Chapter 6

Conclusions

6.1 Connections and themes

The projects described in this thesis, while all focused towards locus discovery, have been more or less distinct. I have investigated the historical and statistical foundations of complex disease genetics in chapters 1 and 2, studied the utility of genotype imputation in Chapter 3, described the discovery of new inflammatory bowel disease (IBD) risk loci via custom genotyping in Chapter 4, and investigated genetic risk factors in a multiplex IBD family in Chapter 5. Each chapter contained a unique dataset, and in each case I investigated this dataset using the methods most relevant to that data type.

Despite this, certain topics have come up multiple times throughout the chapters. For instance, the importance of the Crohn's disease *NOD2* locus has come up in almost every chapter: as an important development in the history of disease genetics in Chapter 1, one of the few common loss-of-function risk variants in Chapter 3, as a pilot locus for fine-mapping in Chapter 4, and as an important contribution of risk in the family discussed in Chapter 5.

However, beyond specific topics like this, certain wider themes have emerged as relevant to all the individual projects, and possibly to the field of complex disease genetics as a whole.

One of the major themes has been the economics of experimental design in disease genetics. In this field we do not design the "ideal" experiment, we design the experiment that has the highest power or utility given the availability of datasets, technology, samples and statistical methods. Sometimes the driving considerations have been explicitly financial: the design of the Immunochip was born from economic arguments about the relationships between sample size, power, unit cost and bulk buying. Other experiments have been driven by the exploitation of unique sample resources, such as the relatively low-cost sequencing of very large multiplex families. Still others have been about leveraging external datasets: genotype imputation using sequencing datasets is a prime example of statistical methods, combined with external datasets, adding substantial value to existing studies. Success in complex disease genetics is largely dependent on being able to recognise the potential for such "high-value" studies as new resources become available.

A related theme is the appearance of "next-generation" datasets that can add value to the genetic data used in locus discovery. The studies in Chapters 3 and 5 would have been essentially impossible without the use of genome-wide external datasets of population sequencing and genome function respectively. While the list of 163 loci in Chapter 4 would probably have been obtainable without external datasets, the transformation of this locus list into biological hypotheses would have been essentially impossible. In essence, these datasets allow a quantitative understanding of biological function throughout the genome, and as they improve and become more integrated with the genetic datasets, our ability to discover and characterise disease associations can only grow.

Another theme has been the role of theory in complex disease genetics. The statistical techniques described in Chapter 2 have been used throughout this thesis. The scale of data generated in modern complex disease genetics means that these techniques are the only way to analyse this data, making understanding the assumptions and models implied by these methods especially important. However, we have also seen the role of biological theory in the analysis of this data. In Chapter 5, it was knowledge of biological function (both genome function and gene expression) that allowed us to reduce the number of candidate causal mutations to a manageable number. This union of statistical and biological theory was particularly important in Chapter 4, where both bioinformatic interrogation and biological insight were required to transform a long list of loci into a set of more concrete biological hypotheses. I believe that this integration between biological and statistical theory will be increasingly important as the field moves forward, as I will discuss later in this chapter.

Finally, a continued theme throughout this thesis has been the historical trajectory of locus discovery. In Chapter 1 I took an unashamedly teleological view of the field, describing how experiments (if not discoveries) are often predicted far in advance. While I do not believe that science in general is an onwards march of progress, there have always been certain discoveries that have been foreseen long before they came pass. Given sufficient time, the source of the Nile will be discovered, the moon will be walked on, and the genetic differences that lead to disease susceptibility will be identified. The question is not if but when.

In this spirit of teleology, I will spend the rest of this chapter considering the future of locus discovery, looking ahead to the next crop of experiments that are already underway, and those that will appear in the more distant future. We will see the same themes I have discussed above appear again, in many cases becoming more important as the scale of the data and the level of data integration in human disease genetics increases.

6.2 A next-generation GWAS using low-coverage sequencing

As I have discussed in this thesis, we can attempt to map low-frequency and rare disease alleles in a variety of ways. I have presented projects that use imputation, custom genotyping within known loci and sequencing in multiplex families to identify low-frequency risk variants. As discussed in the introduction, other groups are also using exome sequencing, targeting sequencing of candidate loci and custom genotyping of low-frequency coding variation to the same end. All of these experiments look at a restricted class of variants or samples. Sequencing in families can only identify variants present in those families, and imputation can only identify variants that are in the reference sets and can be inferred from common variation. Targeted sequencing or genotyping is only as powerful as the selection of targets, and any "surprising" variants (e.g. large effect size regulatory variation) will be missed.

A more "hypothesis-free" way of discovering low-frequency risk variation is to extend the genome-wide approach that has been so successful in GWAS using next-generation sequencing. Complete (high-coverage) sequencing is currently too expensive to produce the sample sizes required, but the imputation framework described in Chapter 3 can be applied to incomplete (or low-coverage) sequencing data to allow us to infer the genotypes of nearly all variants in the genome at a fraction of the cost of complete sequencing. This opens up the possibility of affordable whole-genome sequencing of case and control cohorts with sample sizes large enough to detect low-frequency associations.

In order to put this technique into practice, two research groups from the

Wellcome Trust Sanger Institute, in collaboration with the UK IBD Genetics Consortium (UKIBDGC), designed a large low-coverage sequencing project of IBD cases. The project is funded by the Wellcome Trust and the MRC, and will sequence 5000 cases (3000 CD and 2000 UC). This data will be combined with the 4000 UK10K cohort controls to produce (to my knowledge) the largest whole-genome sequencing case-control dataset ever produced.

The cases picked for sequencing underwent a detailed selection procedure to maximise the likelihood of detecting associations. Samples were selected for sequencing on the basis of family history (at least three affected family members, or one affected first-degree relative), age of onset (diagnosed before age 17) and severity of disease (more than three surgical interventions). Other samples were prioritised on the basis of having attached functional data, including gene expression and epigenetic assays, in order to allow functional studies to be performed using the whole-genome sequencing data.

The aim of the dataset is to identify suggestive ($p < 10^{-5}$) associations to replicate in a larger cohort. The experiment will have 77% and 55% power to detect low-frequency (MAF of 1%) associations of intermediate effect size (OR = 2) that are unique to CD or UC respectively. For shared associations this power rises to 85%, with 26% power for risk alleles with a frequency below 0.5%. There are more than 7000 additional UKIBDGC cases (2500 CD and 4500 UC) ready for use in replication, with others to come, which if combined with a large number of controls will have high power to replicate associations down to at least 0.5%.

While sequencing for the final dataset is not yet complete, we have run a small pilot project to test and refine the methodology. This involved sequence data from 4249 samples with a mean coverage of 3.7X, and focused on a 40Mb region of chromosome 16 that contained the *NOD2* locus (a positive

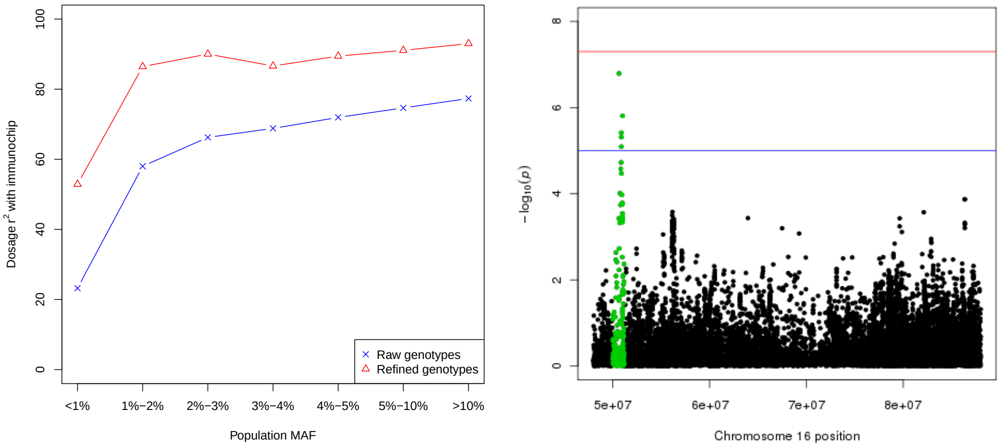

Figure 6.1: a) The correlation between allele dosage as calculated from the sequencing data and from the Immunochip data, before and after imputation genotype improvement. b) Manhattan plot of variants with MAF between 1% and 5% for the last 40Mb of chromosome 16 after extensive QC. Association testing was carried out using SNPTest on the imputation posteriors. The green dots are variants in the *NOD2* region.

control for low-frequency association). Samples were genotyped, and SNPs and indels called, using the pipeline described in Chapter 5. We then used the imputation program Beagle (Browning and Browning, 2007) to refine the genotype likelihoods, which substantially improved the accuracy of the calls when measured by concordance with Immunochip data on the same samples (Figure 6.1a). Overall, the refined genotypes had an r^2 of approximately 87% with the true genotypes at sites with a minor allele frequency of between 1% and 2%. Substantial further QC on both SNPs and samples was required to produce a clean enough dataset to allow association testing. After filtering, association tests at low-frequency variants (MAF of between 1% and 5%) yielded a clean Manhattan plot with the *NOD2* region showing clear evidence of association (Figure 6.1b). This demonstrates that the approach is sound, that specific low-frequency variants can be detected and that with enough filtering false positives rates can be controlled

The analysis of the full dataset will involve numerous methodological challenges, in addition to the significant computation burden. The probabilistic genotypes need to be well-calibrated to allow association testing, and false positive associations generated by countless new error modes will need to be identified. Standard tests, such as burden tests, will need to be redeveloped to deal with the uncertainty in the data. However, if these problems can be overcome, this experiment will allow the first well-powered whole-genome survey of low-frequency IBD risk variation to date.

6.3 Towards the ideal locus discovery experiment

I said in Chapter 1 that human disease genetics is a field where future experiments are anticipated relatively far in advance. The genome-wide linkage studies of the late 1990s and early 2000s were anticipated since at least the development of RFLP linkage in the 1970s. Likewise, the power of GWAS was foreseen since at least the 1990s. In both cases the conceptual framework was present, waiting for the technology and the sample collections to make them a reality. In the same way, we can now anticipate what the next (and possibly final) locus discovery experiment may look like in the future.

Much as the original set of GWAS experiments were followed by meta-analyses and international replication experiments, it seems reasonable that in the next few years the various international disease genetics consortia will combine their sequencing data into meta-analyses. These are likely to be very heterogeneous analyses, combining information from targeted, exome and whole-genome sequencing across a range of technologies. Doubtless this will then be followed by replication in tens of thousands of samples. The power of these projects will depend on the coverage of their component studies, but it is likely that a large number of low-frequency and rare associations will be identified at this point.

Beyond this, we start to move towards what disease geneticists refer to as the "right" or "ideal" locus discovery experiment. The cost of sequencing has fallen dramatically in recent years, and the speed and ease of sample preparation seems set to rise dramatically. We are on the verge of the $1000 genome, and it seems likely that the next decade will bring the cost of whole-genome sequencing down to $100 a sample or below. Within 20 years it is likely that a WGS experiment including hundreds of thousands of samples would have a price tag measured in the low millions of pounds, and be as technically feasible as GWAS.

Even in the absence of a concerted effort from researchers, it seems likely that such datasets will become available eventually. Cheap and readily available genome sequencing is already being used in clinical genetics practice to diagnose genetic disease (Worthey et al., 2011b; Rios et al., 2010), to guide cancer treatment (Link et al., 2011), and as a cost-effective form of carrier testing (Bell et al., 2011). It is likely that a relatively large proportion of patients will undergo routine whole-genome sequencing in the not-too-distance future, and many of these patients will consent to their data being used for research. The cost of the "ideal" WGS experiment may well end up being covered by the budgets of public healthcare services and private insurance companies.

Let us imagine that, sooner or later, researchers will have access to high-quality genome sequencing from 100,000 IBD cases (around a third of the patients in the UK), including sporadic and familial cases, as well as sequence data from their parents and an arbitrarily large number of other healthy controls. What could this ideal dataset tell us about the genetics of IBD?

Firstly, as discussed in Chapter 4, this dataset would have a very high power to fine-map associations with odds ratios larger than 1.1. If any IBD loci exist that have not yet been fine-mapped by other projects, a dataset of this size and completeness would allow the vast majority to have the causal variant determined.

Secondly, this dataset would allow us to characterise a large proportion of the common, low-effect size variants that contribute to polygenic risk, detecting most common risk variants with odds ratios > 1.03. Distinguishing these variants from the effects of very subtle population stratification may be difficult, but sequence data is also available on the parents this can be easily overcome. These polygenic risk loci are likely to cover a significant proportion of the genome, and be extremely difficult to fine-map, making them hard to interpret. However, they could be combined with other external datasets to perform detailed network and pathway analyses, in the same manner as the Immunochip loci were used in Chapter 4.

Thirdly, almost all low-frequency risk variants (MAF $> 1\%$) with an odds ratio of greater than 1.15, and all the rare (MAF $> 0.01\%$) mutations with odds ratios greater than 3, could be identified via this dataset. Aggregation tests for rare variants (Neale et al., 2011) would also allow us to identify genes or other functional units that carry extremely low-frequency risk mutations. This high power and completeness of data would allow us to ask questions about the biological properties that lead to some genes, parts of genes or classes of variation to carry risk variants, while others do not (using the techniques described for loss-of-function variants in Chapter 3 and *NOD2* coding variants in Chapter 4).

Fourthly, the family data would allow us to identify high penetrance familial mutations in IBD. The sequencing of parents would allow us to detect the contribution of de novo point, indel and structural mutations to IBD (in a similar fashion to recent studies of autism (Neale et al., 2012)). It would also let us identify extremely rare near-Mendelian mutations that are shared only by a handful of families, using the techniques described in Chapter 5.

Finally, this dataset would allow the full power of genetic risk prediction to be utilised, via standard risk prediction using established variants, and via identity-by-state and identity-by-descent methods. Assuming 50% of liability variance is captured by the risk score, we would be able to define a "high-risk" group who are more likely to than not to develop the disease (which would catch 3% of cases), and a "medium-risk" group that have a 1 in 6 chance of developing the disease (catching another 28% of cases). Even in these extreme cases genetic risk prediction would not give any guarantees, and some 10% of cases will still have a lower genetic risk than the population mean. However, the information provided may still be a significant aid to diagnosis and prevention, particularly in combination with non-genetic risk prediction.

While we are dreaming up ideal datasets, we can also imagine what other functional information that may come attached to our ideal genetic dataset. Today many IBD patients undergo measurement of certain cytokine and antibody concentrations, and the battery of tests is always increasing. It has recently been shown that gene expression data from CD8+ T-cells can be used to predict disease prognosis in IBD (Lee et al., 2011), and it stands to reason that such assays will become standard procedure in the future. Sequencing assays of epigenetic data, such as of methylation, transcription factor binding or open chromatin, are also rapidly becoming important in research, and may eventually become clinical tools. It seems likely that our ideal dataset will be accompanied by at least some data on gene expression, epigenetic marks and other relevant biological quantities. Combining the WGS data with this functional data will both allow us to reconstruct how the genetic risk factors act biologically.

6.4 Beyond locus discovery

This thesis has largely been focused on methods for discovering disease-associated loci. In this chapter I have discussed even more ways that we can discover risk loci, both in the near and more distance future. However, as I discussed in Chapter 1, risk loci in and of themselves are not inherently valuable to science or society. Despite all the debate about missing heritability, very few scientists have a deep and abiding desire to increase the "heritability explained" counter up to 100%. It is only when these loci can improve our understanding of disease biology, or directly impact patient care, that they really start earning the investment put into finding them. It is in the follow-up of these disease-associated loci that the real biological discovery starts to take shape.

In Chapter 2 I described the field of complex disease genetics as inherently statistical, and I stand by that statement. However, after disease loci have been identified, the task of following them up has historically been passed on to our less statistically, and more experimentally minded colleagues. For instance, the discovery of the *NOD2* loci via linkage was followed up by a decade of experimental work, establishing and investigating the biological links between *NOD2*, IBD and immunity (Shaw et al., 2011).

However, this is not a sustainable approach. The GWAS era ended the days when disease loci could be counted on one hand (and numbered in a universally recognised fashion: "IBD5", "IDDM2", "BRCA1"). Now risk loci are numbered in the hundreds, encompassing thousands of genes. Understanding the function of these loci has moved from something that can be established in the lab, and into the domain of statistical genetics described in Chapter 2. This mirrors developments in other fields, such as the rise of gene expression profiling and functional sequencing assays that have made functional biology into a high-throughput science (The ENCODE Project Consortium, 2012).

Many of the more striking discoveries that I have described in this thesis have come from the integration of genetic and functional data. As I hinted at in the previous section, future genetic studies are likely to become increasingly tied in with functional assays, allowing GWAS-style studies at all levels of disease biology. This will also be of benefit to purely experimental scientists following up these experiments, as the unit of follow-up will change from a gene name to a more detailed biological mechanism, pathway or hypothesis.

The next great challenge of statistical genetics in the coming decade will be to take the techniques and philosophy that have driven locus discovery, and turn them to the task of understanding the biological mechanisms of disease risk. This will require new models and new methodology, but perhaps more importantly it will require statistical geneticists to engage with disease biology, and experimental biologists to engage with statistical models. As I have seen throughout my thesis, complex disease genetics is an inherently statistical field, but it is also an inherently biological one.