 1, GWAS have also detected over a dozen type 2 diabetes susceptibility genes with robust evidence for replication. Towards the end of my PhD I was able to use publicly available data from GWAS to include in meta-analyses with my own candidate gene data. This not only increased the power of my studies to detect susceptibility loci but provided an independent dataset that could be analysed for evidence of replication.

Though GWAS have detected a number of reproducible susceptibility loci, this approach still has important limitations. In genome-wide studies, the hundreds of thousands of tests lead to substantial type 1 error but, on the other hand, adjustment for multiple tests results in a very stringent significance level ($P \sim 10^{-7}$), and inflation of the type 2 error rate. In other words, genome-wide studies will tend to capture the “low hanging fruit”, SNPs with larger effects on type 2 diabetes in the populations tested, but may ignore truly associated variants with more modest effects on disease risk. For example, the

candidate gene study described in Chapter 6 did not impose a very strict P value cut-off in discovery cohorts and consequently modest associations between WFS1 SNPs and type 2 diabetes were pursued in replication cohorts. However, WFS1 was not prioritised for replication in the first wave of GWAS. Another limitation of genome-wide approaches stems from the use of custom-made genotyping arrays, such as Affymetrix and Illumina SNP chips, which only cover between ~40-70% Phase II HapMap SNPs with $r^2 \geq 0.8$ (Dong et al. 2007). Therefore, many genes and functional non-coding regions will not be well covered in genome-wide association studies.

In contrast, more thorough characterisation of the variation in a given region can be achieved by a candidate gene study, which might resequence the gene to discover novel variation and population-specific patterns of LD. Though this can be an expensive and time-consuming study design for susceptibility gene discovery, it can act as a complimentary approach to GWAS. Indeed, candidate gene studies are beginning to be used to extend results from genome-wide analyses by focusing on the effects of known susceptibility genes in distinct subgroups, such as different ethnic groups or cohorts with data on metabolic quantitative traits, and on SNPs with less dramatic *P* values in GWAS.

8.3 And future

There are several outstanding challenges remaining in the field of complex disease genetics. Firstly, SNPs that have been associated with type 2 diabetes and other complex diseases to date are not necessarily causal variants but instead they represent genomic regions which are sometimes hundreds of kilobases away from known genes. The fine-mapping of association signals and the identification of true causal variants will be a necessary but potentially arduous task, especially when LD in the region of association is strong making it difficult to distinguish between the effects of different variants on disease risk. I experienced this difficulty while attempting to refine the

association signal between *WFS1* and type 2 diabetes (described in Chapter 7). I sequenced exons, splice junctions and conserved non-coding regions of *WFS1* in a subset of cases and controls to discover novel variation and genotyped tagging SNPs in 854 cases and 1242 controls. Though several previously untested SNPs were nominally associated with type 2 diabetes risk, high correlation between SNPs made it difficult to refine the signal any further. However, I did detect a nominal association between a previously untested 3'UTR variant and type 2 diabetes risk that was stronger than rs10010131 in Cambridgeshire and EPIC alone. This will require replication in further cohorts but could potentially impact protein function through mRNA stability, processing and transport. Identification of the true functional variant(s) involved in complex disease may require genotyping of variants across the region in populations characterised by different and/or weaker patterns of linkage disequilibrium and/or extensive resequencing efforts to cover the entire region between recombination hotspots flanking the association signal to make sure all putative disease variants are tested.

Secondly, an important goal of the complex disease genetics research community is to expand knowledge of the underlying biology of disease and non-disease states, and through this the identification of genes and pathways that could be targeted for therapeutic intervention. Type 2 diabetes susceptibility genes *PPARγ* and *KCNJ11* are proofs-of-principle as they are also targets for thiazolidinediones (insulin sensitising drugs) (Berger et al. 1996) and sulphonylureas (insulin secretagogues) (Sturgess et al. 1988) respectively. However, the biological function of certain predisposing genes and how they contribute to disease remains elusive. For example, very little was known about the fat mass and obesity associated gene, *FTO*, when it was detected in a GWAS. Bioinformatics analysis of the *FTO* sequence revealed that it shared motifs with members of the Fe(II)- and 2-oxoglutarate-dependent dioxygenase family, and appeared

to have a role in DNA methylation (Gerken et al. 2007; Sanchez-Pulido and Andrade-Navarro 2007). Studies of *FTO* expression showed its presence in human adipose tissue, where the protective genotype was associated with higher rates of lipolysis (Kloting et al. 2008; Wahlen et al. 2008). In rodent models *FTO* mRNA is also abundant in hypothalamic nuclei and its expression is regulated by nutrional status, suggesting a possible role in the regulation of energy balance (Fredriksson et al. 2008; Gerken et al. 2007). The story of FTO shows how results from GWAS can inspire new studies into the biological function of a gene, and potentially identify new pathways that can be targeted for drug discovery. Further studies in model organisms will of course be important in elucidating gene function, and may be required to discover the function of non-coding intergenic SNPs by chromosome engineering (Wallace et al. 2007).

A third challenge will be to elucidate the role of rare variants in complex disease. Candidate and genome-wide association study designs have thus far focused on common alleles (MAF>0.05) which do not effectively tag rarer variants. However, it is perfectly plausible that rare variants with moderate effects on disease risk that are somewhere between the effect size seen for common SNPs (OR<1.4) and fully penetrant Mendelian disease mutations, may collectively contribute a substantial portion of inherited susceptibility to complex disease. Such variants have already been found to influence risk of colorectal adenomas, reviewed in (Bodmer and Bonilla 2008), circulating lipid levels (Cohen et al. 2004; Cohen et al. 2006; Romeo et al. 2007), schizophrenia (Walsh et al. 2008) and blood pressure (Ji et al. 2008). These studies employed deep resequencing to compare the frequency of newly discovered and known rare variants in candidate genes between disease cases and controls as well as individuals at opposite extremes of continuous trait distributions. In a review of rare variant analyses in the literature, Bodmer and Bonilla show that odds ratios for rare variants are generally >2,

with an average of 3.74 (Bodmer and Bonilla 2008). As sequencing technology becomes faster and cheaper, future studies of this kind may test the entire genome rather than limiting themselves to candidate genes. I performed deep resequencing of *WFS1* in 1235 type 2 diabetes cases and 1668 controls and identified 83 rare (MAF<0.01) missense and nonsense changes. However, there was no statistically significant association between any single variant and disease status, and the cumulative frequency of these variants, weighted according to their likelihood of having deleterious effects on protein function, was not enriched in cases compared to controls. Furthermore, I found two nonsynonymous variants with moderate frequencies in the cohort (MAF>0.01 and <0.05) but these showed no statistical association with disease. Much larger sample sizes would be required to detect very modest effects (OR < 1.2) of rare variants on complex disease susceptibility. However, such variants are unlikely to importantly contribute to disease at a population-wide level and therefore such studies may not be cost-effective.

Finally, many studies of Mendelian disease and complex traits, including my own, have tended to focus on point mutations, SNPs and small insertions/deletions rather than larger structural genetic changes. CNVs are ubiquitous in the genome and are already reported to influence a few complex disease phenotypes such as familial breast cancer, autism and autoimmune diseases such as systemic lupus erythematosus (Fanciulli et al. 2007; Frank et al. 2007; Sebat et al. 2007; Willcocks et al. 2008). Structural variations have been shown to explain a substantial portion of the variability in gene expression in HapMap samples, providing a possible mechanism through which many CNVs might impact disease risk (Stranger et al. 2007). Global maps identifying CNV regions, such as the study finished in 2006 identifying ~1500 regions (Redon et al. 2006), will be important for future association studies of type 2 diabetes. Furthermore, a structural

variation analysis group will be looking for both common and rare (MAF<0.01) CNVs as part of the 1000 genomes project (Hayden 2008).

Copy number variations (CNVs) are also a well established cause of Mendelian disease. The development of the paired-end mapping technique for detection of structural variations will aid high-throughput screening of disease cases (Korbel et al. 2007).

Another form of variation understudied in my screening analyses and others are functional non-coding variants. The identification of functional non-coding regions for screening in disease phenotypes could therefore help to identify Mendelian disease mutations. Such studies should be greatly aided by completion of the ENCODE (ENCyclopedia Of DNA Elements) project which aims to catalogue structural and functional components of the genome, including non-coding and regulatory elements (Birney et al. 2007). This information could be used to screen non-coding regions of candidate genes in the severe insulin resistance cohort.

In conclusion, awareness of the caveats of candidate gene association studies and the introduction of genome-wide association studies in recent years has led to the detection of many type 2 diabetes susceptibility loci. Investigators in the field of complex disease genetics still face important challenges, namely the identification of causal variation in associated genomic regions, the investigation of rare variants for impact on disease risk, and the biological mechanisms behind disease association. With the help of new sequencing and genotyping technology, the forging of collaborations between research groups to share resources and expertise, and enthusiasm for elucidating the function of new susceptibility genes and variation, the next few years should see progress towards understanding the genetic architecture of common type 2 diabetes and Mendelian forms

of insulin resistance, as well as the biological pathways underlying their development. It is hoped that with this knowledge we can target therapies to effectively combat these devastating diseases.