CHAPTER 1

INTRODUCTION

Copy number variation (CNV) is a prevalent form of genetic variation wherein deletions and duplications of large (typically greater than 1kb) segments of the genome lead to variable number of copies of such segments among different individuals. The functional impact of copy number variants travels along the path of manifestation of genetic information from DNA, through intermediate molecular and cellular phenotypes, to individual organismal phenotypes, and onwards towards evolutionary change [1].

At the DNA level, CNVs can encompass part or all of one or multiple genes, or regulatory elements that act in *cis* or *trans* to coding sequences, thus leading to alteration of structure or abundance of transcripts and proteins. Lupski *et al* [2] summarized six types of molecular mechanism by which a CNV can affect functional sequences (Figure 1.1), including (*i*) dosage changes, (*ii*) disruption of coding sequence, (*iii*) gene fusion, (*iv*) position effect, in which the CNV has effects on expression/regulation of genes near the breakpoint, potentially by removing or altering a regulatory sequence, (*v*) unmasking a recessive allele or functional polymorphism, and (*vi*) transvection effect, in which the deletion of a gene and its surrounding regulatory sequences affects the communication between alleles.

Gene expression is the first step on the path of manifestation, and propagates the disruption of functional DNA sequences into molecular phenotypes. Stranger *et al* [3] have verified that an appreciable minority of the variation in transcript abundance

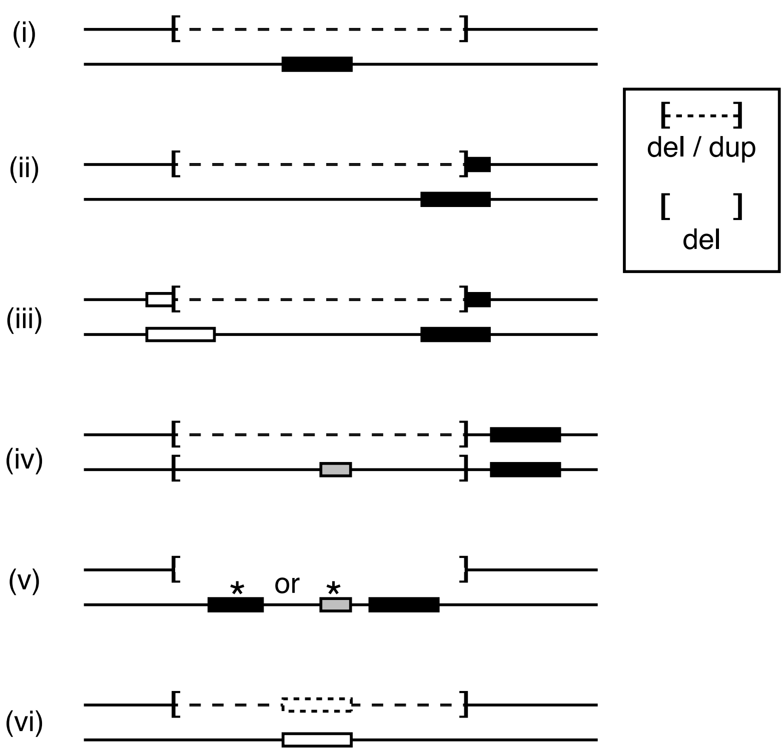

Figure 1.1: Molecular mechanisms for CNV's impact on functional sequences. Adapted from Lupski *et al* [2].

in cell-lines can be explained by CNVs, but they also demonstrated that expression at all CNV-affected loci is not equally responsive to underlying DNA dosage, and that expression can be sensitive to disruption of regulatory sequences as well as changes in dosage caused by full length deletion or duplication. It should be noted, however, that the numbers and types of tissues studied, the sensitivity of transcript profiling and the resolution and accuracy of CNV detection hinder the drawing of robust quantitative conclusions from these kinds of studies.

The impact of CNVs at the protein level is less clear, as technologies for quantitative profiling of protein abundance in parallel are less mature, although detailed characterization of protein changes caused by individual CNVs is not uncommon. For example, chromosome translocation that leads to truncation of *DISC1* has been known to cause Schizophrenia [4] and the truncated *DISC1* has been found up-regulated in patients of Schizophrenia at transcript level [5], however the truncated protein has not been detected in those patients [6], although the introduction of truncated protein in mice led to phenotypes resembling severe Schizophrenia in human [7].

Molecular phenotypes are propagated into cellular phenotypes by the perturbation of cellular networks of interacting genes and proteins. Although the current knowledge of human protein-protein or genetic interactions is far from complete and the direction, strength and consequence of such interactions is even less well understood, it is believed that some perturbations may be buffered by the network such that there is no change in outputs, others may render the network more sensitive to other genetic and environmental perturbations, others may perturb the network outputs but be buffered at higher levels of physiology and others may cause fundamental errors in organismal function. While mapping genes disrupted by CNVs in patients with a given disease onto such networks has identified enrichments of CNV-affected genes in parts of a network that relate to specific, aetiologically-relevant, pathways and complexes [8, 9], the actual network output in response to such perturbation has not been measured directly.

The impact of CNVs on function at the level of an entire organism is the primary focus of genetic disease and complex trait association studies. A large number of genetic diseases, especially neurodevelopmental disorders, have been shown to be caused by large rare CNVs (*e.g.* [9–11]). Conversely, common CNVs appear to account for a very small fraction of common disease susceptibility alleles [12, 13].

At a population level, the functional impact of CNVs is revealed by the imprint of natural selection in their genomic distribution and allele frequencies. Conrad *et al* showed that negative selection removing deleterious alleles from the population is greatest for deletions that remove exonic sequences, and is much milder on duplications and deletions of non-exonic sequences[12]. In addition, dosage-sensitive genes have been shown to be preferentially located in regions of the genome with lower rates of deletions and duplication [14]. Population studies of individual CNVs have suggested that a minority of genic CNVs might confer a selective advantage in certain environments (*e.g.* [15]), and at an evolutionary level, some copy number differences between species have been suggested to have been adaptive (*e.g.* [16, 17]).

Rather than explore all possible molecular mechanisms by which a CNV might exert a functional impact, and all levels of biology along the path to manifestation outlined above, in this thesis I focus primarily on the causal role of CNVs in disease, with a particular emphasis on CNVs that result in unambiguous loss of function of encompassed genes. Each chapter is self-contained, and so most of the relevant introductory material is presented within each chapter.

Chapter 2 describes the development of a CNV discovery and quality control (QC) pipeline for Affymetrix 6.0 genotyping array data. The chapter first assesses the performance of several existing CNV discovery algorithms on Affymetrix 6.0 data and then describes CNV call and sample QC procedures developed to produce robust CNV call sets for subsequent analyses.

Chapter 3 describes the functional impact of CNVs on the proportion of coding sequences that are most sensitive to DNA dosage alteration. The chapter first describes the computational identification of the tendency of exhibiting haploinsufficiency for human protein coding genes, which then leads to the description of a pathogenicity scoring scheme for genic CNVs. The chapter finally describes a probabilistic diagnostic framework for CNVs that can incorporate various aspects of the knowledge of the variant and harness population distribution of variant pathogenic scores conditioned on that knowledge.

Chapter 4 describes the investigation of the role of CNVs in severe early onset obesity. The chapter is organized in two parts of which the first describes the analyses of an initial and smaller patient cohort and the second describes the analyses of a following and larger patient cohort. The impacts of both rare and common CNVs were examined.