CHAPTER 5

DISCUSSION

In this thesis I explored the functional impact of copy number variation using both a disease association approach and a prediction-based approach with a focus on heterozygous LOF CNVs. Initially, I developed an informatics pipeline for robust discovery of CNVs from large numbers of samples genotyped using the Affymetrix whole-genome SNP array 6.0, to support both the association-based and prediction-based study. For the disease association strategy, I studied the role of both common and rare CNVs in severe early-onset obesity using a case-control design, from which a rare 220kb heterozygous deletion at 16p11.2 that encompasses *SH2B1* was found causal for the phenotype and an 8kb common deletion upstream of *NEGR1* was found to be significantly associated with the disease, particularly in females. Using the prediction-based approach, I characterized the properties of haploinsufficient (HI) genes by comparing with genes observed to be deleted in apparently healthy individuals and I developed a prediction model to distinguish HI and haplosufficient (HS) genes using the most informative properties identified from these comparisons. An HI-based pathogenicity score was devised to distinguish pathogenic genic CNVs from benign genic CNVs. Finally, I proposed a probabilistic diagnostic framework to incorporate population variation, and integrate other sources of evidence, to enable an improved, and quantitative, identification of causal variants. As a demonstration, I applied the framework to CNVs that are rare and of unknown inheritance, and CNVs that occur *de novo*.

CNV discovery is fundamental to all CNV-related analysis. It is worth considering the limitations of the CNV discovery that underpins this thesis. With over nearly 2M probes both targeting known common CNV regions and distributed throughout the genome, Affymetrix 6.0 is arguably one of the better commercially available single array platforms for genomewide CNV detection. Birdseye is highly tailored to Affy6 data and, in my benchmarking, produced the best call set compared to other tested CNV discovery algorithms. However, CNV discovery using Birdseye is not perfect and is affected by various technical issues like sample quality and batch size (see Chapter 2). Sensitive and robust CNV discovery is limited to larger CNVs, which may have some impact on the subsequent analyses.

In Chapter 3, in the analysis of genomic burden of rare CNVs, the majority of the burden was concentrated in the largest variants, greater than 500kb. However, the ability to investigate CNV burden of smaller CNVs might be confounded by the calling of smaller CNVs being less robust and it is prone to biases between collections. In addition, the less robust CNV calling of smaller CNVs might have caused the greater proportion of the nominally associated rare CNVs <50kb being rejected by manual examination of intensity profile, compared to rare CNVs >50kb (data not shown). However, it may be less a problem for common CNVs, since they are well-tagged by SNPs irrespective of their size and the impact of smaller CNVs could be investigated by imputing them using reference haplotypes containing CNVs, such as those generated by the 1000 genomes project.

In Chapter 4, the collation of haplosufficient genes and the generation of the population distribution of pathogenicity score for non-causal variants also depend on CNV discovery. The impact of less robust calling of smaller CNVs on the collation of HS genes is likely minor since (*i*) the requirement of being found in at least two individuals should remove many false positives and (*ii*) considering just the larger genic CNVs provides sufficient information to assemble a sizeable training set. The impact of the limitations of CNV discovery on probabilistic diagnosis is probably minor because these limitations mainly affect the lower end of the distribution of pathogenicity scores of non-causal variants, whereas it is mainly the high end of this distribution that overlaps with the distribution for causal variants and thus could influence the resultant posterior probabilities.

In Chapter 4, I primarily considered the functional impact of deletions that are obviously LOF. The gene-based predicted probability of exhibiting HI and the pathogenicity score derived from such prediction is useful for interpreting LOF CNVs, and LOF sequence variants. I also showed that these pathogenicity scores may be useful for interpreting whole gene duplications as many HI genes are triplosensitive as well. Intragenic duplications are harder to interpret on the basis of that their interpretation requires knowledge of the precise variant structure, and array data do not contain any information on the location of the duplicated segment. These LOF-based scores are not likely to be so useful for interpreting other classes of CNV functional impact, for example, gain of function changes. Interpreting sequencebased CNVs should become more straight-forward given the greater information on the precise structure of the new allele. Interpreting the functional impact of smaller CNVs will be challenging, but can draw upon some of the finer annotations used for predicting the functional impact of point mutations, such as: base conservation, physical and chemical properties of amino acids, protein domain structure, spatial location relative to the active site. Full interpretation of individual genomes is going to require measures of functional impact, ‘pathogenicity scores’ for all classes of variation.

The predictive framework that I developed for characterizing a set of genes/variants by comparison with a contrasting type of genes/variants and training a classification model using the most informative characteristics drawn from a broad range of evolutionary, genomic and functional properties could be applied to other classes of putatively functional variants. Although current networks of interacting proteins and genes are far from complete, network centrality and network proximity to other known HI genes were among the most significantly differentiated properties between known HI and HS genes. The latter was also the most informative predictor in the selected prediction model. Incorporating network information is likely to be of considerable utility in the development of pathogenicity scores for other types of variation. As an exemplar, in a recent study of LOF variants discovered in the pilot phase of the 1000 genomes project (MacArthur *et al*, in press), I applied the same strategy to distinguishing recessive disease genes and dispensable genes (those disrupted by homozygous LOF SNPs, indels and CNVs) and the model achieved an AUC of 0.83 in cross-validation. Critical to the success of the prediction of haploinsufficiency and recessive LOF genes was the collation of a large body of population data. Exome sequences are only now becoming available for sample sizes of thousands. We can expect these population data to be invaluable for the development of improved pathogenicity scores for genic sequence variation.

Full characterization of the role of CNV in the genetics of obesity, or indeed any trait, will require an integrated analysis of the full range of genetic variation: sequence and structural variation, coding and regulatory variants. For example, pooling rare CNVs and point mutations in the same functional elements could increase statistical power to detecting associated loci. Moreover, causal recessive genes harboring LOF deletions of one allele and deleterious point mutations in the other allele could be missed if considering CNVs alone. Finally, point mutations and CNVs may also interact in *cis* or *trans* to have a functional impact that is impossible to appreciate from study of the CNV in isolation. The generation of exome sequences from many of the severe early onset obesity cases studied here, as part of the

UK10K project, promises to enable these kinds of integrated analyses.