THE FUNCTIONAL IMPACT OF COPY NUMBER VARIATION IN THE HUMAN GENOME

This dissertation is submitted for
the degree of Doctor of Philosophy,
by

NI HUANG

Wellcome Trust Sanger Institute
Darwin College, University of Cambridge

December 2011

PREFACE

I hereby declare that this dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except specifically indicated in the text and acknowledgements. No part of this dissertation has been submitted for a degree or diploma or other qualification at the University of Cambridge or any other university. This dissertation does not exceed the 60,000 words excluding bibliography and appendices.

ACKNOWLEDGEMENT

This thesis would not have been possible without the constant guidance, help, encouragement and support of my wonderful supervisor, Matthew Hurles, to whom I am forever in debt. It is hard to overstate how much I have benefited from his knowledge, insight, patience and enthusiasm. Without him, I would have been lost.

I wish to thank members of my thesis committee: my secondary supervisor, Gos Micklem, Inês Barroso and Helen Firth for the helpful advices. I wish to thank Sadaf Farooqi and Elena Bochukova at Addenbrooke's Hospital, for their collaboration and the inspiring discussions. I am grateful to Richard Redon, whose enthusiastic supervision during my first rotation drew me to the study of copy number variation, and Chris Tyler-Smith and Yali Xue, whose teaching and caring made my second rotation an unforgettable experience.

Special thanks to people at the graduate program at Sanger Institute, Alex Bateman, Christina Hedberg-Delouka and Annabel Smith for the warm support and caring that helped me through my most difficult period during the writing of this thesis. Lots of thanks to all members of team 29 and 19 past and present at the Sanger Institute for all the help, encouragement and friendship.

I also thank the Wellcome Trust for the generous financial support.

Finally, I wish to thank my parents, whose love and understanding supported me throughout this four-year endeavor abroad, as always.

SUMMARY

The functional impact of copy number variation in the human genome

Ni Huang

Copy number variation (CNV) is a class of genetic variation where large segments of the genome vary in copy number among different individuals. It has become clear in the past decade that CNV affects a significant proportion of the human genome and can play an important role in human disease. With array-based copy number detection and the current generation of sequencing technologies, our ability to discover genetic variants is running far ahead of our ability to interpret their functional impact. One approach to close this gap is to explore statistical association between genetic variants and phenotypes. In contrast to the successes of genome-wide association studies for common disease using common single nucleotide polymorphism (SNP) as markers, the majority of disease CNVs discovered so far have low population frequencies and are mainly involved in rare developmental disorders. Another strategy to improve interpretation of genomic variants is to establish a predictive understanding of their functional impact. Large heterozygous deletions are of particular interest, since *i*) loss-of-function (LOF) of coding sequences encompassed by large deletions can be relatively unambiguously ascribed and *ii*) haploinsufficiency (HI), wherein only one functional copy of a gene is not sufficient to maintain normal phenotype, is a major cause of dominant diseases.

This thesis explored both approaches. 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 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.

PUBLICATIONS

Publications arising from work associated with this thesis:

1. E. G. Bochukova*, **N. Huang***, J. Keogh, E. Henning, C. Purmann, K. Blaszczyk, S. Saeed, J. Hamilton-Shield, J. Clayton-Smith, S. O'Rahilly, M. E. Hurles, and I. S. Farooqi. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, 463:666–70, 2010.
2. **N. Huang**, I. Lee, E. M. Marcotte, and M. E. Hurles. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet*, 6:e1001154, 2010.
3. N. J. Prescott, K. M. Dominy, M. Kubo, C. M. Lewis, S. A. Fisher, R. Redon, **N. Huang**, B. E. Stranger, K. Blaszczyk, B. Hudspith, G. Parkes, N. Hosono, K. Yamazaki, C. M. Onnie, A. Forbes, E. T. Dermitzakis, Y. Nakamura, J. C. Mansfield, J. Sanderson, M. E. Hurles, R. G. Roberts, and C. G. Mathew. Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. *Hum Mol Genet*, 19:1828–39, 2010.
4. S. Nik-Zainal, R. Strick, M. Storer, **N. Huang**, R. Rad, L. Willatt, T. Fitzgerald, V. Martin, R. Sandford, N. P. Carter, A. R. Janecke, S. P. Renner, P. G. Oppelt, P. Oppelt, C. Schulze, S. Brucker, M. Hurles, M. W. Beckmann, P. L. Strissel, and C. Shaw-Smith. High incidence of recurrent copy number variants in patients with isolated and syndromic Müllerian aplasia. *J Med Genet*, 48:197–204, 2011.
5. D. G. MacArthur, S. Balasubramanian, A. Frankish, **N. Huang**, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M.-M. Suner, T. Hunt, I. H. A. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, 1000 Genome Project Consortium, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, and C. Tyler-Smith. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335:823–8, 2012.

*Join first authors

TABLE OF CONTENTS

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

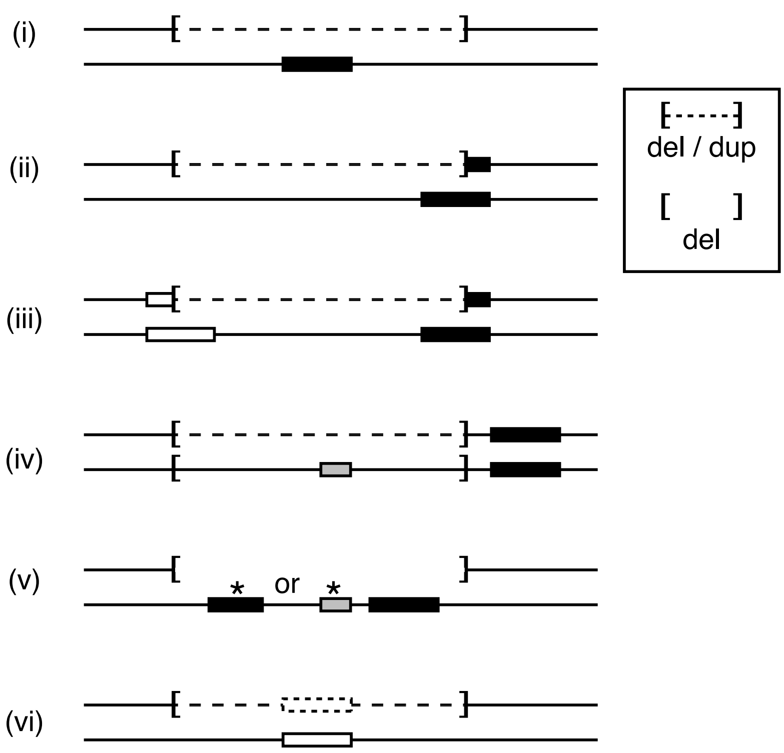

Figure 1.1: Molecular mechanisms for CNV's impact on functional sequences. Adapted from Lupski *et al* [2].

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.

CHAPTER 2

DEVELOPMENT OF A CNV DISCOVERY PIPELINE FOR AFFYMETRIX 6.0

2.1 Introduction

2.1.1 CNV discovery using microarrays

There are two major types of data that serve as the source of CNV discovery using microarrays: two-channel array-Comparative Genomic Hybridization (CGH) and genotyping arrays. The difference in the nature of the array affects data normalization, the models underpinning CNV discovery algorithms and interpretation of results.

Array-CGH hybridizes two differentially labeled DNA samples, often one test sample and one reference sample, together on the same array and the difference in DNA dosage between the two samples is reflected by the difference in fluorescent intensity between the two channel. A log ratio of the intensity is often calculated for each probe and its significant deviation from zero is an indication of copy number differences between the test and reference samples in regions targeted by the corresponding probes in the test sample relative to the reference. For this type of data, technical variation among different probes is internally controlled. Algorithms essentially find outliers in ratio space. However, the derived copy number difference is always relative and the choice of the reference affects the translation of relative copy number difference into absolute copy number.

Genotyping arrays are primarily composed of pairs of oligonucleotide probes that target the same locus but different alleles of each selected SNPs. Only one DNA sample is hybridized to the array and DNA dosage is reflected by the intensity of the probes and also partially/indirectly by the intensity ratio between the two probes targeting the same SNP. Some genotyping arrays also contain non-variable probes similar to array-CGH probes, which only provide intensity information that reflects absolute DNA dosage due to the single-channel design. For this type of data, technical variation among different probes needs to be removed explicitly in data analysis. Algorithms work in intensity space and absolute copy number can in principle be determined once probe dosage response has been calibrated.

2.1.1.1 Affymetrix Genome-wide human SNP array 6.0

Affymetrix genome-wide human SNP array 6.0 (Affy6) is an array platform that aims to perform both high-density SNP genotyping and high resolution CNV discovery simultaneously. It was developed between Affymetrix and the Broad Institute [18]. The array has 906,600 SNP probe sets and 946,000 copy number probe sets. The latter includes 202,000 probe sets targeting 5,677 CNV regions from the Database of Genomic Variants at high density and the rest spread evenly along the genome [19]. Each SNP probe set contains multiple oligonucleotide features that are identical copies of one of the two probes targeting the two possible alleles. Each copy number probe set contains multiple identical features targeting the same genomic location 1. For simplicity, I will use 'probe' to refer to 'probe set' when describing analyses that use only summarized probe set intensities or their derivatives.

After hybridization, washing and scanning, a .CEL file is produced for each sample genotyped with Affy6, which contains information including probe locations and intensities. Affymetrix developed a suite of command line tools called Affymetrix Power Tools (APT) [20] for extracting information from the .CEL files and common downstream analysis such as SNP calling and CNV discovery. A number of CNV calling methods can also be applied to Affy6 data once the probe intensities have been extracted by APT.

2.1.2 CNV discovery algorithms

Regardless of the type of the array, the typical data summary that is input into CNV discovery algorithms is often a sequence of values (intensity or log ratio) with ordered spatial coordinates along a chromosome. For genotyping arrays, a second sequence of values (measuring the relative intensity of the two alleles, often called 'B allele frequency') sharing the same spatial coordinates with the first sequence of values is available. CNV discovery aims to solve the problem of finding spatial segments with values sufficiently different from adjacent segments as a result of belonging to one of a finite set of copy number states that is different between adjacent segments.

Many CNV discovery methods have been developed. Except for a few methods that use empirical cut-off values [21, 22] or hierarchical clustering [23], most of them can be placed into one of the following two broad categories: segmentation-based (change-point-finding) methods and hidden-markov-model-based (HMM-based) methods.

2.1.2.1 Segmentation-based methods

This category of methods search for change points in an ordered sequence of values that define segments having different distribution of values (often measured by having different means). Circular binary segmentation is a typical method belonging to this category proposed by Olshen *et al* [24] that recursively test if a new segment or breakpoint should be introduced inside an existing segment based on the differences in the distribution of values between the newly introduced segment and the rest of the existing segment or between the two resulting segments separated by the proposed breakpoint. Jong *et al* [25] proposed a method that models the values along a chromosome as a sequence of normal distributions with different parameters and used the genetic algorithm to find the spatial boundaries that maximize the likelihood that actual values are drawn from those normal distributions. Similarly, Hupe *et al* [26] modeled segments of different copy number states along a chromosome as a piecewise constant function and estimated the parameters of the function using adaptive weights smoothing. Pique-Regi *et al* [27] further formulated piecewise constant function as linear combinations of step functions and used sparse Bayesian learning (SBL) to obtain the best linear combination that fits the data. All segmentation-based methods have certain measures to restrict over-segmentation during maximization of model fitting. This usually involves substituting log likelihood with Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC) or similar criteria as the target function for optimization to penalize the use of more parameters [25–27] and/or a separate pruning step after segmentation is done to eliminate spurious breakpoints [27].

2.1.2.1.1 GADA GADA is the implementation for the method described in [27]. It has a SBL step that provides initial breakpoints of which level of sparseness is controlled by the parameter a (the larger the more sparse) and a backward elimination step that removes spurious breakpoints of which stringency is controlled by the parameter T (the larger the more stringent).

2.1.2.2 HMM-based methods

HMM-based methods model the ordered sequence of values as a sequence of observed states that are determined by a chain of discrete hidden states, each one of which is determined probabilistically by its previous hidden state(s). The key parameters of a HMM include the number of hidden states K, the vector of initial state probability π, the state transition probability matrix A and collection of emission probability functions B. Fridlyand *et al* [28] applied unsupervised HMM to array-CGH data. B was assumed to be a collection of Gaussian distributions, each corresponding to a hidden state, and the initial parameters of B were estimated through clustering. Parameters (π, A, B) were optimized using the EM algorithm. The number of states K was chosen to minimize an AIC-like criteria to balance between model fitting and restricting the total number parameters. Finally, The states were merged into segments with user-defined criteria. Marioni *et al* [29] improved the model by using distance-aware transition probabilities to account for heterogeneity in probe density. Guha *et al* [30] modified the model to use a fixed 4-states HMM and incorporated Bayesian learning in which each state represented a predefined copy number state, informative priors were imposed on model parameters and MCMC was used in learning model parameters and generating copy number states. In this way, segmentation and classification was performed simultaneously. Shah *et al* [31] modeled the emission probability distribution of each state as a mixture of two Gaussian distributions with one component representing values generated from the given state and the other representing outliers, which improved the robustness of CNV calling. Methods designed for SNP genotyping arrays further exploit the additional B allele frequency (BAF) information. QuantiSNP is an objective Bayes HMM-based algorithm highly tailored to Illumina Beadarray data [32]. Similar to Marioni *et al* and Shah *et al*, state transition probabilities were adjusted for local probe distance and outliers were considered in modeling emission probabilities. Emission probabilities for BAF were modeled alongside log R ratio. Parameters were estimated using the EM algorithm with hyper-parameters of the conjugate priors for the emission model estimated from a reference dataset with known copy number. The program calculates a Bayes factor for each CNV called that facilitates ranking and post-filtering of CNV calls. PennCNV is another widely used HMM-based program for the CNV analysis of Illumina Beadarray data [33]. Its underlying model is very similar to QuantiSNP, except it incorporates population B allele frequency in the emission model for BAF and it has an *a posteriori* validation step using family information if available.

2.1.2.2.1 Birdsuite Birdsuite is a software suite highly tailored to the Affy6 data that integrates SNP and CNV calling. Its CNV discovery component, Birdseye, is an HMM-based program. Unlike most HMM-based methods, Birdseye receives predefined parameters for emission probability distributions from Canary and Birdseed, the components of Birdsuite that run prior to Birdseye that estimate copy number at known CNVs and genotype SNPs respectively. Those parameters are probe-specific and are estimated using the EM algorithm during the running of Canary and Birdseed with priors learned from samples of known genotype. The state transition probabilities are also pre-defined, distance-dependent and tuned to the probe density of Affy6. Birdseye uses the Viterbi algorithm to determine the most probable chain of states and produces a LOD score for each segment representing the strength of evidence [18].

2.1.3 CNV calling pipeline

CNV calling algorithms solve the specific problem of identifying genomic segments with likely aberrant copy number from input sequences of values with ordered spatial coordinates. However, the process that takes raw data generated by microarray experiments and produces CNV calls ready for downstream analysis involves many other steps that can affect the quality of the final set of CNV calls remarkably. The structure of a typical CNV calling pipeline is demonstrated in Figure 2.1. After raw intensities have been extracted, a pre-processing step is usually mandatory prior to CNV calling. In this step, various normalization procedures may be applied to remove technical biases or variation between channels of an array, across probes of different spatial location or genomic context, or across different array experiments, etc. Normalized intensities may be organized and transformed to the format required by the calling algorithms. Technical failures may also be identified and removed at this stage. After intensities have been properly normalized and transformed, multiple calling algorithms may be applied to complement or support one another. Next, resultant CNV calls are subjected to post-processing that usually involves computing quality control (QC) metrics and filtering calls and samples based on those QC metrics. Additional procedures such as merging CNV calls may be necessary depending on the calling algorithm that has been applied. Various visualization tools are often an essential part of the pipeline that facilitates quality control and the selection of filtering thresholds. It is crucial to assess the performance of such pipeline with an independent CNV dataset, ideally of higher quality and from the same samples, based on which the pipeline may be optimized.

The extent of completeness in implementing the above pipeline varies among current CNV calling programs. Some have pre-processing capabilities, such as the APT utility apt-copynumber-workflow, which handles intensity extraction, normaliza-

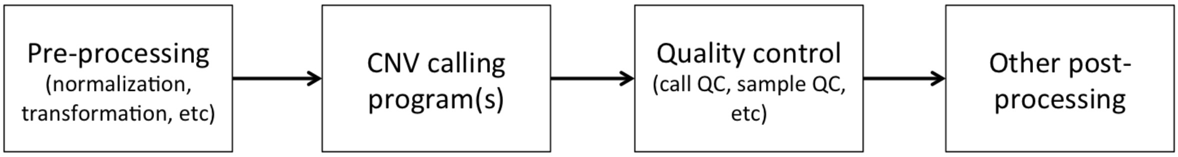

Figure 2.1: Simplified diagram of a typical CNV calling pipeline

tion, probe set summarization and CNV calling. Some are dedicated CNV callers that do not implement any pre- or post-processing at all, such as GADA, which simply outputs segmentation on receiving an input sequence of log ratios. The usability of output CNV calls also varies. Again taking apt-copynumber-workflow as an example, instead of delivering genomic segments with copy number, it only outputs the inferred copy number state of every probe set. GADA provides genomic segments but does not assign copy number state. Birdseye provides the most usable calls of the three, as it not only produces genomic segments with copy number state, but also produces the statistical confidence of the called CNVs that facilitates post-processing. Regardless of the above differences, current CNV programs provide little post-processing and QC functions, whereas robust and consistent post-processing is of vital importance in producing a reliable CNV call set, especially in studies with a large sample size and/or multiple datasets. A few simple post-processing methods have been applied in previous CNV studies [33, 34], which were limited to filtering CNV calls by number of probes and size or removing samples with large variance in probe intensities or apparent mosaic chromosomes.

In the result section of this chapter, I will first compare the performance of three CNV calling programs on Affy6 data. I will next describe a CNV pipeline I developed for Affy6 data that features an effective sample QC. Finally, I will demonstrate the application of this pipeline to several Affy6 datasets that will be further discussed in later chapters. The implementation details are provided in the methods section.

2.2 Materials and methods

2.2.1 Extracting probe intensities and re-producing the scanned image

Raw probe intensities and probe IDs were extracted from .CEL files using the APT command 'apt-cel-extract'. The positions of a probe on the array can be derived from its probe ID using the equations provided by Affymetrix [20]:

$$x = (probeID - 1) \bmod N_{\text{column}}$$

$$y = \text{floor}\left((probeID - 1) / N_{\text{column}}\right)$$

where, N_{column} stands for the number of columns of the probe array, which is 2680 for Affy6 [20]. I wrote an R script to calculate the positions, re-order the probes according to their positions and plot the scanned image using heat map colors. The brighter the color, the higher the intensities.

2.2.2 Extracting and normalizing probe set intensities

I extracted probe set intensities from .CEL files and normalized them across samples on the same sample plate using the APT command 'apt-probeset-summarize' with the option 'quant-norm.target=1000,pm-only,plier.optmethod=1,expr.genotype=true'. This command first extracted probe intensities from all input sample .CEL files, then applied quantile normalization to adjust all samples to the same distribution with a median probe intensity value of 1000 and lastly summarize probe set intensities from composing probes using the PLIER (probe logarithmic intensity error) estimation with the 'perfect match only' option [20].

2.2.3 Transform probe set intensities into log ratios

Let $x_{i,j}$ denotes the summarized and normalized intensity of probe set i in sample j. The log ratio $y_{i,j}$ was calculated as:

$$y_{i,j} = \log_2 \frac{x_{i,j}}{\text{median}\left(x_{i,*}\right)}$$

, where median $\left(x_{i,*}\right)$ is the median value of all samples on the same plate.

2.2.4 Calculating log-ratio-related sample QC statistics

Noise level and extent of spatial waviness (autocorrelation) of array data are two important factors that remarkably affect CNV analysis as will be described later. I used median absolute deviation (MAD) of probe sets log ratios as the measure of noise level and sum of auto-correlation (SAC) along the chromosomes as the measure of spatial waviness. For sample j:

$$MAD_j = \text{median}\left(\left|y_{*,j} - \text{median}(y_{*,j})\right|\right)$$

$$SAC_j = \sum_{k=1}^{n=5} \left| \frac{\mathrm{E}\left[(Y_{i,j} - \mu_{y_j})(Y_{i+k,j} - \mu_{y_j})\right]}{\sigma_{y_j}^2} \right|$$

2.2.5 Correction for spatial auto-correlation

For each sample, correction was done by chromosomes using the method developed by [35]. Briefly, a loess curve was fitted to the $\log_2$ ratios along a chromosome with a window size containing 10% of the probes in the chromosome and the $\log_2$ ratios were replaced by the residuals (Figure 2.2).

2.2.6 Storage and retrieval of normalized intensity data

Normalized probe set intensities were stored as a probe-set-by-sample table with probe set name and chromosomal location in HDF format [36]. Each table contained

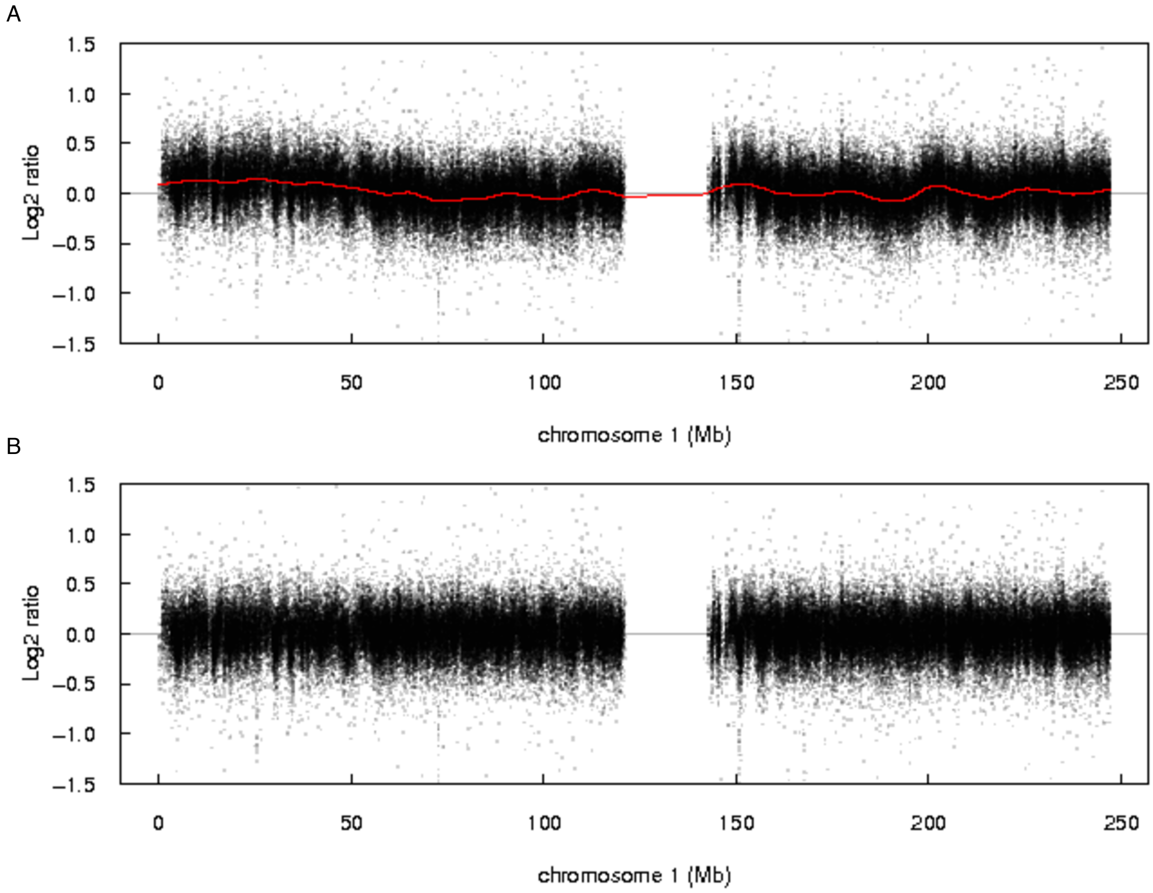

Figure 2.2: Demonstration of correcting spatial auto-correlation, showing $i\log_2$ ratio profile across chromosome 1 of a sample with a SAC of 0.7766 (top 0.15%) before (A) and after (B) correction for spatial auto-correlation. The red curve in (A) is the loess curved fitted with a window size of 14k probes (10% of all probes targeting chromosome 1)

a plate of samples. I developed a python utility using the PyTables package [37] to write such table in HDF format and create an index in the column containing chromosomal locations. The resulting HDF file was similar to an in-file database. I also wrote a python utility for retrieving probe set intensities by chromosomal coordinates from such files.

2.2.7 The CNV call format

This is the format of plain text files in which CNV calls were recorded. Each CNV is described in one row, of which fields are separated by tab and the first seven fields are required. The required fields are chromosome, start coordinate, end coordinate, number of probes contained, average log ratio, sample ID and copy number change. Additional fields may be appended to the end of row.

2.2.8 Merging split CNV calls

CNV calls on the same chromosome were sorted by genomic coordinates and scanned through, each time taking a pair of adjacent calls. The two adjacent calls were merged into one, if:

1. Both calls have the same genotype
2. The number of probes separating the two calls < 100
3. The ratio of the number of probes separating the two calls to the number of probes in the merged call $< 10\%$
4. The probe density between the two calls > 1 probe per 5kb
5. The absolute difference in average $\log_2$ratio between the two calls < 0.15

The scan and merge process was repeated until no CNV calls could be merged.

2.2.9 CNVE clustering

To combine CNVs called in different individuals into CNV events (CNVEs), I used a hierarchical-clustering-like method described in [12]. Briefly, pairwise reciprocal overlap (RO) were first calculated among CNVs overlapping at least 1bp and CNV pairs with greatest RO were merged into a CNVE if RO>50%. Then, unmerged CNVs having a RO>50% with all CNVs already merged into this CNVE were iteratively merged in order of best RO. This method guarantees that the ROs between all pairs of CNVs belonging to a CNVE are greater than 50% and when a CNV has RO>50% with CNVs of multiple CNVEs, it is merged to the one with better RO. The boundaries of a CNVE were defined as those enclosing the minimum genomic interval that encompasses 90% of belonged CNVs.

2.2.10 Definition for different overlap criteria

Given two genomic intervals A and B on the same chromosome defined by coordinates $[start_A, end_A]$ and $[start_B, end_B]$, the length of overlap $L = \min(end_A, end_B) - \max(start_A, start_B) + 1$. Simple overlap is defined as $L > 0$. Overlap relative to interval A is: $O_A = L/(end_A - start_A + 1)$. Overlap interval $A >$ 50% is defined as $O_A >$ 0.5. Reciprocal overlap $>$ 50% is defined as: $O_A >$ 0.5 and $O_B >$ 0.5.

2.2.11 Heuristic quality score for APT and GADA CNV calls

The quality score Q is defined as $N_{probe} \times \left|\overline{logRatio}\right|$.

2.3 Results

2.3.1 Comparing discovery programs for Affy6 data

2.3.1.1 Test pipeline for assessing CNV calling programs

To test the performance of apt-copynumber-workflow, GADA and Birdsuite, three test pipelines were constructed.

As stated in section 2.1.3, apt-copynumber-workflow is a fairly standalone program that handles both pre-processing (intensity extraction, normalization, probe set summarization, transformation into log ratio) and CNV discovery. However, it only produces copy number state for individual probe sets. Therefore, an extra step was added to merge adjacent probe sets into one CNV call if they had the same copy number state and their copy number were not equal to 2 and to calculate other information such as average log ratio that were required by the CNV call format (as described in section 2.2.7). Then CNVs were filtered by size, number of probes, and probe density and samples with excessive CNV calls were removed.

Since GADA is a dedicated CNV caller, apt-probeset-summarize was used to handle intensity extraction, normalization and probe set summarization. Probe set intensities were then transformed into log ratio as described in method. Unlike apt-copynumber-workflow, wherein intervention is not possible between pre-processing and CNV calling, an extra step that corrects spatial auto-correlation was added before running GADA, as it reduces the long range waviness in the data (Figure 2.2). Since GADA only performs segmentation but not copy number assignment, thresholds were applied to the distribution of average log ratio of segments to distinguish CNV calls and segments with normal copy number. The resulting CNV calls were stored in CNV call format and filtered using the same criteria as for calls made by apt-copynumber-workflow. Samples with excessive CNV calls were also removed.

Birdsuite calls apt-probeset-summarize to handle pre-processing. Since it works with probe set intensities instead of log ratios, transformation was not needed. Only output from the Birdseye algorithm were passed on for downstream analyses as Canary performs CNV typing at known CNVs rather than CNV discovery. The Birdseye calls were filtered using the LOD score in addition to the same criteria as above and samples with excessive CNV calls were removed.

2.3.1.2 Comparing general characteristics of call sets

I used the above test pipelines to call CNVs in 270 HapMap1 individuals. A number of program parameters and filter parameters were tested as listed in Table 2.1. I examined the median number of calls per sample, the median size of calls, the number of CNVEs, the fraction of singletons and overlap with published CNV datasets [12, 34] (Table 2.1).

Table 2.1: **Summaries of call sets produced by test piplines**

Program	Program or filter parameter	Median #call per sample	Median call size (kb)	Deletion-to-duplication ratio	#CNVE	%Singleton CNVE	%Overlap rate*	%Overlap rate†
APT	Default	64	14.6	2.82	2861	49.2	27.2	25.5
GADA	a=1 T=7 M=5‡	88	14.0	2.54	4514	49.3	23.8	23.2
GADA	a=1 T=8 M=5	73	15.2	2.61	3500	47.7	27.9	26.1
GADA	a=1 T=9 M=5	61	17.3	2.67	2833	47.2	30.0	27.9
GADA	a=1 T=10 M=5	51	19.4	2.80	2307	45.8	31.8	29.6
Birdseye	LOD$\geq$5	86	14.8	4.03	3469	48.5	30.9	24.7
Birdseye	LOD$\geq$10	59	23.4	4.20	2176	46.9	35.3	28.6

* Proportion of calls reciprocally overlapped by common CNVs described in McCarroll *et al* [34].

† Proportion of calls reciprocally overlapped by ng42M CNVs described in Conrad *et al* [12].

‡ For a and T see section 2.1.2.1.1; M defines the minimal required number of probes.

There were a few characteristics that were shared by or similar among all pipelines. For example, (*i*) all pipelines called 50–90 CNVs per individuals, (*ii*) the median call size of the majority of call sets were all in the range of 15–20kb, (*iii*) the proportions of singletons were close to 50%, (*iv*) one quarter to one third of the CNVEs were found previously, (*v*) all three CNV calling methods were better at calling large recurrent deletions as indicated by increased call size and deletion-to-duplication ratio and decreased proportion of singletons with increasing stringency (GADA T7 to T10, Birdseye LOD5 to LOD10) of calling. The deletion-to-duplication ratio differed remarkably between call sets produced by Birdseye and the other two programs. This is likely due to that Birdseye calls from intensities whereas the other two calls from log ratios (see section 2.4.2).

2.3.1.3 Benchmark by a high quality call set

Comparing with one or more independent high quality datasets generated from the same samples can provide direct assessment of the performance of the calling algorithms. Previously, a set of tiling resolution CNV calls were produced from 40 individuals, 19 of which were from the HapMap1 individuals, using a set of Nimblegen CGH arrays that collectively contained 42M probes [12] (referred to as ng42M). I used this dataset as a gold standard to benchmark GADA T9, GADA T10 and Birdseye call sets, as the rest of the call sets were apparently of lower quality. The fraction of ng42M CNVs reciprocally overlapped >50% by test call set in the same individual was used as a measure of sensitivity and the fraction of test call set overlapped by ng42M CNVs in the same individual was used as a measure of specificity (Table 2.2–2.9).

Breaking down by CNV size, sensitivity generally increased as call size increased in all three call sets as expected. Specificity, however, was highest in the middle ranges (10kb to 100kb) and sharply dropped to roughly 10% for CNVs above 500kb. This

Table 2.2: **Proportion of ng42M calls reciprocally overlapped by GADA T9 calls**

Frequency	(1kb,10kb]	(10kb,20kb]	(20kb,100kb]	(100kb,500kb]	(500kb,+∞]	All Classes
(0,1]	2.35%	9.38%	19.17%	19.15%	10.00%	5.63%
(1,5%]	1.67%	11.22%	7.30%	9.09%	0.00%	3.57%
(5%,10%]	1.86%	13.25%	5.08%	4.85%	50.00%	3.67%
(10%,100%]	0.78%	5.14%	5.09%	7.26%	16.67%	2.30%
All Classes	1.10%	6.77%	6.26%	7.63%	15.63%	2.83%

Table 2.3: **Proportion of GADA T9 calls reciprocally overlapped by ng42M calls**

Frequency	(1kb,10kb]	(10kb,20kb]	(20kb,100kb]	(100kb,500kb]	(500kb,+∞]	All Classes
(0,1]	44.44%	50.00%	66.67%	77.78%	50.00%	54.64%
(1,1%]	69.57%	70.59%	42.86%	58.33%	0.00%	57.60%
(1%,5%]	45.83%	52.94%	44.83%	45.65%	0.00%	45.64%
(5%,10%]	49.25%	51.43%	48.39%	40.00%	33.33%	47.13%
(10%,100%]	31.21%	55.74%	56.61%	51.79%	4.35%	44.82%
All Classes	42.86%	55.06%	52.22%	49.37%	10.26%	47.51%

Table 2.4: **Proportion of ng42M calls reciprocally overlapped by GADA T10 calls**

Frequency	(1kb,10kb]	(10kb,20kb]	(20kb,100kb]	(100kb,500kb]	(500kb,+∞]	All Classes
(0,1]	1.73%	7.59%	17.92%	19.15%	10.00%	4.82%
(1,5%]	1.11%	10.20%	6.74%	9.09%	0.00%	3.01%
(5%,10%]	1.54%	12.05%	3.73%	3.88%	0.00%	3.00%
(10%,100%]	0.63%	4.21%	4.77%	7.04%	16.67%	2.04%
All Classes	0.86%	5.70%	5.76%	7.36%	12.50%	2.46%

Table 2.5: **Proportion of GADA T10 calls reciprocally overlapped by ng42M calls**

Frequency	(1kb,10kb]	(10kb,20kb]	(20kb,100kb]	(100kb,500kb]	(500kb,+∞]	All Classes
(0,1]	59.26%	30.00%	65.52%	85.71%	50.00%	60.00%
(1,1%]	58.33%	63.16%	48.72%	63.64%	0.00%	55.66%
(1%,5%]	51.85%	51.16%	44.62%	50.00%	12.50%	48.12%
(5%,10%]	42.11%	53.57%	40.00%	40.48%	0.00%	41.72%
(10%,100%]	33.09%	56.86%	62.05%	54.35%	5.26%	48.56%
All Classes	43.92%	53.64%	55.32%	51.35%	8.33%	48.95%

Table 2.6: **Proportion of ng42M calls reciprocally overlapped by Birdseye LOD5 calls**

Frequency	(1kb,10kb]	(10kb,20kb]	(20kb,100kb]	(100kb,500kb]	(500kb,+∞]	All Classes
(0,1]	4.69%	18.30%	26.67%	14.89%	10.00%	9.19%
(1,5%]	3.90%	13.27%	9.55%	9.09%	0.00%	5.69%
(5%,10%]	2.67%	19.88%	5.42%	7.77%	0.00%	5.00%
(10%,100%]	1.31%	7.40%	7.45%	9.90%	16.67%	3.41%
All Classes	1.96%	10.27%	8.82%	9.87%	12.50%	4.27%

was likely caused by the resolution difference between Nimblegen 42M arrays and Affymetrix 6.0 arrays, which has two implications: (*i*) Nimblegen 42M arrays have much better power to detect small to middle size CNVs, (*ii*) One large Affy6 call may be called as multiple smaller CNVs in ng42M. Actually if using 'overlap ng42M >50%' as the criteria instead of 'reciprocal overlap >50%' (see section 2.2.10), specificity also increased as call size increased and reached close to 100% above 500kb (data not shown). Breaking down by CNV frequency, both sensitivity and specificity decreased as frequency increased. This might reflect the enrichment of common CNVs in duplicated regions of the genome and the impaired performance of CNV discovery algorithms in such regions, and suggests a need for different strategy for calling common CNVs. Overall, Birdseye LOD10 call sets achieved the high-

Table 2.7: **Proportion of Birdseye LOD5 calls reciprocally overlapped by ng42M calls**

Frequency	(1kb,10kb]	(10kb,20kb]	(20kb,100kb]	(100kb,500kb]	(500kb,+∞]	All Classes
(0,1]	62.00%	69.57%	51.61%	66.67%	0.00%	60.19%
(1,1%]	72.41%	61.54%	65.79%	66.67%	50.00%	67.65%
(1%,5%]	60.29%	79.63%	47.44%	39.13%	0.00%	56.96%
(5%,10%]	45.88%	69.57%	36.96%	40.63%	33.33%	45.50%
(10%,100%]	43.88%	51.32%	53.20%	48.23%	0.00%	48.91%
All Classes	52.06%	60.79%	51.63%	46.58%	18.18%	52.41%

Table 2.8: **Proportion of ng42M calls reciprocally overlapped by Birdseye LOD10 calls**

Frequency	(1kb,10kb]	(10kb,20kb]	(20kb,100kb]	(100kb,500kb]	(500kb,+∞]	All Classes
(0,1]	2.83%	12.95%	22.92%	14.89%	10.00%	6.75%
(1,5%]	1.89%	12.24%	8.43%	7.27%	0.00%	3.90%
(5%,10%]	1.46%	15.06%	5.08%	6.80%	0.00%	3.61%
(10%,100%]	0.97%	4.21%	6.29%	8.58%	11.11%	2.62%
All Classes	1.27%	6.77%	7.54%	8.62%	9.38%	3.20%

est specificity (55.59%) and the second highest sensitivity (3.20%) of the three. It was particularly better in calling smaller events (1k to 20kb). It had lower specificity than GADA call sets in middle to large size ranges, but it might be affected more severely by the array difference discussed above as having a much larger median call size.

I investigated how sensitivity and specificity changes as a function of call filters. For Birdseye calls, LOD score was a natural quality filter. As APT and GADA did not compute a per call confidence/quality score, a heuristic formula (see section 2.2.11) previously shown to be monotonically related to false positive rate [38] was used. To account for the fact that ng42M calls were relative to a certain reference individual, sensitivity and specificity was calculated based on both direct comparison of Affy6 CNV calls to ng42M CNV calls in the same individual and comparison of

Table 2.9: **Proportion of Birdseye LOD10 calls reciprocally overlapped by ng42M calls**

Frequency	(1kb,10kb]	(10kb,20kb]	(20kb,100kb]	(100kb,500kb]	(500kb,+∞]	All Classes
(0,1]	90.48%	70.59%	62.96%	66.67%	0.00%	72.46%
(1,1%]	73.33%	61.11%	75.86%	66.67%	50.00%	70.33%
(1%,5%]	67.16%	87.80%	55.74%	47.37%	0.00%	63.64%
(5%,10%]	54.00%	91.67%	44.44%	42.86%	0.00%	52.10%
(10%,100%]	49.74%	44.05%	54.15%	46.46%	0.00%	49.85%
All Classes	57.98%	62.21%	55.92%	47.60%	10.00%	55.59%

Affy6 CNV calls to ng42M CNVEs. Birdseye calling plus LOD score filtering outperformed other call sets from other algorithms filtered using the heuristic score under most stringency thresholds by yielding more calls while achieving higher specificity (Figure 2.3).

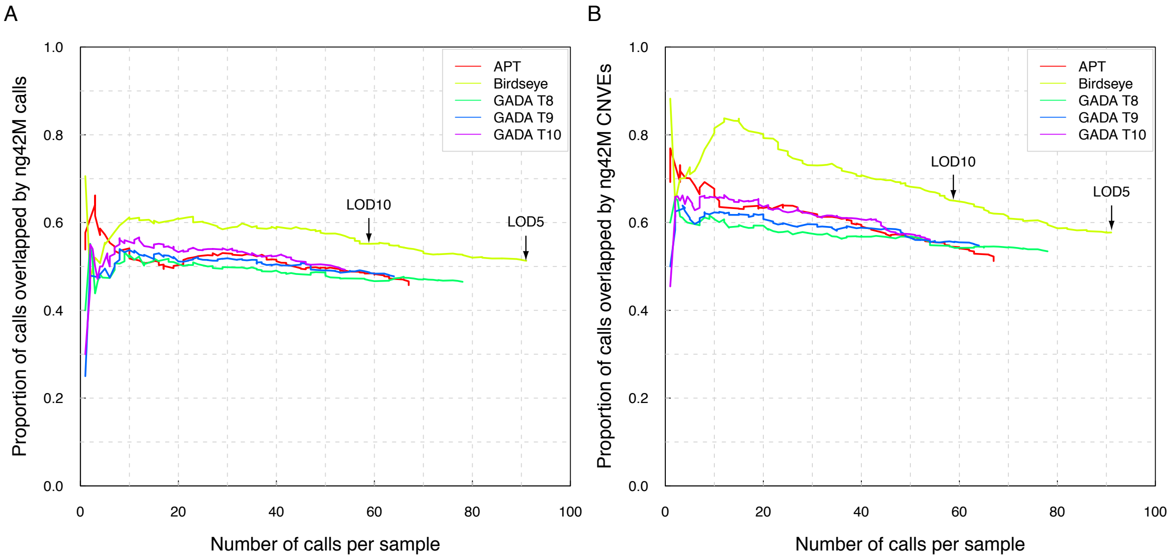

Figure 2.3: Sensitivity, measured by median of number of calls per sample, versus specificity, measured by proportion of calls reciprocally overlapped by ng42M CNVs in the same sample (A) or by ng42M CNVEs (B), under shifting call filters. LOD score (for Birdseye calls) and number of probes times absolute $\log_2$ ratio (for APT and GADA calls) thresholds increase from right to left.

Since calculation of specificity and sensitivity was based on the definition of overlap, I examined if the superior performance of Birdseye was independent of overlap threshold. By fixing the call filter and shifting the reciprocal overlap threshold used to define a test call as being present in the gold standard dataset, a series of specificity value were calculated. Again, Birdseye call sets out-performed other call sets and Birdseye LOD10 had higher specificity than Birdseye LOD5, as expected (Figure 2.4).

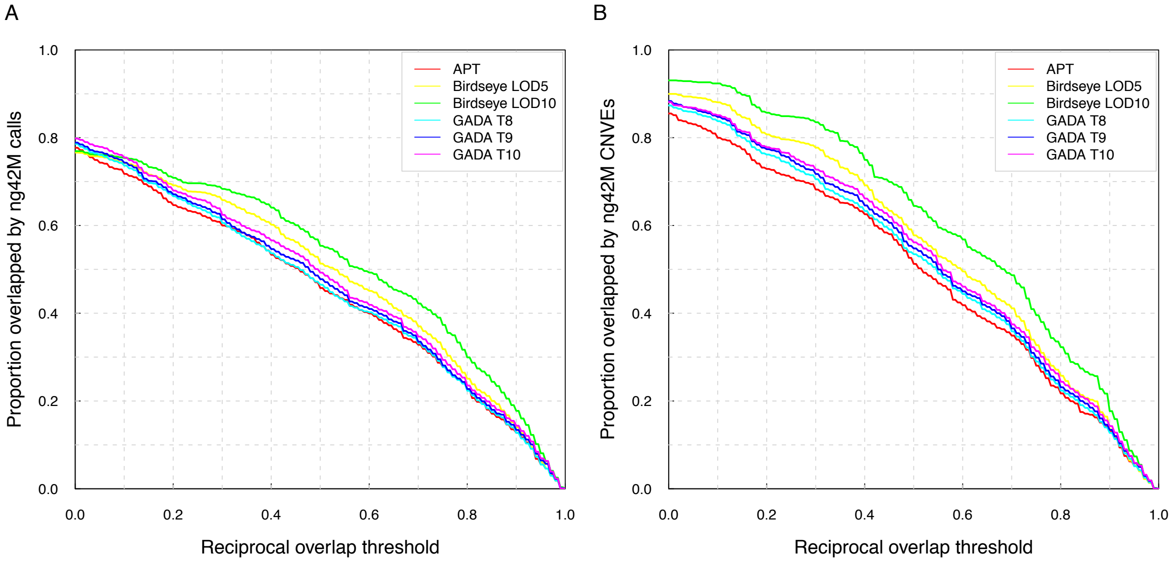

Figure 2.4: Specificity, measured by proportion of calls reciprocally overlapped by ng42M CNVs in the same sample (A) or by ng42M CNVEs (B), as a function of reciprocal overlapping threshold.

Based on the above comparisons, Birdsuite was chosen as the core CNV discovery program around which the production pipeline was built.

2.3.2 Implementing a CNV discovery and QC pipeline for Affy6 data

I developed a robust CNV calling and quality control pipeline for Affymetrix 6.0 data around Birdsuite. The pipeline is able to process thousands of samples automatically, providing robust CNV calls ready for downstream analysis and visualization for manual examination of calling quality. Below I have used the WTCCC2 control dataset as an example when demonstrating certain features of the pipeline.

2.3.2.1 Pre-calling QC

Defects in array experiment can sometimes be visually apparent when simply looking at the scanned image and could lead to the exclusion of the array before entering the CNV discovery process. As the actual scanned images are not available for many array experiments, they were regenerated from the .CEL files (see section 2.2.1). Those with defects such as contamination (Figure 2.5A) and global low-hybridization (Figure 2.5B) can be easily distinguished from those with scanned images of typically normal experiments (Figure 2.5D). Usually, small-scale contamination has little impact on the overall quality of data, but abnormally low hybridization often causes increased noise level. Samples with such defects could also be identified by other QC metric at later stages. Therefore, except for some very rare cases (Figure 2.5C), the role of scanned images generated at this step was mainly to aid the investigation of the cause of low quality data rather than dropping samples before CNV calling.

2.3.2.2 Pre-processing and CNV calling

CNVs were called by plate using Birdsuite with default parameters. The normalized and summarized probe set intensities produced by apt-probeset-summarize, which was called by Birdsuite to handle pre-processing, were stored in the form of in-file database (see section 2.2.6). A copy of the intensities was transformed to log ratios and were used to calculate average log ratio for each CNV call and log-ratio-related

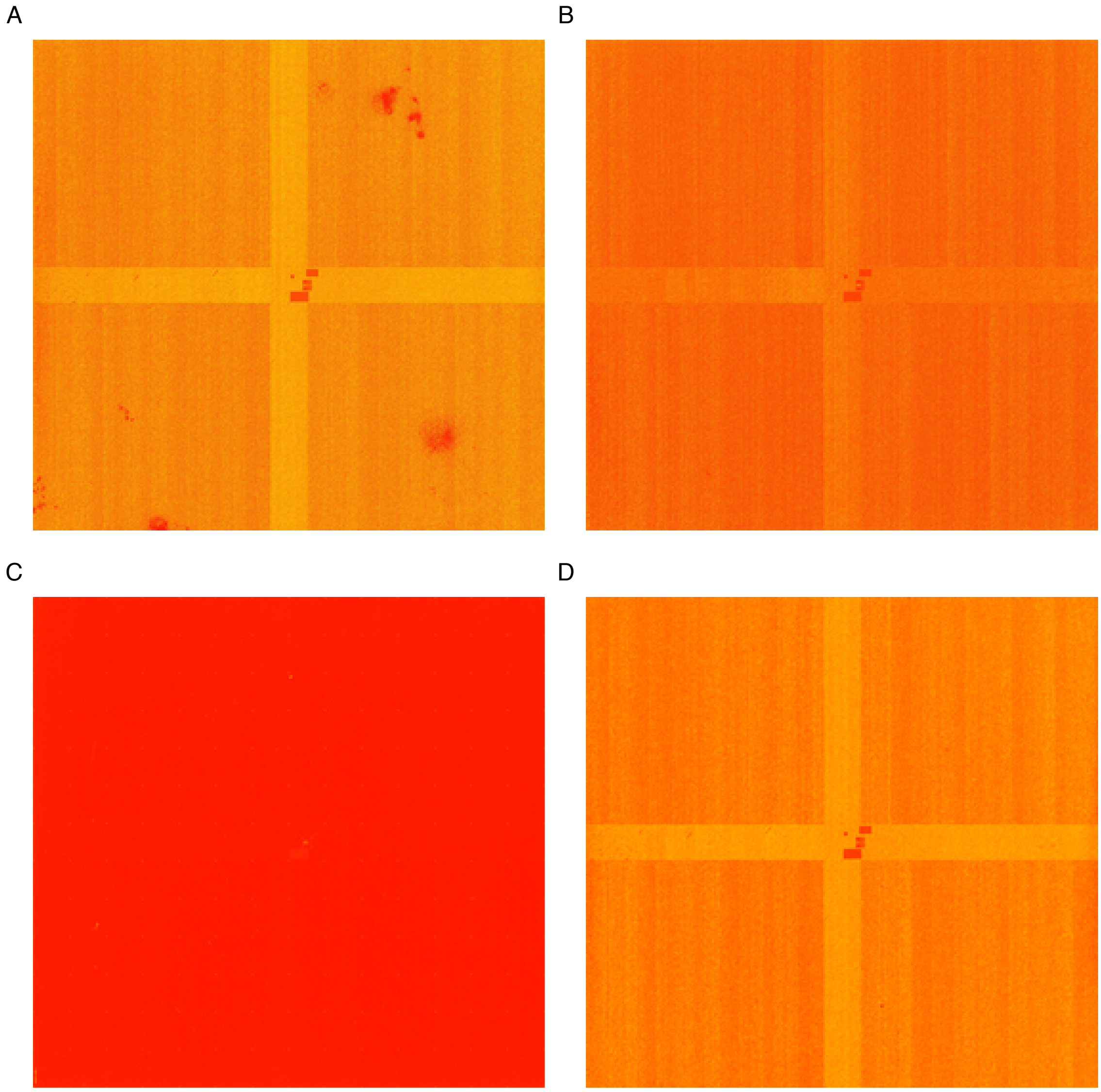

Figure 2.5: Regenerated scanned images of four samples. An Affy6 chip contains close to 7 million probes organized in a 2572 × 2680 matrix, each represented as a dot in heat colors. Brightness is proportional to $\log_2$ intensity. The four scanned images are examples of contamination (A, scattered abnormal low intensity regions in top right, bottom right and bottom left zone), global low-hybridization (B, globally darker color and blurred border between the central cross region and the rest of the chip), failed experiment (C, no hybridization signal at all) and normal scanned image (D, bright and clear cross region with the rest of chip being relatively homogeneous).

sample QC statistics for each sample. CNVs called from all plates were pooled and stored in CNV call format.

2.3.2.3 CNV call QC

A number of summary statistics were calculated and visualized to aid the decision of filter parameters (Figure 2.6). Considering that Birdsuite does not yet correctly segment Y chromosome and CNV calling in X chromosome is problematic due to the presence of pseudoautosomal regions and the difference in neutral copy number between male and female, and in order to remove the lower end tail in the distribution of call size, number of probes and probe density, I set up the following criteria to filter CNV calls:

1. Autosomal
2. LOD score ≥ 10
3. Number of probes ≥ 5
4. Size $\geq$ 1kb
5. Probe density $\geq$ 1 per 10kb

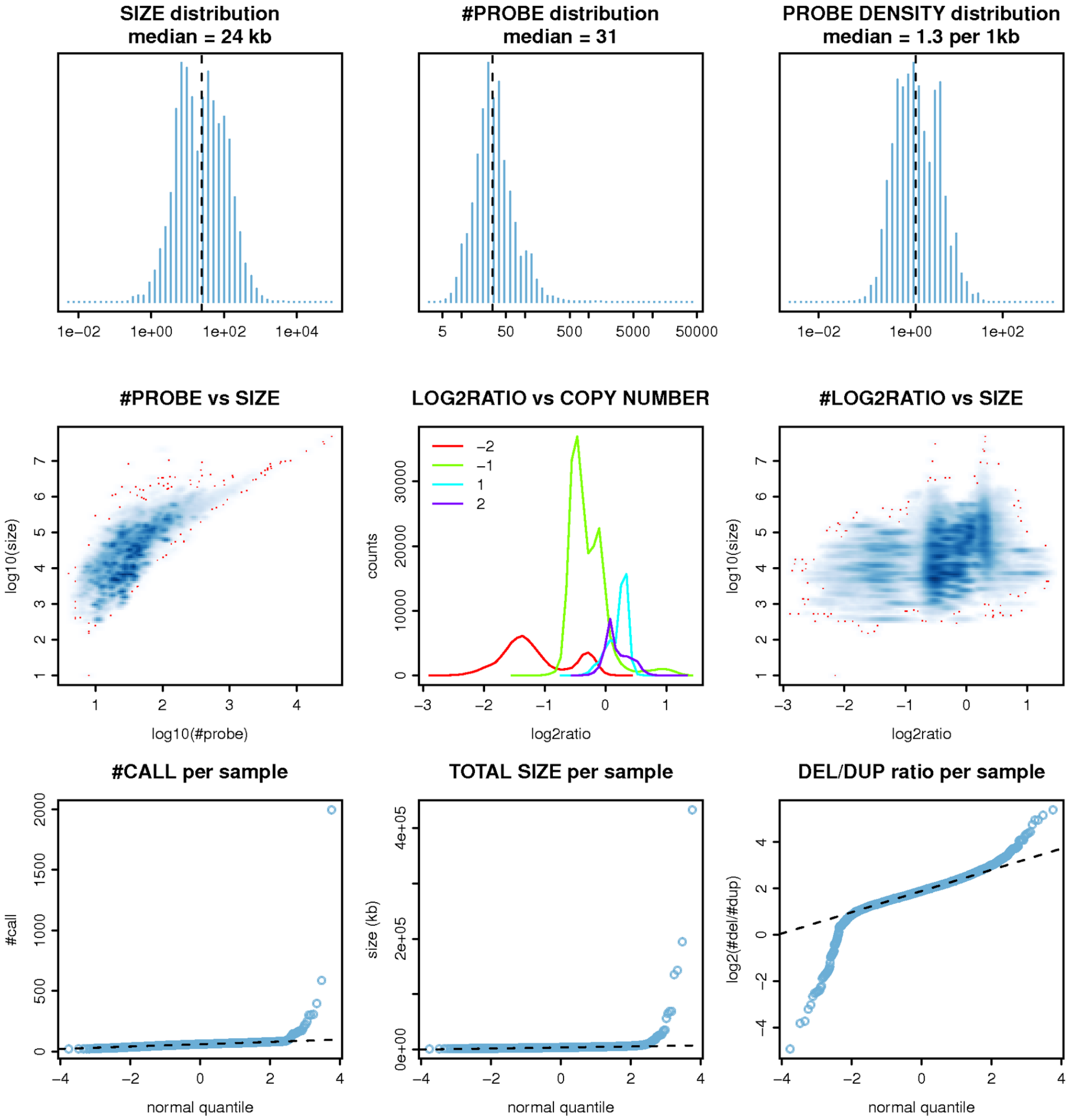

Figure 2.6: Summary statistics of the call set before call QC.

2.3.2.4 Sample QC

Even after the above filters, the number of CNVs called in some samples were a few orders of magnitude greater than most of the other samples. Obviously, this variation could not be explained solely by natural variation in the number of CNVs carried by an individual, but were more likely technical artifacts. There can be two types of such artifacts. The first type is caused by differential sensitivity, wherein specificity of CNV calling is good and similar across samples, but due to some samples being noisier than others, fewer CNVs can be called with a level of confidence that reaches a pre-defined common threshold. The second type is more severe, wherein the data quality of some samples is so poor that CNV calling program starts to produce large number of false calls. This was observed when running samples with strong waviness as indicated by having high spatial auto-correlation through the GADA test pipeline without correction. The number of apparently false CNVs was effectively reduced after correcting for spatial auto-correlation, suggesting waviness was indeed the cause of such artifact (Figure 2.7). As Birdseye works on probe set intensities rather than log ratios, such correction could not be performed, which might explain the excessive CNV calls. To account for both types of artifact, I used a linear function to model the negative correlation between the number of calls per sample and the sample's median absolute deviation (MAD) of log ratios, a measure of the level of noise in the data, wherein the parameters of the linear model were estimated using samples with a MAD<0.3 and a SAC in the bottom 90% in order to exclude the influence of samples with extreme level of noise or spatial auto-correlation. Samples which after correction were more than four MADs from the fitted linear model were removed (Figure 2.8).

In addition to the number of calls per sample, deletion-to-duplication ratio (DDR) should also be relatively stable across samples given the same CNV discovery algorithm and a reasonably large number of CNVs called per sample. Indeed, most samples had a DDR between 1 and 16 with a median at about 4 (Figure 2.9). The majority of samples that fell outside this range had a DDR below the lower bound and many of them also had high SAC and coincided with those having excessive CNV calls. This indicates the abnormally high number of calls per sample and low DDR was predominately driven by over-calling of false duplications in samples with strong spatial correlation. In practice, the median and MAD of the distribution of log-transformed DDR were calculated and samples falling more than four MADs from the median were removed (Figure 2.9).

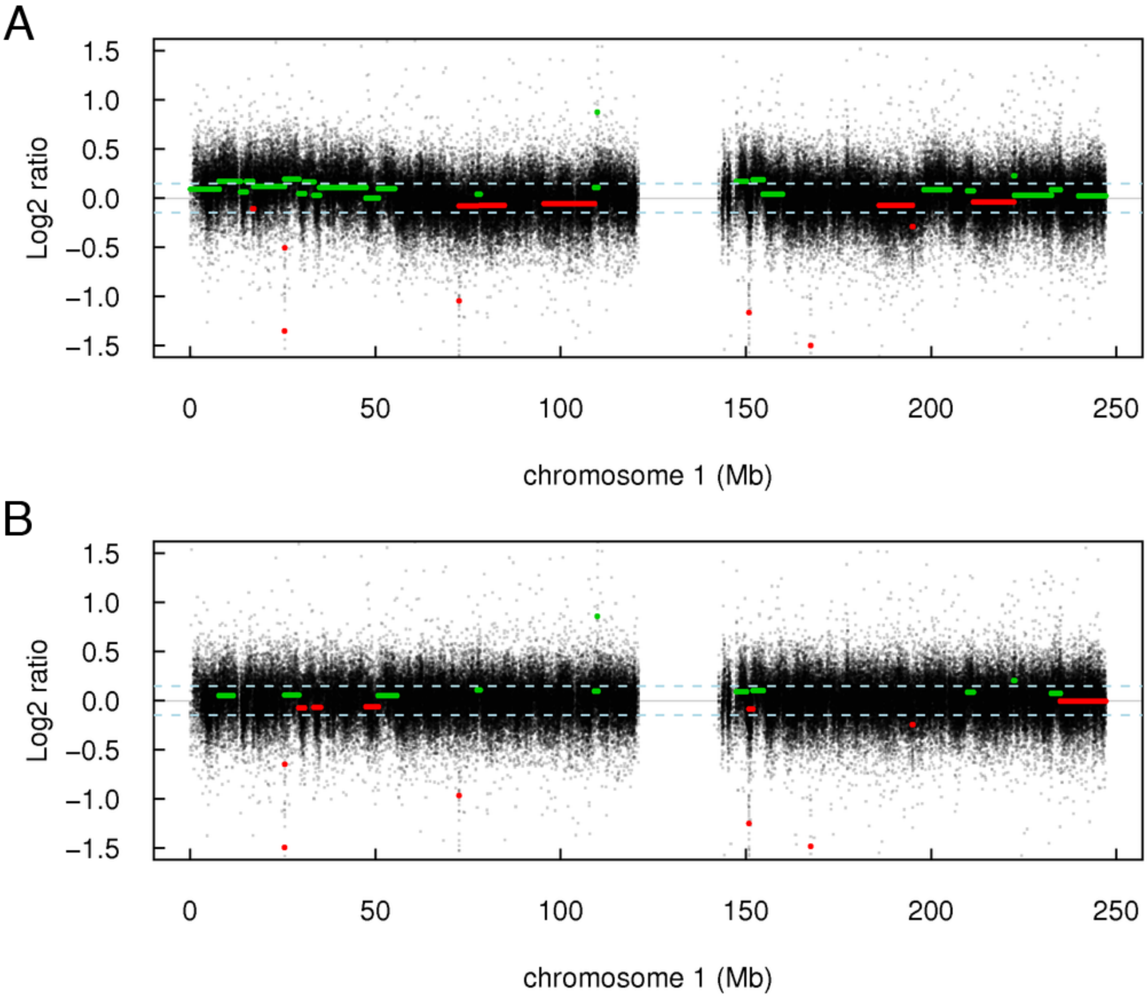

Figure 2.7: Example of over-calling due to spatial waviness along the chromosome. Black dots are $\log_2$ ratios across chromosome 1 of a sample with a SAC of 0.78 (top 0.15%) before (A) and after (B) correction for spatial auto-correlation. Horizontal segments are deletions (red) and duplications (green) called by GADA. The average $\log_2$ ratios of the CNV calls are represented by the horizontal position of the segments. Those located between -0.15 and 0.15 (light blue dashed lines) will be filtered out.

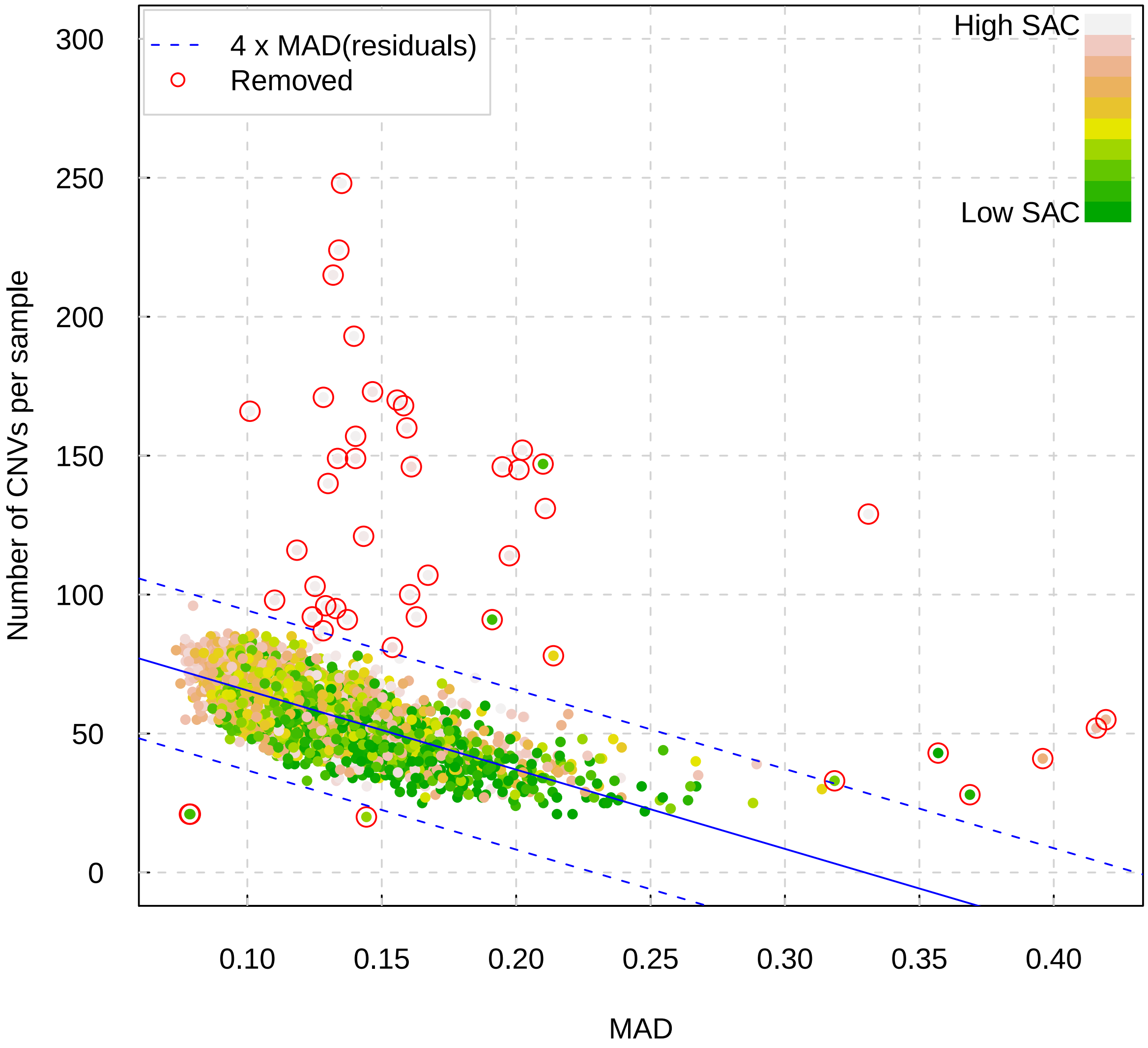

Figure 2.8: Number of CNV calls per sample as a function of the sample's level of noise. Each solid colored dot represents a sample. The sample's level of spatial auto-correlation is coded by terrain color, where the more greenish the lower the level of spatial auto-correlation. The blue solid line denotes the fitted linear model. The blue dashed lines represent four times the MAD of the residuals away from the fitted line. Dots encircled by red are samples to be removed for falling outside the region bordered by the blue dashed lines.

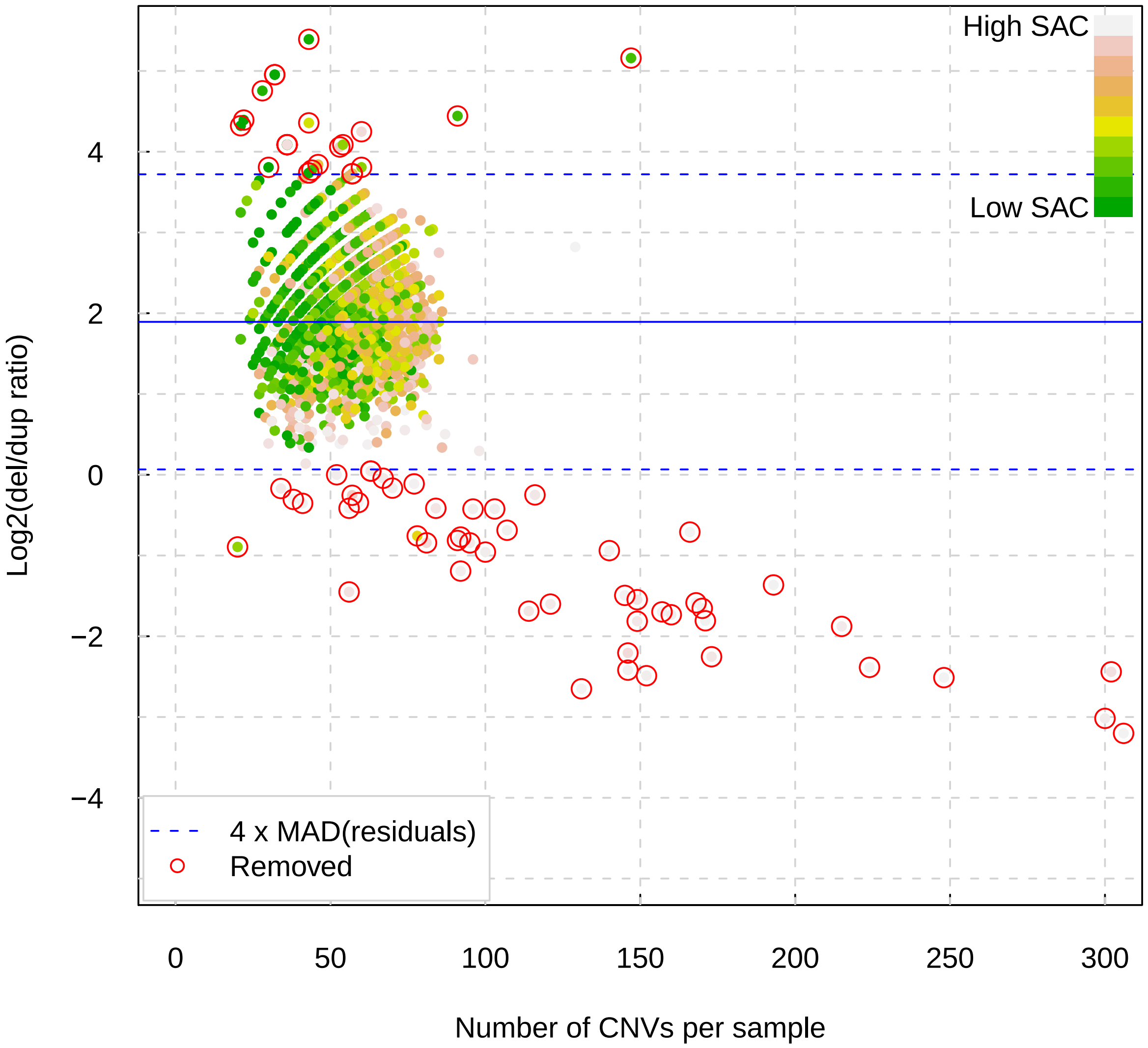

Figure 2.9: Distribution of deletion-to-duplication ratio. Each solid colored dot represents a sample. The sample's level of spatial auto-correlation is coded by terrain color, where the more greenish the lower the level of spatial auto-correlation. The blue solid line denotes the median of $\log_2$ deletion-to-duplication ratio. The blue dashed lines represent four times the MAD of the residuals away from the median. Dots encircled by red are samples to be removed for falling outside the region bordered by the blue dashed lines.

2.3.2.5 Merge split CNV calls

Birdseye sometimes incorrectly split large CNVs into multiple smaller calls due to just a few probes in the middle that did not meet the expected level of dosage-responsiveness (Figure 2.10). I added an *ad hoc* step to merge these split calls based on the number of probes between adjacent calls, the ratio of the number of probes between adjacent calls to the number of probes in the merged call, the probe density between adjacent calls and the absolute difference in log ratio between adjacent calls (see section 2.2.8). My selection of merging parameter values was guided by the distribution of the above metrics (Figure 2.11) and visual inspection of the merged calls.

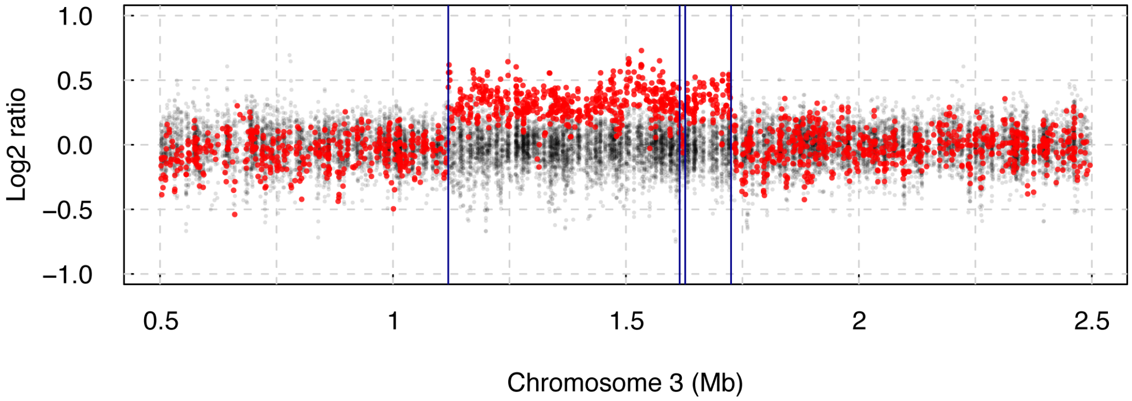

Figure 2.10: Birdseye split a duplication of 620kb into two duplication calls of 500kb and 110kb, respectively. The sample carrying the duplication is highlighted in red. Other samples in the same plate are in black. Blue vertical lines denote the boundaries of the two duplication calls.

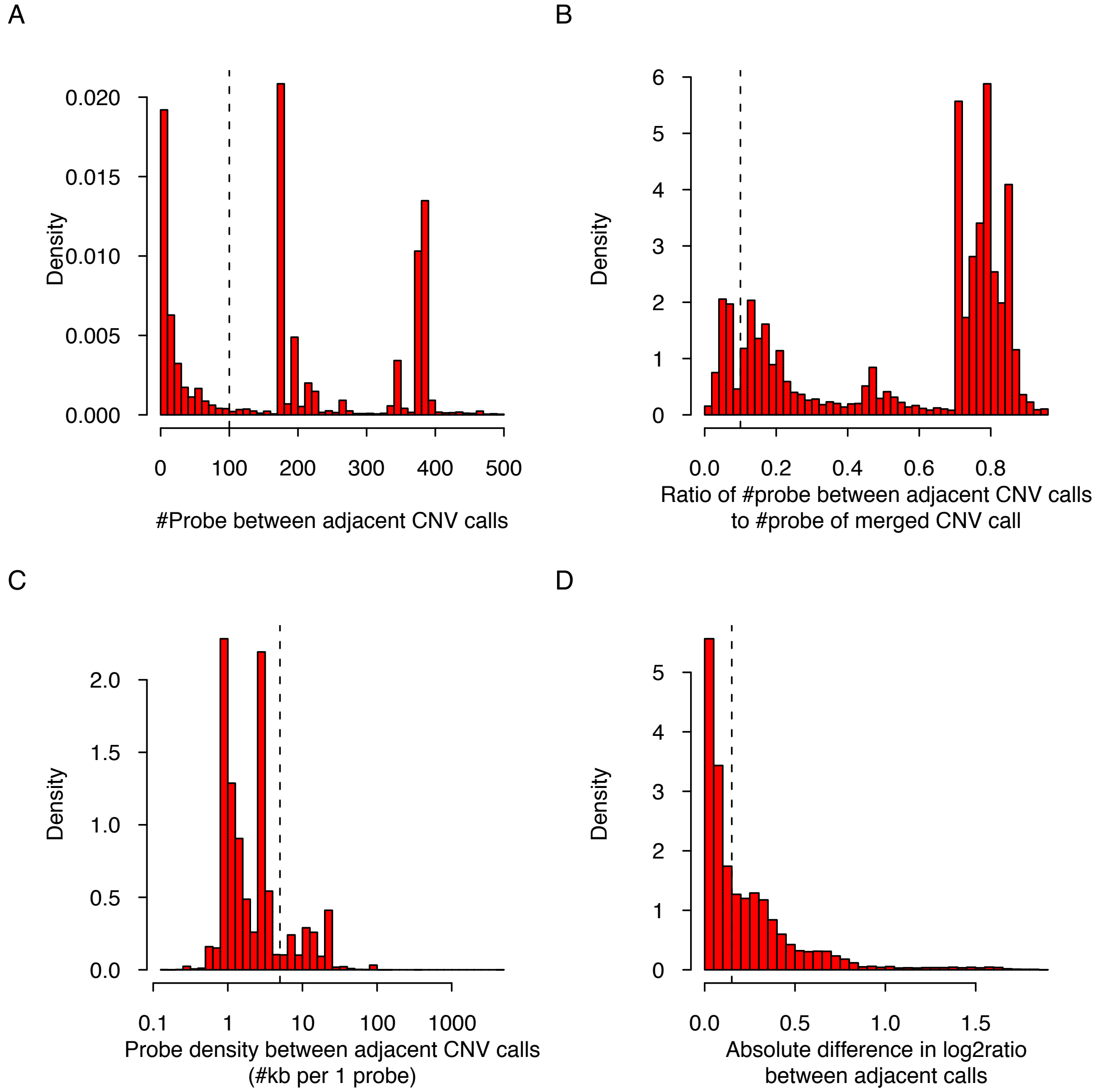

Figure 2.11: The distribution of variables used as metric for deciding if adjacent calls should be merged. Vertical dashed lines mark the thresholds.

2.3.2.6 Cluster CNVE and calculate CNVE frequency

I observed frequently that CNV calls discovered in one sample had extensive overlap with CNV calls in multiple other samples, which likely indicate these variants were identical and probably result from a single ancestral mutation event. Under such a scenario, any slight differences in location were probably just technical fluctuations in the precision of CNV discovery. Even if two overlapping CNV calls originated independently in different individuals and had real slight differences in their locations, operationally treating them as a single event was reckoned reasonable in most analyses considering the highly similar genomic content they encompass and the utility of knowing the frequency of a CNV of a particular genomic interval. I used a clustering-like algorithm to merge such CNVs into CNVEs (see section 2.2.9) and the frequency of a CNVE was calculated as the number of individuals carrying this CNVE divided by the sample size. This call frequency is not the same as an allele frequency, but is nevertheless useful in downstream analyses to distinguish between common and rarer CNVs.

2.3.3 Application of the pipeline to process Affy6 datasets

I applied the above CNV discovery pipeline to the following cohorts genotyped using Affy6:

1. 5,989 UK individuals recruited as common controls in the Wellcome Trust Case Control Consortium 2 project (referred to as WTCCC2).
2. 1,442 American individuals of European ancestry recruited as controls for the GAIN study of Schizophrenia and Bipolar disease (referred to as GAIN_EA, Genetic Association Information Network, European Ancestry)
3. 226 prenatal samples with major ultrasound abnormalities or multiple soft markers detected by standard two dimensional ultrasonography, (referred to as AFD, Abnormal Fetal Development)
4. 334 UK patients with sever-early-onset obesity, half of which also had developmental delay (referred to as SCOOP1, Severe Childhood-Onset Obesity Project 1)
5. 1,386 UK patients with severe early-onset obesity (referred to as SCOOP2)

The biological interpretation of the CNVs identified in these cohorts is described in other chapters in this thesis. Here I focus on the performance of the CNV discovery pipeline across a range of different datasets, generated in different laboratories. I compared the QC metrics and CNV statistics of these cohorts, examined the reproducibility of CNV discovery using this CNV calling pipeline and investigated if commonly adopted QC metrics for SNP genotyping are also appropriate for CNV QC.

2.3.3.1 Comparing QC and CNV statistics

I first examined the distribution of level of noise and spatial autocorrelation in the five datasets. Small but statistically significant differences in the distribution of the level of noise were observed both between control cohorts and between controls and cases (Figure 2.12A). The level of noise (as measured by the MAD of log ratios) was lower in WTCCC2 as compared to GAIN_EA ($p = 2.2\times10^{-47}$, two-sided Mann-Whitney U test, same for the following), AFD ($p = 8.3\times10^{-9}$), SCOOP1 ($p = 8.6\times10^{-13}$) and SCOOP2 ($p = 1.1\times10^{-5}$). These differences largely explained the differences in the distribution of number of calls per sample among the different cohorts (Figure 2.12B, r = -0.88, p = 0.04). Significant differences in the distribution of spatial autocorrelation were also observed. SCOOP1 and SCOOP2 samples had significantly greater median spatial autocorrelation ($p = 6.3\times10^{-60}$ and $p = 2.0\times10^{-100}$, respectively, as compared with WTCCC2) and more samples with very high spatial autocorrelation than the rest of the cohorts (Figure 2.12C). This explained their lower sample QC pass rate (Figure 2.12D, r = -0.93, p = 0.02).

Table 2.10: **Summaries statistics of CNV call sets**

Cohort	#Sample pass QC	Median #call per sample	Median call size (kb)	Deletion-to-duplication ratio	#CNVE	Median CNVE size (kb)	%Singleton CNVE	Average plate size
WTCCC2	5897	58	23.6	3.73	12295	37.9	62.8	84
GAIN_EA	1419	49	27.0	3.83	4493	42.9	63.4	85
AFD	224	50	22.3	5.13	1432	33.1	58.9	38
SCOOP1	292	55	23.5	4.09	2173	32.7	61.7	67
SCOOP2	1289	56	23.3	3.91	5277	37.9	64.5	87

Next I compared the summary statistics of the final filtered call set of the different cohorts (Table 2.10). As discussed above, GAIN_EA and AFD produced fewer calls per sample due to having noisier intensities ($\log_2$ ratios). GAIN_EA had larger CNVs and CNVEs, possibly due to lower sensitivity to smaller events, but there could be other contributing factors. The differences in the proportion of CNVEs seen only in one sample (singletons) might be partly explained by differences in the sizes of the cohort and possibly the impact on sensitivity of differences in average batch (plate) sizes (r = 0.97, p = 0.006), since Birdseye receive parameters of the

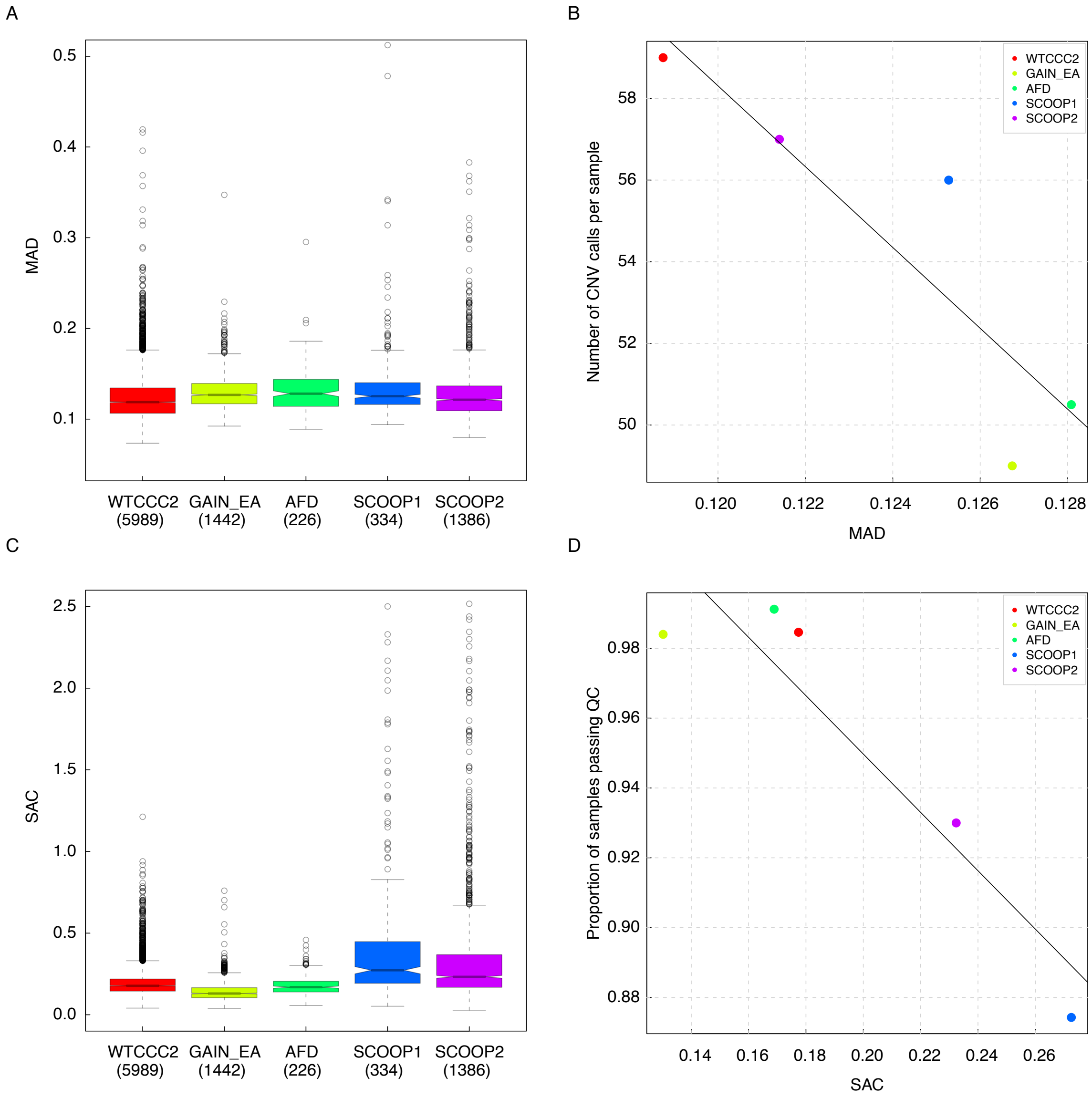

Figure 2.12: Comparison of QC statistics. (A) Distribution of the level of noise. (B) The number of CNV calls per sample as a function of the level of noise. The points represent the median value for each cohort. (C) Distribution of spatial autocorrelation. (D) Sample QC pass rate as a function of the level of spatial autocorrelation. The x value of each point is the median SAC for each cohorts.

emission probability distribution from Canary and Canary could overestimate the variance of the intensity distribution of the neutral copy state when given a smaller number of samples, which would lead to under-calling of singletons. The disease cohorts (AFD, SCOOP1 and SCOOP2) had greater deletion-to-duplication ratios, which might reflect ture biological differences, but more likely is due to technical biases, as duplications become more difficult to call than deletions with noisier data and smaller plate sizes.

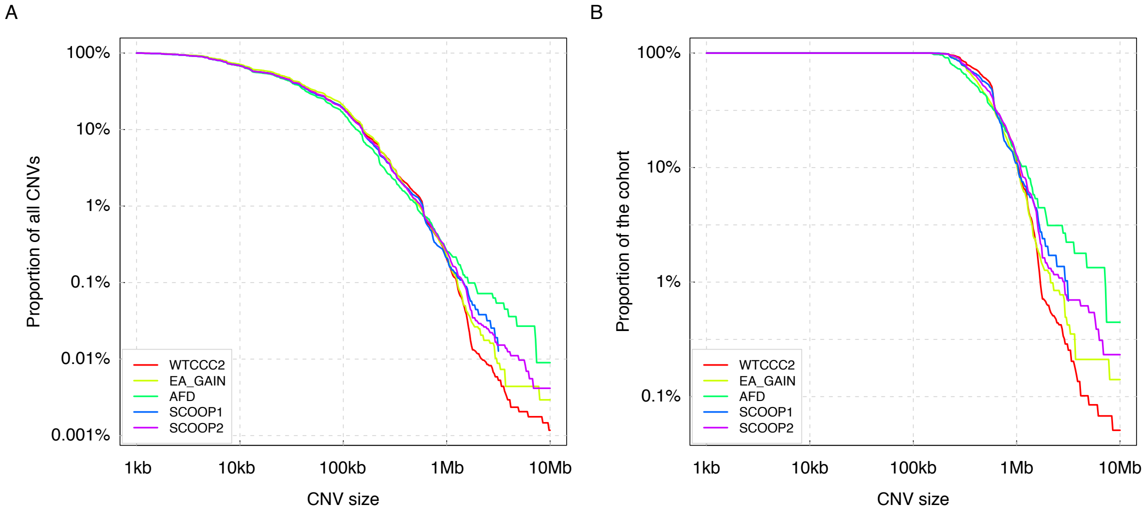

Figure 2.13: Proportion of large CNVs relative to all CNVs as a function of size threshold (A) and proportion of the cohort carrying large CNVs as a function of size threshold (B). Both CNV sizes and proportions are in log scale.

Finally, I investigated if there is difference in the distribution of large CNV calls in the call sets. As the calling of large CNVs should be least affected by technical issues, this could provide insights into the biological characteristics of the cohorts. For CNVs exceeding a certain size threshold, I calculated their proportion relative to all CNVs and the proportion of the cohort carrying such CNVs. The proportions remained relatively similar until the threshold reached 1Mb, beyond which disease cohorts (AFD, SCOOP1 and SCOOP2) had both greater proportion of large CNVs and greater proportion of individuals carrying such CNVs (Figure 2.13).

2.3.3.2 Reproducibility of CNV discovery using Affy6 plus the pipeline

There were 55 SCOOP1 patients genotyped for a second time using Affy6 as part of SCOOP2 (46 passed sample QC both times), which provided an opportunity to investigate the reproducibility of CNV discovery using the CNV discovery pipeline I developed. Samples of 46 of those patients passed QC in both datasets. For each individual, I defined a CNV called in one dataset 'reproduced' if it reciprocally overlapped >50% with a CNV called in the same individual in the other dataset. Replicate rate was defined as:

$$\frac{N_{\text{reproduced,SCOOP1}} + N_{\text{reproduced,SCOOP2}}}{N_{\text{SCOOP1}} + N_{\text{SCOOP2}}}$$

On average, a replicate rate of 76.8% was achieved. As expected, due to differences in sensitivity and specificity, the replicate rate was much higher for deletions than duplications (Table 2.11). I further interrogated if the concordance ('replicate rate') between CNV sets called in samples from the same individual was higher than that between CNV sets called in samples from different individuals. To do this, for each of the 46 SCOOP1 samples, I calculated replicate rates with 100 randomly chosen SCOOP2 samples. This verified that the observed level of reproducibility between samples from the same individual was not a mere coincidence that could be achieved by pairing randomly chosen samples (Figure 2.14).

Table 2.11: **Replicate rate of CNV discovery using Affy6 and the Birdsuite pipeline**

Type	(1kb,10kb]	(10kb,20kb]	(20kb,100kb]	(100kb,+∞]	All Classes
Duplication	40.00%	63.08%	46.20%	60.97%	54.56%
Deletion	80.08%	76.15%	87.46%	81.25%	82.10%
All Classes	79.19%	74.93%	78.65%	70.82%	76.83%

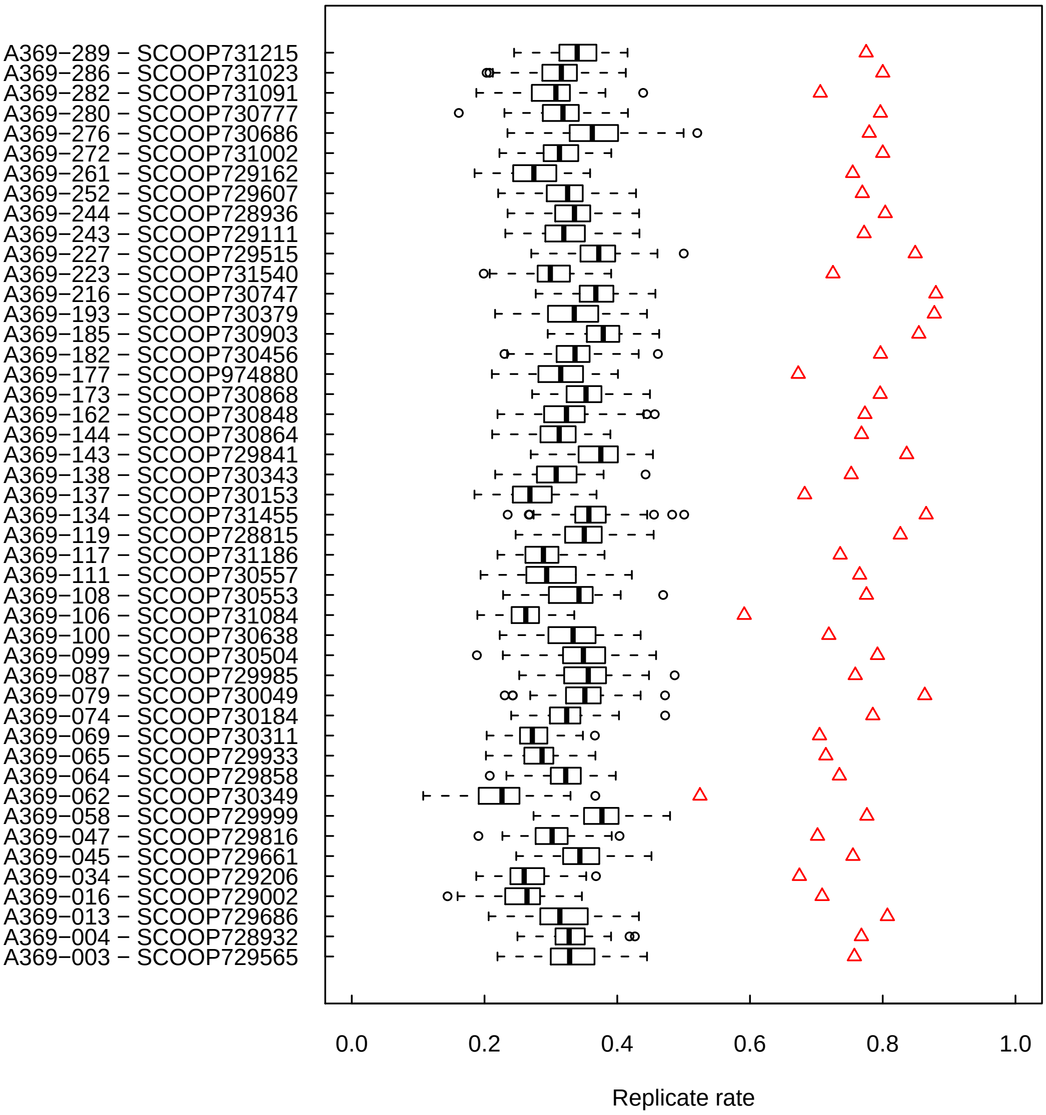

Figure 2.14: Comparing replicate rate between randomly chosen pairs of samples and samples that are true replicates. The distributions of replicate rate between randomly paired samples are presented by boxes whereas replicate rate between true replicates as printed in the labels are presented by red triangles.

2.3.3.3 SNP genotyping QC metrics are not suitable for CNV QC

SNP call rate and the level of heterozygosity are sample QC metrics that are frequently used in SNP GWAS where samples having low SNP call rate or being outliers in the distribution of the level of heterozygosity were removed [13, 39]. I investigated if these metrics are also appropriate for identifying samples to be removed for CNV analyses. I examined the distribution of SNP call rate and the level of heterozygosity against the two metrics I used for CNV sample QC, the number of calls per sample and deletion-to-duplication ratio (DDR). As shown in (Figure 2.15), the majority of samples having low SNP call rate or being outliers in the distribution of the level of heterozygosity yielded similar number of CNV calls or DDR to samples with high SNP call rate and normal level of heterozygosity. Samples with extreme number of calls and DDR are close to the mode of the distribution of SNP call rate and the level of heterozygosity, implying that filtering samples for downstream CNV analyses by using these SNP-based QC metrics would not be useful.

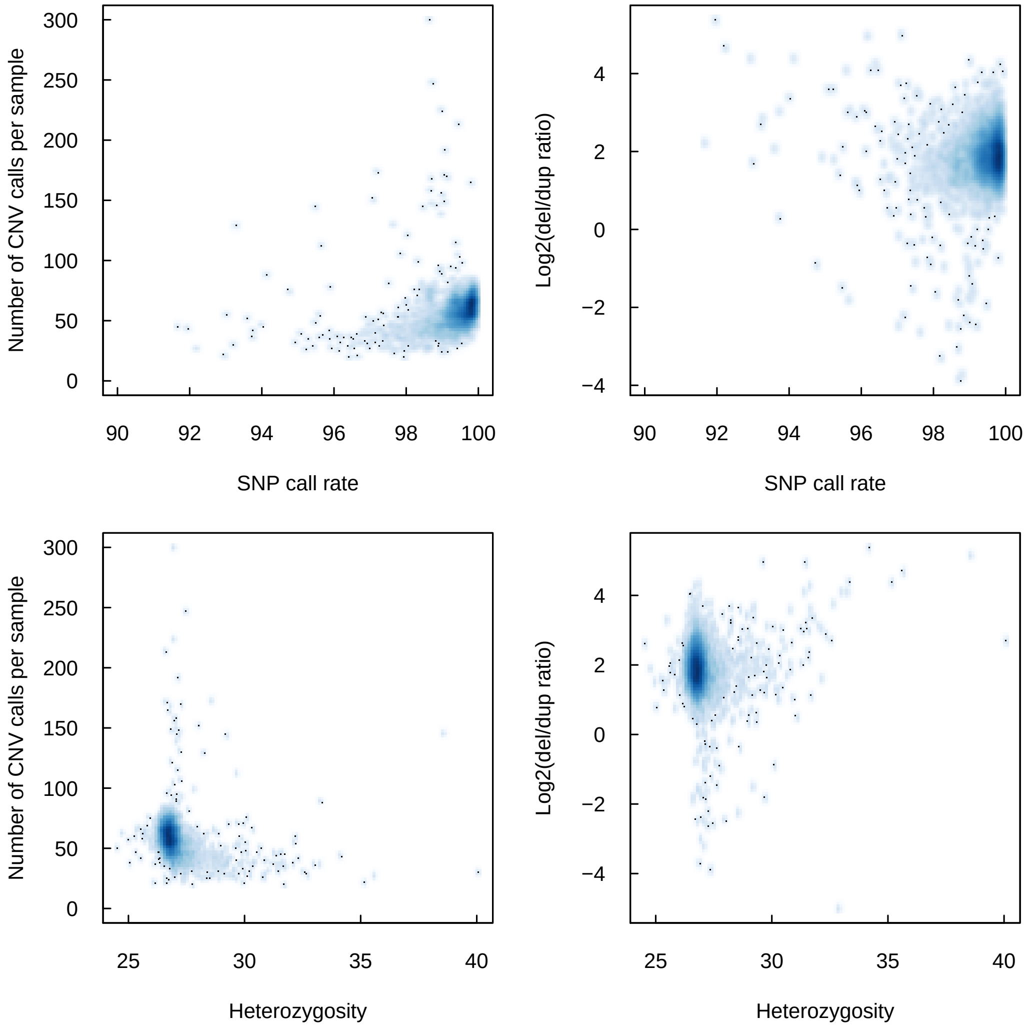

Figure 2.15: Distribution of SNP QC metric against CNV QC metric.

2.4 Discussion

In this chapter, I described the development of a CNV calling pipeline for Affy6 genotype data. I first compared the performance of three CNV calling programs on Affy6 data. Next, I built a production pipeline around the selected program, Birdsuite, which incorporated a number of QC metrics and post-processing procedures. Finally, I applied the pipeline to several Affy6 datasets to produce filtered CNV call sets for further analysis. I have demonstrated that the pipeline I developed generated high quality CNV call sets across a range of different datasets.

2.4.1 Storage of CNV data

The amount of data a single microarray experiment can produce increases linearly with the increase in resolution and coverage of the array. With 6,892,960 oligonucleotide features on the slide, Affy6 yields nearly seven million raw intensity values per experiment, which are summarized to more than 2.7 million intensity values corresponding to nearly one million copy number probe sets and one million SNP probe sets. The summarized and normalized intensities, together with minimal annotations required for CNV calling, including probe set ID, type and genomic location, take up close to 2Gb of disk space for 96 samples if stored in plain text format. Such data are required not only for CNV calling but for QC and various visualizations as well, in which scenario fast access to intensity values of certain samples within a genomic window of interest is needed. Due to hardware limitations, I used a special HDF file to store and manage such data. Although the genomic coordinates were indexed to facilitate fast query, the speed is still affected and limited by disk performance. However, relationships among samples and other annotations associated to probes and samples have to be managed separately. Ideally, all those data should be stored in an efficiently designed database with an index loaded in memory.

2.4.2 Log ratio versus intensity

Birdsuite outperformed GADA and APT for CNV discovery from Affy6 data in the comparisons I performed. Its CNV discovery component, Birdseye, is the only program of the three that works on intensities rather than converted log ratios. In principle, the conversion to log ratio should reduce the variation across different probe sets. Actually, this explains why most general-purpose (multi-platform) CNV discovery programs expect to work with log ratios: for segmentation-based methods, simple piecewise constant function can be used to model log ratio profiles along a chromosome and for HMM-based method, the same parameters for emission model can be used for all probe sets. By comparison, algorithms working with intensities produced from single channel genotyping arrays need special treatment to handle the larger variance across probe sets (*e.g.* Birdseye uses probe-set-specific emission parameters), which often limits their application to other types of arrays. However, calling CNVs from intensities also has advantages over log ratios. First, strictly speaking, log ratios can only indicate comparative loss or gain of copy number relative to a reference rather than an actual genotype. For the APT and GADA test pipeline, log ratio were converted from intensities using a population (all samples in a plate) median as the reference, which could deviate from diploidy in common CNV regions and lead to a more balanced 'deletion' 'duplication' ratio, as shown in Table 2.1, which, if taken at face value, could give a misleading view of the nature of CNVs in an individual's genome. Second, discriminating high copy number states is much harder using log ratios than using intensities especially when the reference's copy number is greater than two.

2.4.3 CNV discovery QC filter parameters

Due to differences in array specification, CNV discovery algorithm and purpose of investigation, there is little consensus in what filters should be applied for CNV discovery. Without an independent and high quality CNV call set in the same individuals, previous studies often have had to rely on simulated data or indirect measures [24, 26, 28]. Rather than using stringent filtering, I have instead used instead fairly permissive filters in the production pipeline, as Birdseye calls with a LOD ≥ 10 already have reasonably high specificity even for smaller events (in the size range of 1kb to 10kb (Table 2.9) or having 5 to 10 probes (data not shown)), as judged in the comparisons with the ng42M call set.

2.4.4 CNV discovery sample QC

The sample QC method I developed for this pipeline is relatively simple yet effective. By quantifying the two primary data quality factors that affect CNV discovery performance, spatial auto-correlation and noise, the method is able to clearly distinguish samples of acceptable quality, in which the number of CNV calls per sample follows an expected inverse correlation with the level of noise, and samples that are apparent outliers to this trend. This pattern of separation has been consistently observed in several Affy6 datasets, and in principle should be applicable to CNV discovery pipelines for other arrays and sequence data as well.

The QC methods can still be improved. In analyses described in later chapters in this thesis, I found that data quality was not equally poor throughout the entire genome and good quality CNV calls at specific loci could still be salvaged in some of the samples that had failed the QC thresholds described here. Therefore, a finer QC procedure that filters by chromosomes rather than by samples might prove useful.

2.4.5 CNV clustering versus joint calling

An *ad hoc* CNV clustering step was deployed at the end of the discovery pipeline to combine CNVs called from each individual that likely correspond to the same mutation event and to calculate a lower bound on the numbers of individuals carrying such events in a population. Based on reciprocal overlap, the generality of this method ensures it can be applied to all CNV call sets produced by various CNV discovery pipelines. However, a better solution would be to statistically model CNVEs and to call CNVs jointly from multiple individuals rather than calling CNVs one individual at a time. As pointed out by Zöllner [40], sharing information across individuals should not only increase the sensitivity of calling common CNVs but also make estimation of the border of CNVE more accurate.

2.4.6 Merging split CNV calls

There have been a number of reports that large CNVs are sometimes incorrectly split by both HMM-based methods [11, 33, 41] and segmentation-based methods [12, 27]. In this pipeline, I introduced a merging step after call QC and sample QC. For some large CNVs, since each individual split calls produced by Birdseye did not meet the call QC thresholds, they could not be caught by the merging step and hence missed from the final call set. An alternative design would be to merge CNV calls prior to any call filtering. However, it would be difficult to derive a LOD score for the merged CNV call to allow it to be filtered along with other, unmerged calls. A neater solution would be to reduce the probability of splitting large CNVs at the discovery stage. HMM-based methods such as Birdseye currently often use distance-aware transition probabilities wherein the likelihood of a probe having a different copy number state from its previous probe increases as the distance to the previous probe increases. These distance-aware transition probabilities are independent of the location of the probe. With the current knowledge of spatial distribution of CNVs across the genome, location-aware transition probabilities could be introduced. With the availability of large amount of Affy6 data, one can calculate a signal-to-noise ratio for every probe and weight probes on their signal-to-noise ratio during CNV calling. Such information has been used in segmentation-based method [42] and can also be incorporated into the Viterbi algorithm for HMM-based methods.

2.4.7 Application of this pipeline

This pipeline has been successfully applied to a number of cohorts genotyped using Affy6, ranging from apparently healthy genomes and patient genomes with subtly different patterns of CNV from controls (see following chapters), and should be applicable to the majority of disease cohorts. However, this pipeline was not designed for CNV discovery in cancer genomes, since (*i*) the emission model parameters of Birdseye were estimated only for the pre-defined states corresponding to copy number of integer 0–5, and (*ii*) spatial auto-correlation and level of noise as measures of data quality only applies to genomes with a limited amount of CNV and the assumption that the number and the deletion-duplication-ratio of CNVs discovered in one sample should be relatively comparable among individuals does not hold for cancer genomes.

CHAPTER 3

COPY NUMBER VARIATION AND SEVERE EARLY-ONSET OBESITY

3.1 Introduction

3.1.1 The genetics of obesity

Obesity is a medical condition in which an excess of body fat has accumulated to the extent that it may have an adverse effect on health. The addition to its social and psychological effects, the incidence of obesity is highly correlated with increased morbidity of type II diabetes, hypertension, coronary artery disease, many forms of cancer and reduced life expectancy [43]. Today, nearly one fifth of the UK population can be defined as being clinically obese by having a body mass index (BMI) greater than 30 [44]. The role of 'environmental' factors in the development of obesity is apparent, as the increasing prevalence of obesity is coupled with an increase in dietary energy intake and a more sedentary lifestyle over past decades. However, the heritability of BMI estimated from studies of large number of monozygotic twins adopted as infants and raised separately in unrelated families ranges from 0.4 to 0.8 [45–47], indicating a strong genetic determinant in relative body weight.

3.1.1.1 Physiological basis of body weight control

Although the nature of such genetic determinants of obesity has not been fully understood, they must act through the long-term control of energy intake and expenditure. Such control resides in the part of the brain called the hypothalamus through regulation of appetite by the leptin-melanocortin signaling pathway (Figure 3.1). This pathway was largely characterized through genetic studies in mice [48–50]. Leptin is an adipose-derived hormone that circulates through blood. It interacts with leptin receptors on first order neurons at the hypothalamic arcuate nucleus, activating proopiomelanocortin(*POMC*)-producing neurons and suppressing neuropeptide-Y(*NPY*)/Agouti-related peptide(*AGRP*)-producing neurons. The former leads to the cleavage of *POMC* into α-melanin stimulating hormone (α*-MSH*) that exerts catabolic actions through melanocortin-4-receptor (*MC4R*) and melanocortin-3-receptor (*MC3R*) and the latter causes decrease in food intake. The *NPY/POMC*-producing neurons also project to the hypothalamic paraventricular nucleus, which has long been identified as a 'satiety center' [51]. In this way, the long-term energy balance is maintained by the feedback between body fat and regulation of appetite and catabolism via leptin.

3.1.1.2 Monogenic and syndromic obesity

Human mutations throughout the leptin-melanocortin signaling pathway have been found to produce Mendelian disorders in which severe obesity is the most obvious phenotype [53–55]. The majority of those mutations are dominant. Obesity is usually developed in childhood with some patients rapidly gaining weight just weeks after birth, and most are accompanied with hyperphagia [56]. Obesity caused by congenital deficiency of leptin can be effectively treated by administration of leptin [57], but defects in later steps of the pathway currently have no targeted therapy.

Apart from the monogenic form of obesity that is primarily caused by mutations in appetite-controlling pathways, at least 20 rare syndromes are also characterized by obesity [43]. Most of these obesity syndromes are distinguished by the presence of mental retardation, such as Prader-Willi syndrome, Pseudohypoparathyroidism

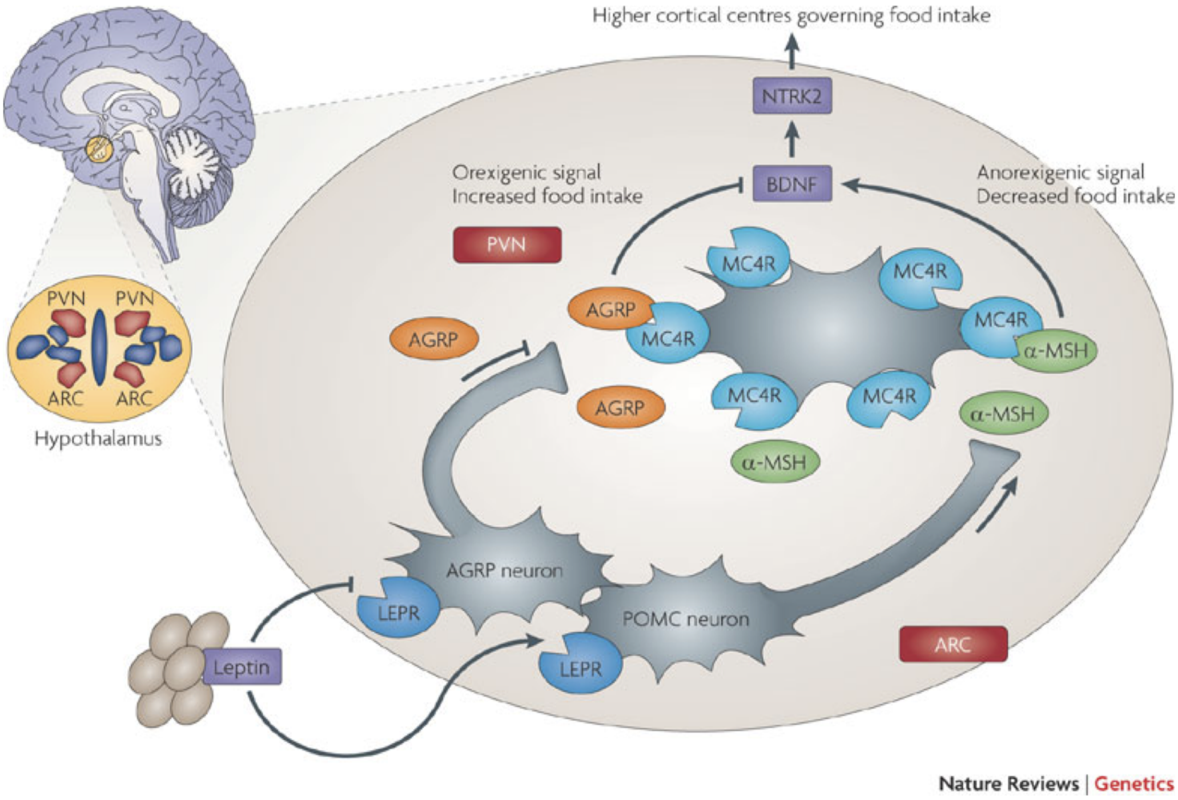

Figure 3.1: The leptin-melanocortin pathway. ARC: arcuate nucleus; PVN: paraventricular nucleus. Figure taken from Walley *et al* [52].

type1A (PHP1A) syndrome, Bardet-Biedl syndrome (BBS), etc. The causes of these syndromes are diverse, and both discrete point mutations and large chromosomal abnormalities have been shown to play a role [43, 58, 59].

3.1.2 Previous discoveries of obesity related loci

Genes and mutations discovered so far only account for a small fraction of extreme early onset obese cases. For example, mutations in MC4R, despite being the most common known cause of monogenic obesity, is found in only 1-6% of obese individuals from different ethnic groups, and the frequency is lower in cases with a less severe phenotype [60]. There has been continued effort to search for novel genes and variants that might cause obesity and account for the heritability of relative body weight. Much progress has been made in recent years, especially for population variation in BMI.

3.1.2.1 Family-based linkage studies

This method involves the genotyping of families of a proband using polymorphic markers throughout the genome and calculating the degree of linkage of each marker to the disease trait. A number of loci have been found to be linked to common or severe obesity, such as 2p21-p23 [61, 62], 3q27 [63, 64] and 20q11-q13 [65, 66]. However, these linkage intervals are large and have proven to be difficult to replicate due to issues in sampling, phenotyping and statistical power, and hence linkage studies have been more or less superseded by genome-wide association studies in recent years.

3.1.2.2 Genome-wide association studies (GWAS)

This method entails genotyping a large number of common polymorphic markers throughout the genome in large cohorts of unrelated cases and controls and tests the association of each marker with the trait in question. In 2007, *FTO* became the first gene found to be associated with BMI by GWAS [67]. This finding was replicated in multiple cohorts, with an estimated increase in BMI caused by one copy of the risk allele being 0.2–0.4kg/m^2 [68–70]. A year later, a second association signal, a SNP downstream of *MC4R*, was found and replicated in cohorts of individuals of European descent [71]. In 2009, a meta-analysis of 15 GWAS for BMI in cohorts of European descent conducted by the GIANT consortium not only replicated associations at *FTO* and *MC4R*, but also discovered six new associated loci at which several of the likely causal genes are expressed or known to act in the central nervous system [72]. More recently, 18 more BMI-associated loci were discovered by GWAS in even larger cohorts [73]. However, all confirmed associated loci together only explain $\sim$1.45% of the variance in inter-individual BMI [73], while further increasing sample size using current genotyping chip designs is likely to find only common variants of even smaller effect size.

3.1.2.3 Candidate gene association testing

The candidate gene approach involves genotyping polymorphic markers or gene resequencing in a candidate gene of putative relevance to obesity in cases and controls. Such candidates can come from current knowledge of the etiology of the disease, or genomic intervals where linkage or association was found by whole genome approaches.

3.1.3 CNV-disease association

In principle, a disease with a genetic etiology can be caused by any type of genetic lesion; some of these lesions will be SNPs and some will be CNVs [1]. Large chromosomal abnormalities have been known to cause both inherited and sporadic diseases long before the discovery of the genome-wide prevalence of CNVs in the general population. Some of these abnormalities are cytogenetically detectable and many are flanked by long segmental duplications that make the region susceptible to re-arrangements mediated by Non-Allelic Homologous Recombination (NAHR). Well-known examples include the 22q11.2 deletion, which is responsible for the DiGeorge syndrome [74], the 17p11.2 duplication, which is responsible for Charcot-Marie-Tooth syndrome type1A [75].

Following the discovery of common CNVs in the general population [38, 76–80], their functional impact has been fervently sought after with the hope that some of them might explain part of the 'missing heritability' left by SNP GWAS. A few disease associations with common CNVs have been reported, such as deletions upstream of IRGM, which is associated with Crohn's disease [81], a multi-allelic CNV at CCL3L1, which influences susceptibility to HIV-1/AIDS and rheumatoid arthritis [82, 83] and a $\sim$43kb deletion upstream of *NEGR1*, which is associated with increased BMI [72]. However, a comprehensive study of disease association of all common CNVs $>$500bp undertaken by the WTCCC revealed that except for a limited number of loci, the vast majority of common CNVs that could be genotyped using current technology do not associate with the studied diseases and are unlikely to have substantial impact on common diseases in general. For the small number of loci that do exhibit association, the CNVs are typically well-tagged by common SNPs and have been captured by previous SNP GWAS, indicating that common CNVs are unlikely to explain the 'missing heritability' for common diseases [13]. Therefore, much attention has shifted towards rare CNVs in rare diseases, wherein variants might be expected to have larger effect sizes and are unlikely to be fully captured by common SNPs.

Studies in moderately rarer neurodevelopmental disorders, such as schizophrenia, have been especially fruitful. In addition to observations of an increased genome-wide burden of large and rare CNVs that disproportionally disrupt neurodevelopmental pathways in patients compared to controls, associations involving *de novo* or recurrent CNVs at specific loci, including deletions at 1q21.1, 15q13.3 and 22q11.2 and duplications at 16p11.2 were discovered and replicated [11, 84–86]. Similar findings have been reported for autism and related phenotypes, including specific associated CNVs, increased genome-wide CNV burden and functional enrichments within CNV-disrupted genes [9, 87–90]. While some of the discovered disease-CNVs are highly penetrant, others may act as predisposing factors and exacerbate phenotype in association with other large rare CNVs [91].

In this chapter, I will describe two CNV case-control studies on severe early-onset obesity. The first one involves a relatively small patient cohort that is enriched for patients with syndromic forms of obesity (Section 3.3.1). The second study involves a larger cohort of patients with only severe early-onset obesity (Section 3.3.2). The first study only investigated the role of rare CNVs, whereas both common and rare CNVs were examined in the second study.

3.2 Materials and methods

3.2.1 Patient and control data

The 1,656 UK obese patient samples are from the SCOOP (Severe Childhood Onset Obesity Project) cohort, a selected subset of patients recruited to the Genetics of Obesity Study (GOOS) on the basis of severe obesity defined as a BMI standard deviation score (BMI sds) >3 and onset of obesity before 10 years of age [92]. They have normal karyotype and do not have mutations in *LERP*, *POMC* and *MC4R* as determined by prior sequencing conducted at the Metabolic Research Laboratories, Addenbrooke's hospital. Some of these patients were ascertained with developmental delay in addition to obesity. The 1,656 samples were divided into three sub-cohorts: 959 obese-only patients of self-reported European ancestry, referred to as SCOOP1, 325 patients of self-reported European ancestry, of which 143 have developmental delay in addition, referred to as SCOOP3, and the remaining 374 of patients out of which 219 have developmental delay in addition and 15 self-reported as being of non-European ancestry, referred to as SCOOP2. SCOOP3 were referred to as SCOOP1, and SCOOP1 and SCOOP2 were referred to as SCOOP2 in Chapter 2. The initial study described in Section 3.3.1 only investigated SCOOP3, whereas the following study described in Section 3.3.2 included all of the three sub-cohorts.

The 7,431 apparently healthy individuals are drawn from two sources. The first set includes 5,989 UK individuals recruited as common controls in the GWAS of 13 diseases undertaken by the Wellcome Trust Case Control Consortium 2 (WTCCC2), of which $\sim$50% of samples are from the 1958 British Birth Cohort and $\sim$50% of samples are from the UK Blood Service Control Group. The second set of 1,442 control individuals, all of European-American ancestry, are from a subset of a control cohort used in a GWAS of schizophrenia and bipolar disease undertaken by Genetic Association Information Network (GAIN). Samples from both patients and controls were previously genotyped on Affymetrix genome-wide human SNP array 6.0. Affymetrix 6.0 .CEL files for cases were obtained from the Metabolic Research Laboratories, Addenbrooke's hospital and the Microarray facilities, Sanger Institute, and .CEL files for controls were from the Wellcome Trust Case Control Consortium

2 for WTCCC2 controls and from the Database of Genotype and Phenotype (dbGaP) through accession number phs000017 and phs000021 for GAIN controls.

3.2.2 Permutation test of CNV burden

To assess the significance of altered CNV burden in cases compared to controls, I randomly permuted the 'case' 'control' labels of samples 10,000 times. To control for confounding factors that might be correlated with affected status such as data quality, measured by median of absolute deviation (MAD) of sample $\log_2$ ratio, and number of all CNVs called per sample (NCPS), permutations were conditioned on these factors, *i.e.* pooled case and control samples were stratified into MAD or NCPS deciles and labels of affected status were only permuted within each decile.

3.2.3 Identifying ethnic outliers

I called the genotypes of $\sim$1M SNP probes included in the Affymetrix 6.0 array using 'Birdseed', the SNP genotype calling module of 'Birdsuite' for all case and control samples, together with the 270 HapMap1 samples (90 European, 90 African and 90 East Asian) and 74 HapMap3 Indian samples. The genotypes were coded as 0, 1 and 2 for loci with homozygous reference alleles, heterozygous alleles and homozygous alternative alleles, respectively. 10,827 SNPs that are at least 20kb apart along the genome were selected as markers to exclude strongly correlated markers as well as to reduce computational load. A Euclidean distance between each pair of individuals was calculated using these markers and the distance matrix was supplied as the input for multidimensional scaling (MDS). Individuals were projected using the first two dimensions that represented inter-population genetic variation. The 'genetic distance' to Europeans was calculated as the distance in the projected space between each individual and the center of the CEU cluster. An empirical genetic distance threshold was adopted above which individuals were regarded as non-Europeans.

3.2.4 Defining CNVEs for test of enrichment

CNV calls in cases and controls were pooled together and then divided into deletions and duplications. CNVEs (see Chapter 2, page 34 and 15, for definition) were clustered from pooled deletions and duplications separately. Each deletion (or duplication) CNVE with a carrier frequency of $<1\%$ was treated as a locus for the test, at which the number of cases and controls carrying deletions (or duplications) covering $>50\%$ of the bases of the CNVE were counted (Figure 3.2).

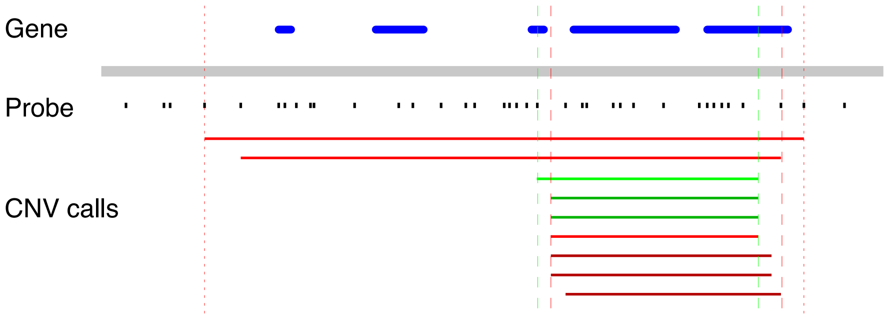

Figure 3.2: Illustration of the unit of test. Red horizontal lines represent deletions and green horizontal lines represent duplications. Control CNVs are in darker colors. Green dashed lines mark the CNVE clustered from duplications. Red dashed lines mark the CNVE clustered from the smaller deletions. Red dotted lines mark the CNVE clustered from the larger deletions. Three tests, each for one CNVE (two deletion CNVEs and one duplication CNVE), will be performed for the illustrated region.

3.2.5 Performing common CNV case-control association testing

For each tested CNVE (genomic window), for each sample a single CNV measurement that summarized the measurements of all probes within the window was generated to perform the test. Three probe measurements (intensities, $\log_2$ ratios relative to plate median and $\log_2$ ratios relative to cohort median) and three methods of summarization (mean, median and first principal component) were considered. The first principal component was calculated from the probe-by-sample matrix. This summarization method accounts for the differences in informativeness among different probes (*e.g.* probes located within the CNVE but outside the actual CNV in the specific sample are less informative of the genotype of the CNV). The resulting principal component usually down-weights probes of which measurements are uncorrelated with the remainder and isolates the variation across samples of different copy number. The summarized measurements (mean, median and first principal component) were then analyzed using the R package CNVtools, which implements a likelihood ratio test that models the distribution of summarized values as a Gaussian mixture and compares the goodness of fit with or without association to affected status [93]. The method takes a pre-defined number of CNV genotypes, models the parameters of the Gaussian mixture using a generalized linear model in which the mean and variance of CNV measurements is dependent on copy number, affected status and other sources of differential errors, such as batch effects, and uses a EM algorithm to obtain the maximum likelihood estimates of the model parameters. I considered the number of CNV genotypes ranging from 2 to 4, which covers the majority of scenarios. Since no single combination of probe measurement, method of summarization and pre-defined number of CNV genotypes worked best for all CNVEs, the test was run under all combinations of settings, therefore yielding 27 test results for each common CNVE (Figure 3.3). These results were subjected to manual examination and the one with most appropriate genotype clustering was selected as the final result. For a small proportion of CNVEs of which meaningful genotype clustering could not be produced under all combinations of settings, the test was re-run with manually tweaked settings. CNVEs that failed manual tweaking were removed from further analyses.

3.2.6 Test of functional enrichment

A modified version of gene sets enrichment analysis developed by Raychaudhuri *et al* [94] was used to test if genes functionally related to obesity were affected more frequently in cases relative to controls. The analysis was based on a logistic model that controls confounders by including them as cofactors. I used the model that controls for the number of CNVs called per sample and the average size of CNVs

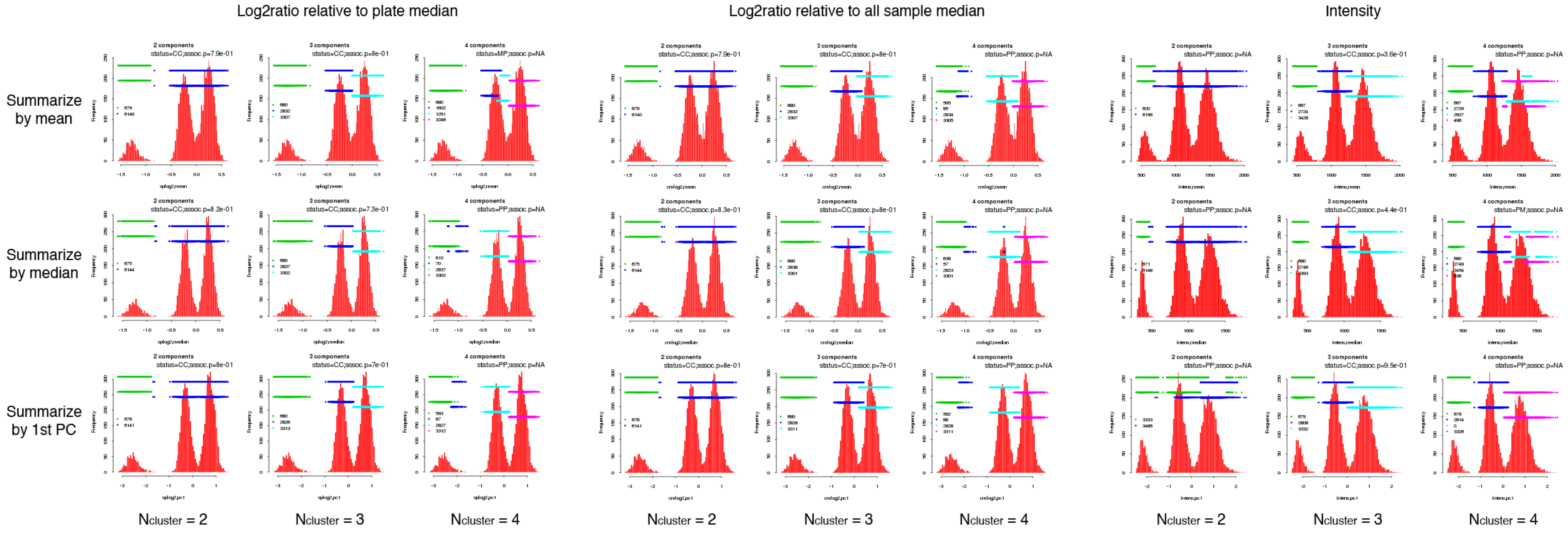

Figure 3.3: An example of test results of one common CNV loci generated by CNVtools under combinations of probe measurement, summarization method and preset number of genotype clusters. These results were subjected to manual examination.

called per sample:

$$\log\left[\frac{p_{i,case}}{1-p_{i,case}}\right] = \theta + \beta_0 \cdot c_i + \beta_1 \cdot s_i + \gamma \cdot g_i + e$$

, where $p_{i,case}$ is the probability that individual i is affected, θ represents the background log likelihood the individual is affected, c_i, s_i and g_i is the number of called CNVs, the average CNV size and the number of CNV affected genes belonging to a gene set of interest in that individual and e is an error term. The analysis tests if γ, the increase in log likelihood per CNV affected gene within the gene set is significantly different from 0. I re-implemented this method in R.

Gene sets were obtained from the Molecular Signatures Database v3.0, which collects annotated gene sets for use with gene sets enrichment analysis [95]. I downloaded the C2 collections which includes canonical pathways, KEGG gene sets, BIOCARTA gene sets, REACTOME gene sets and differentially expressed gene sets in response to chemical and genetic perturbations collected from PubMed.

3.3 Results

3.3.1 Initial analysis of 334 patient samples

CNVs were called from the case (SCOOP3) and control (WTCCC2 and GAIN) cohorts using the pipeline described in Chapter 2 (with slightly different parameters and procedures, as the pipeline continued to improve after this analysis). 15,780 autosomal CNVs from 293 patient samples (including 9 replicates) and 400,736 autosomal CNVs from 7,366 control samples passed QC. For pairs of replicated samples in the cases, the ones with greater level of noise in intensities were removed. The median number of CNVs called per sample (55 vs 55), the median size of CNVs (23.2kb vs 23.1kb) and the deletion-to-duplication (4.09 vs 4.08) ratio were comparable between cases and controls. A summary of call set statistics of cases and controls is presented in Table 3.1.

Table 3.1: Comparison of call set statistics between cases and controls

Cohort	Sample size	#CNV	Median #CNV per sample	Median CNV size (kb)	Deletion-to-duplication ratio	#CNVE	%Singleton
Case	284	15,323	55	23.2	4.09	2,143	63.0
Control	7,366	400,736	55	23.1	4.08	15,399	59.8

For the analysis of this data, I considered assessing three disease models: (*a*) common variants each with small effect, (*b*) a single rare variant with large effect and (*c*) multiple rare variants each with moderate effect. Model *a* has very limited power with such a small patient cohort. Therefore, the analysis was restricted to rare variants (model *b* and *c*).

The frequencies of CNVs were calculated by pooling case and control CNVs together and clustering pooled CNVs into CNVEs (see Chapter2 Methods, page 15).

'Rare' variants were defined as having a carrier frequency <1%. This left 14,645 rare CNVEs (clustered from 51,240 CNVs) out of the total 15,146 CNVEs (clustered from 416,300 CNVs). After filtering out rare CNVEs clustered exclusively from control CNVs, 1,858 CNVEs (clustered from 2,551 case CNVs and 19,764 control CNVs) were left. An overview of the genomic distribution of the CNVs belonging to these CNVEs is shown in Figure 3.4.

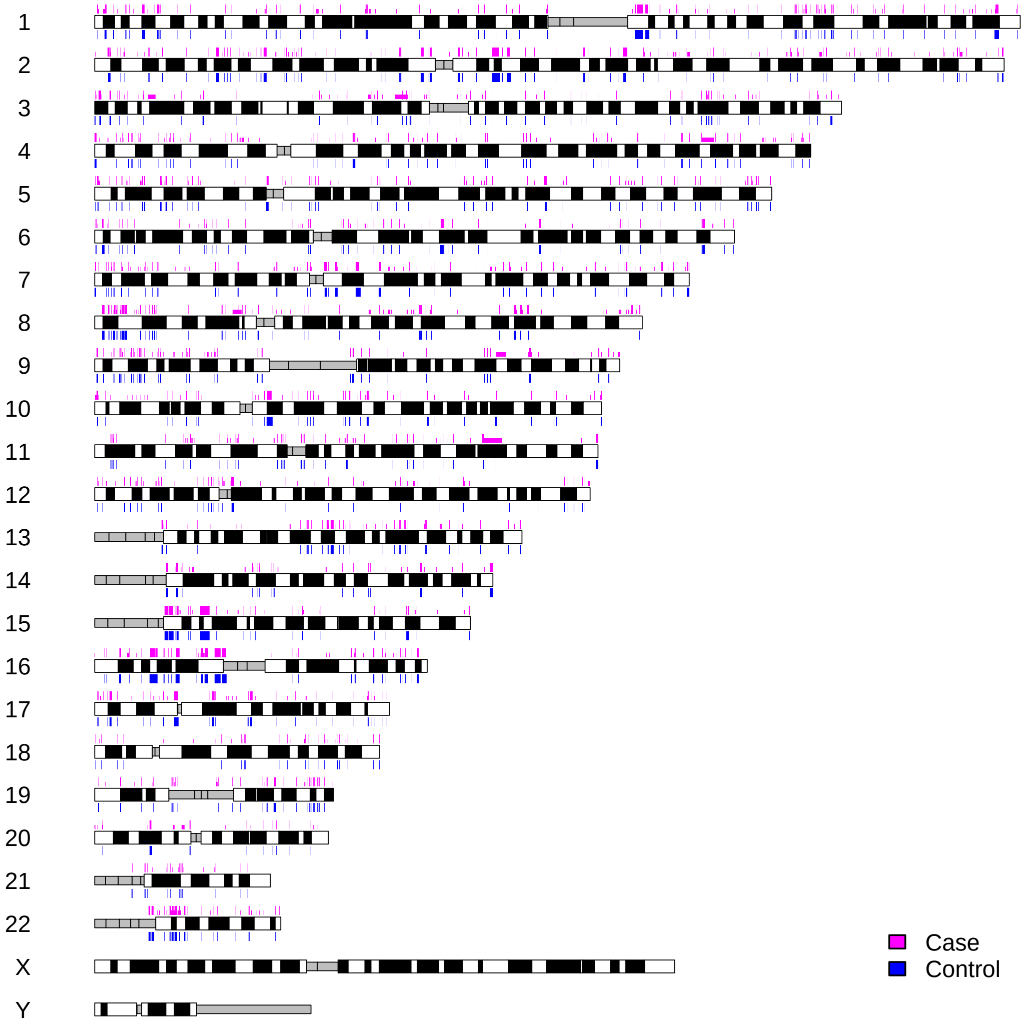

Figure 3.4: Overview of rare CNVs. The lengths of the colored rectangles represent the size of CNVs whereas the heights distinguish recurrent CNVs and singletons by which the former is taller than the latter. No CNV is displayed on chromosome X and Y since only autosomal calls were kept.

3.3.1.1 Specific loci associated with obesity

3.3.1.1.1 Genome-wide testing

Under the disease model in which a single rare variant has a large effect on the phenotype (model *b*), I investigated if there was any locus where rare CNVs were specifically found in cases or significantly enriched in cases. A CNVE-based approach was adopted (see Section 3.2.4). For each locus, the number of cases and controls that carry a CNV overlapping the CNVE >50% was used in a double-sided Fisher's Exact Test to assess the statistical significance of the enrichment. As deletions and duplication differ in their impact on genomic features and the ability to interpret their functional impact, they were treated separately. In total, 1,262 rare deletion CNVEs and 935 rare duplication CNVEs were subjected to the test of enrichment. 502 deletions corresponding to 396 CNVEs observed in 185 cases and 307 duplications corresponding to 256 CNVEs observed in 147 cases were found enriched in cases with a p value under 0.05. To correct for multiple hypothesis testing, I adopted the Bonferroni method, which maintains family-wise false positive rate under α by requiring each individual test to reach a significance level of α/n where n is the number of independent tests. 14 loci at which deletions or duplications were significantly enriched in cases relative to controls left after such correction with only four found overlapping genes (Table 3.2). The deletion at 4p15.31 is located in the first intron (1.1Mb) of some of the longer transcripts of *KCNIP4*, leaving duplications at 8q24.3 and deletions at 16p11.2 the only candidates that affect coding sequence.

Considering the rarity of many of the tested CNVs, the power to detect an association signal that reaches genome-wide significance is low. Therefore, an additional 9 genic CNVs that are case-specific and recurrent were collected (Table 3.3). Except for deletions at 3q28 and 10q11.23 which are intronic, the rest all affect coding sequence.

Table 3.2: Deletions and duplications significantly enriched in cases relative to controls

Loci	Start (kb)	End (kb)	Size (kb)	Type	#Case	#Control	P value	#Overlapped genes
3p12.2	83,228	83,401	173	dup	3	0	5.1×10^{-5}	0
4p15.31	20,981	20,986	5	del	6	1	1.7×10^{-8}	1*
5p11	46,197	46,314	117	del	4	2	2.6×10^{-5}	0
7p14.1	38,261	38,337	77	dup	3	0	5.1×10^{-5}	0
8q23.2–q23.3	112,106	112,213	107	dup	3	0	5.1×10^{-5}	0
8q24.3	143,422	143,656	234	dup	3	0	5.1×10^{-5}	2
10q21.1	54,598	54,611	14	del	7	11	2.0×10^{-6}	0
11q14.1	79,651	79,661	11	del	4	2	2.6×10^{-5}	0
11q14.1	80,550	80,557	7	del	4	0	1.9×10^{-6}	0
13q21.1	56,767	56,787	20	del	4	1	9.0×10^{-6}	0
13q21.31	62,157	62,402	245	dup	3	0	5.1×10^{-5}	0
16p11.2	28,616	28,951	336	del	5	2	1.3×10^{-6}	12
16p11.2	29,425	30,236	811	del	6	4	4.6×10^{-7}	38
21q21.2	23,351	23,356	5	del	5	4	7.6×10^{-6}	0

* Intronic

CNVs affecting coding sequence listed in Table 3.2 and Table 3.3 have been experimentally validated using multiplex ligation-dependent probe amplification performed by E. Bochukova at Metabolic Research Laboratories at Addenbrooke's Hospital. The functional relevance of the majority of them remains unclear at this stage.

3.3.1.1.2 Candidate gene testing

To complement the above association tests, I also used a candidate gene approach which might overcome a lack of power in a whole-genome association test setting. A list of 12 genes (*CRHR1*, *CRHR2*, *LEP*, *LEPR*, *MC3R*, *MC4R*, *MCHR1*, *MTCH2*,

Table 3.3: Case-specific recurrent genic deletions and duplications

Loci	Start (kb)	End (kb)	Size (kb)	Type	#Case	#Control	P value	#Overlapped genes
3p11.2	89,245	89,344	99	dup	2	0	1.4×10^{-3}	1
3q28	193,437	193,452	16	del	2	0	1.4×10^{-3}	1
6p12.1	52,875	52,892	17	del	2	0	1.4×10^{-3}	1
8q24.3	143,250	143,600	350	dup	2	0	1.4×10^{-3}	2
9q31.1	106,401	106,407	5	dup	2	0	1.4×10^{-3}	1
10p15.3	432	877	445	dup	2	0	1.4×10^{-3}	3
10q11.23	52,980	52,985	5	del	2	0	1.4×10^{-3}	1
11q13.4	71,980	72,107	126	dup	2	0	1.4×10^{-3}	2
22q13.33	49,246	49,349	103	dup	2	0	1.4×10^{-3}	10

NTRK2, *PCSK1*, *POMC* and *SIM1*) previously implicated in monogenic obesity was collected from the Human Obesity Gene Map [96] and a list of 8 genes (*BCDIN3D*, *BDNF*, *ETV5*, *FTO*, *GNPDA2*, *KCTD15*, *SH2B1* and *TMEM18*) with nearby SNPs associated with increased BMI was collected from literature [72]. The distributions of CNVs overlapping a 2Mb window based at each above genes were examined. No rare case CNV was found overlapping or near *CRHR1*, *CRHR2*, *LEP*, *MC3R*, *MCHR1*, *NTRK2*, *PCSK1*, *POMC*, *SIM1*, *BCDIN3D*, *BDNF*, *ETV5*, *FTO* and *TMEM18*. A 100kb duplication overlapping the first two exons of *LEPR* was found in one case and a 40kb duplication 237kb upstream of *GNPDA2* and 315kb away from the local peak of GWAS signal (rs10938397) was found in three cases, but they are likely to be irrelevant given their prevalence in controls. A 30kb duplication in the last intron of *CHST8*, 48kb away from *KCTD15* and 82kb away from the local peak of GWAS signal (rs11084753) was found in two cases and three controls with a test p-value of 0.013. A 235kb duplication overlapping the first exon of *PTPRJ* and 185kb away from *MTCH2* was found in one case and is partially (24–29%) overlapped by duplications found in two controls. A 153kb deletion 60kb downstream of *MC4R* and overlapping the local peak of GWAS signal (rs17782313) was found in one case

and is marginally (4–11%) overlapped by deletions found in three controls. Deletions of variable length with a minimal overlapping region of 250kb all encompassing *SH2B1* and the local peak of GWAS signal (rs7498665) were found in five cases and the minimal overlapping region was found deleted in two controls with a test p-value of 1.3×10^{-6}, which was also highlighted by the genome-wide testing approach.

3.3.1.1.3 16p11.2 deletion encompassing *SH2B1*

Both of the above approaches pointed to the heterozygous deletions at 16p11.2 encompassing *SH2B1* found in five unrelated cases out of 284 and two controls out of 7,366. Closer inspection reveals that the deletions fall into two classes: a shorter form of 220kb (28.73–28.95 Mb) and a longer form of $\sim$1.7Mb (28.4–30.1 Mb). The breakpoints of both classes of deletion are embedded within complex, segmentally duplicated regions of 16p11.2 containing directly-oriented, highly-similar ($>$98% sequence similarity) duplicated sequences greater than 15kb in length (Figure 3.5). This observation strongly supports the hypothesis that these deletions arise through non-allelic homologous recombination (NAHR) between duplicated sequences.

Our collaborator E. Bochukova and S. Farooqi at Metabolic Research Laboratories at Addenbrooke's Hospital generated additional genotype and phenotype data on these five families. The shorter 220kb deletion was seen in three patients with severe early onset obesity alone and was inherited from their respective obese parents. The longer $\sim$1.7Mb deletion, which encompasses the 220kb deletion and extends through a 593kb region (29.5–30.1 Mb) where deletions are associated with autism and mental retardation, occurred *de novo*. The two carrying patients had mild developmental delay in addition to their severe obesity. These findings are consistent with a role for the *SH2B1*-containing 220kb region (28.73–28.95 Mb) in severe obesity and the 29.5–30.1 Mb region in brain development. Recently, the 29.5–30.1 Mb region has been discovered to independently associate with obesity in addition to autism and mental retardation [97].

Further experiments undertaken by S. Farooqi *et al* revealed a striking similarity of the phenotype of the patients with the *SH2B1*-containing deletion with human leptin receptor deficient phenotype. Since *SH2B1* encodes an adaptor protein for several members of the tyrosine kinase receptor family including ones involved in leptin and insulin signaling and heterozygous knock-out of *Sh2B1* in mice leads to obesity on a high fat diet, haploinsufficiency of *SH2B1* may be a plausible mechanism underlying the phenotype seen in these patients.

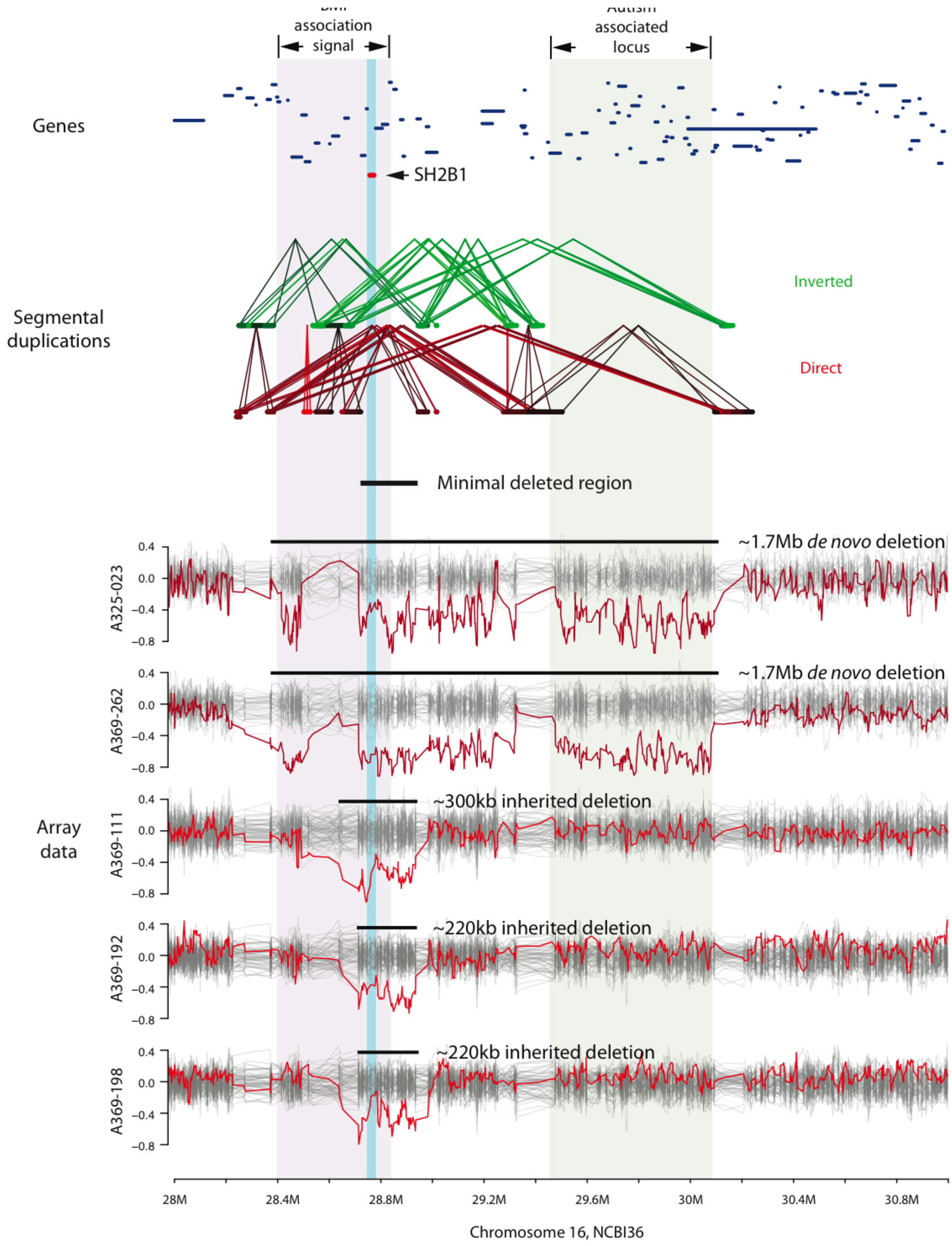

Figure 3.5: Deletions at 16p11.2 overlapping *SH2B1*. Affymetrix 6.0 array data for five patients with deletions at 16p11.2 is shown. Log_2 ratios of the five samples are highlighted in dark red with other samples in the same genotyping plate shown in grey. Annotation of the segmental duplications was taken from the UCSC genome browser and the darkness of color coding represents sequence similarity between the duplicated pairs. Protein-coding genes are represented by dark blue lines; *SH2B1* is highlighted in red and by blue vertical shading. The light pink vertical shading indicates the range of a previous BMI association signal found in two genome wide association studies and the light grey vertical shading indicates the reported autism associated CNV region.

3.3.1.2 Global CNV burden

Despite discovering only a couple of loci at which the locus-specific enrichment of rare CNVs in cases relative to controls reached statistical significance, the finding of many case-specific CNVs and rare CNVs with higher case prevalence might still indicate their contribution to the phenotype that could not be detected individually, but might be detected collectively as a 'burden' of CNVs. Previous study reported increased burden of large (>100kb) and rare (<1%) CNVs in patients with Schizophrenia [11]. Following the same criteria, I explored if there was increased burden of large rare CNVs in patients with severe-early onset obesity relative to controls. To control for the subtle differences in data quality that might lead to differential sensitivity and specificity of CNV calling between cases and controls (Figure 3.6), I used a permutation-based method (see Section 3.2.2) to assess the statistical significance of global burden.

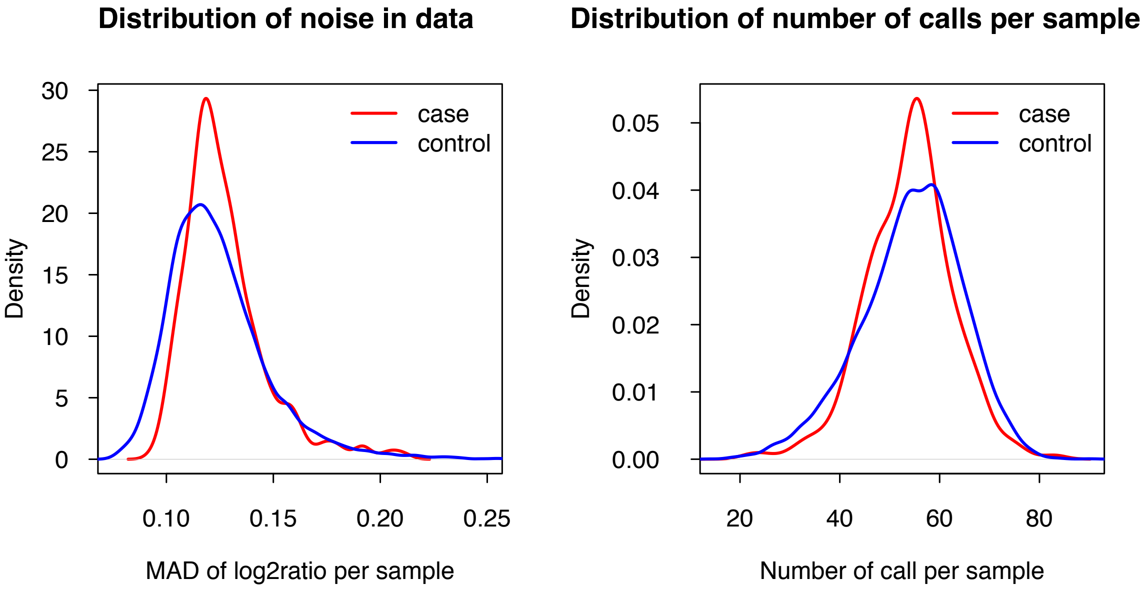

Figure 3.6: Comparison of level of noise and number of CNV calls per sample between cases and controls. Case samples have higher level of noise than controls ($p = 1.2 \times 10^{-3}$, Mann-Whitney test), leading to slightly greater number of calls per sample, though such difference is insignificant (p=0.11).

Since many of the obese patients also had developmental delay and given the previous observation that increased CNV burden is association with neurodevelopmental disorders, I investigated whether the observed increased CNV burden in cases relative to controls was driven by inclusion of patients with developmental delay by performing the analysis separately on the group of patients with obesity and developmental delay, and on the group of patients with obesity but without developmental delay. Collectively, the entire set of cases exhibit a two-fold enrichment of >500kb rare deletions compared to controls ($p = 5 \times 10^{-4}$, Fisher's exact test). A stronger three-fold enrichment is observed in cases with developmental delay in addition to severe early onset obesity ($p = 3 \times 10^{-4}$), whereas the 1.3 fold enrichment in cases with severe early onset obesity alone is not significant (p = 0.24) (Table 3.4).

Table 3.4: Global CNV burden analysis: case enrichment of >500kb rare CNVs

Samples	Type	Case rate	Case/control ratio	P_{MAD}*	P_{NCPS}†
All	Losses and gains	0.2500	1.2996	0.0201	0.0433
	Losses	0.1127	2.0906	0.0005	0.0007
	Gains	0.1373	0.9917	0.4776	0.5800
Severe early-onset obesity only	Losses and gains	0.2089	1.0857	0.3150	0.3905
	Losses	0.0696	1.2917	0.2389	0.2884
	Gains	0.1392	1.0055	0.4790	0.5332
Severe early-onset obesity and developmental delay	Losses and gains	0.2937	1.5417	0.0098	0.0195
	Losses	0.1667	3.1318	0.0003	0.0001
	Gains	0.1270	0.9252	0.5701	0.6801

* Derived from permutation conditioned on MAD of sample $\log_2$ ratio

† Derived from permutation conditioned on number of calls per sample

A more detailed analysis by type, frequency and sizes for rare CNVs $>$100kb yields the following observations: (*i*) a significant 1.1-fold enrichment of rare CNVs $>$100kb is seen in all cases collectively; (*ii*) cases with developmental delay in addition to obesity generally exhibit heavier CNV burden than patients with obesity alone; (*iii*) case enrichment of singleton CNVs is generally stronger compared to recurrent rare CNVs; (*iv*) case enrichment of deletions is generally stronger in larger events ($>$500kb) but the trend seems reversed for duplications of which enrichment of smaller events (100–200kb) is stronger (Table 3.5 & 3.6).

Table 3.5: Global CNV burden analysis of >100kb rare CNVs: event type and frequency

All cases

Type	Frequency	Case rate	Case/control ratio	P_{MAD}	P_{NCPS}
Losses and gains	All <1%	1.9225	1.1297	0.0015	0.0119
	Single occurrence	0.5035	1.5966	0.0000	0.0000
	Recurrent <0.1%	0.5775	1.3901	0.0001	0.0007
Losses	All <1%	0.7430	1.1085	0.0357	0.0877
	Single occurrence	0.1796	1.7246	0.0002	0.0011
	Recurrent <0.1%	0.1937	1.2426	0.0399	0.0778
Gains	All <1%	1.1796	1.1434	0.0055	0.0344
	Single occurrence	0.3239	1.5335	0.0008	0.0084
	Recurrent <0.1%	0.3838	1.4786	0.0004	0.0014
Severe early-onset obesity only					
Losses and gains	All <1%	1.8861	1.0965	0.0352	0.0989
	Single occurrence	0.4937	1.5487	0.0022	0.0069
	Recurrent <0.1%	0.5000	1.2095	0.0396	0.0913
Losses	All <1%	0.7595	1.1284	0.0674	0.1215
	Single occurrence	0.1519	1.4437	0.0646	0.0647
	Recurrent <0.1%	0.2342	1.4909	0.0090	0.0178
Gains	All <1%	1.1266	1.0760	0.1203	0.2298
	Single occurrence	0.3418	1.6004	0.0068	0.0222
	Recurrent <0.1%	0.2658	1.0371	0.3402	0.4749
Severe early-onset obesity and developmental delay					
Losses and gains	All <1%	1.9921	1.1649	0.0043	0.0238
	Single occurrence	0.6032	1.8971	0.0000	0.0002
	Recurrent <0.1%	0.5556	1.3400	0.0062	0.0250
Losses	All <1%	0.7381	1.0997	0.1280	0.2070
	Single occurrence	0.2143	2.0341	0.0012	0.0017
	Recurrent <0.1%	0.1429	0.9182	0.5956	0.6518
Gains	All <1%	1.2540	1.2071	0.0078	0.0328
	Single occurrence	0.3889	1.8292	0.0016	0.0070
	Recurrent <0.1%	0.4127	1.5933	0.0016	0.0062

Table 3.6: Global CNV burden analysis of >100kb rare CNVs: event type and size

All cases

Type	Size (kb)	Case rate	Case/control ratio	P_{MAD}	P_{NCPS}
Losses and gains	100–200	1.1162	1.2279	0.0000	0.0011
	200–500	0.5563	0.9265	0.6769	0.8509
	>500	0.2500	1.2996	0.0201	0.0433
Losses	100–200	0.4190	1.0420	0.1942	0.2713
	200–500	0.2113	0.9862	0.4591	0.5933
	>500	0.1127	2.0906	0.0005	0.0007
Gains	100–200	0.6972	1.3753	0.0000	0.0003
	200–500	0.3451	0.8934	0.7456	0.8721
	>500	0.1373	0.9917	0.4776	0.5800
Severe early-onset obesity only					
Losses and gains	100–200	1.1013	1.2005	0.0077	0.0187
	200–500	0.5759	0.9436	0.6052	0.7298
	>500	0.2089	1.0857	0.3150	0.3905
Losses	100–200	0.4241	1.0471	0.2615	0.3234
	200–500	0.2658	1.2408	0.0861	0.1226
	>500	0.0696	1.2917	0.2389	0.2884
Gains	100–200	0.6772	1.3218	0.0049	0.0133
	200–500	0.3101	0.7829	0.9228	0.9547
	>500	0.1392	1.0055	0.4790	0.5332
Severe early-onset obesity and developmental delay					
Losses and gains	100–200	1.1587	1.2570	0.0007	0.0069
	200–500	0.5397	0.9029	0.6792	0.8193
	>500	0.2937	1.5417	0.0098	0.0195
Losses	100–200	0.4286	1.0601	0.2395	0.3048
	200–500	0.1429	0.6685	0.9478	0.9688
	>500	0.1667	3.1318	0.0003	0.0001
Gains	100–200	0.7302	1.4109	0.0007	0.0039
	200–500	0.3968	1.0332	0.3134	0.4547
	>500	0.1270	0.9252	0.5701	0.6801

3.3.2 Analysis of 1,500 patient samples

Affy6 .CEL files of patient samples belonging to the three subsets (SCOOP1, 2 & 3) were processed together using the pipeline described in Chapter 2, however, with Canary calls (known common CNV genotyping calls) included, as common CNVs were to be interrogated in the analyses. To maintain consistency with a SNP GWAS analysis of these Affy6 data (SCOOP1, 2 & 3) conducted in parallel, only WTCCC2 controls were used for this part of the analysis and samples of patients with developmental delay or with self-reported ethnicity other than European were removed. Replicate samples in the patient CNV set were also removed by excluding the replicate with greater level of noise in array intensities. This left 135,123 CNVs from 1,167 patient samples and 693,468 CNVs from 5,899 control samples.

3.3.2.1 Identification of population ancestry outliers

As population stratification is a well-known factor that can cause spurious association in GWAS [98], I first examined the population structure of the case and control cohorts using MDS (see Section 3.2.3). As expected, the majority of both cases (SCOOP1,2,3) and controls (WTCCC2) are concentrated around the European ancestry reference population (CEU) in the projected space. However, a higher proportion of cases than controls apparently have more diverse population ancestry. Using three alternative arbitrary thresholds on genetic distance to CEU with decreasing stringency (distance to CEU = 5, 10 and 30), 6.4%, 4.8% and 1.3% of cases are regarded as non-European whereas the proportion of controls are only 0.54%, 0.27% and 0.08% (Figure 3.7). The most permissive threshold (distance to CEU = 30) was adopted to only remove cases and controls that are extremely remote in ethnicity relative to Europeans, given that (*i*) all samples have gone through stringent sample QC, (*ii*) systematic inflation in test statistics was very minor even with all samples included (data not shown). This process left 132,839 CNVs from 1,152 cases and 692,256 CNVs from 5,894 controls that entered subsequent analyses of both common and rare CNVs. The summary characteristics of the case and control call set are comparable (Table 3.7).

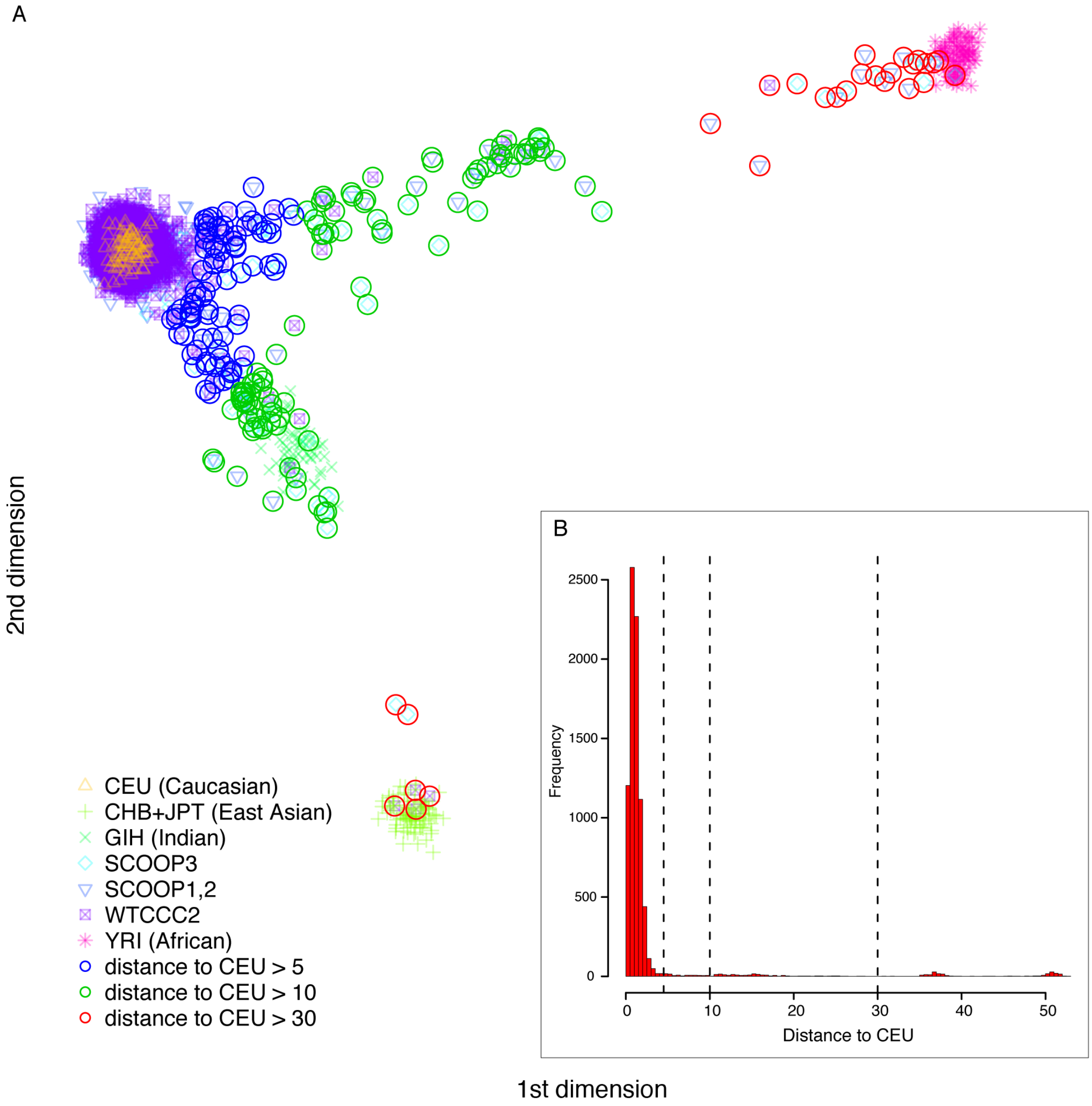

Figure 3.7: Identifying ethnic outliers based on SNP genotypes and MDS projection. (A) Each small symbol represents a sample. The European, East Asian and African populations are well separated and serve as reference points for samples of unknown ethnicity. As a positive control, the Indian population is located approximately at the mid point of the European-East Asian axis, which is consistent with its ethnic and geographical relationship with the two reference populations. All cases and controls, including those failed sample QC or removed by various filters, are displayed. Red, green and blue circles highlight samples regarded as non-European under thresholds of different stringency, shown as dashed lines in (B), the distribution of the distance between the European reference population and all samples (including the reference populations).

Table 3.7: Call set statistics of cases and controls (Birdseye + Canary calls)

Cohort	Sample size	#CNV	Median #CNV per sample	Median CNV size (kb)	Deletion-to-duplication ratio	#CNVE	%Singleton
Case	1,152	132,839	115	14.6	4.09	5,101	62.0
Control	5,894	692,256	117	14.6	3.93	12,568	61.4

3.3.2.2 Common CNV analysis

With the much larger sample size of cases in this second analysis, I first explored if there were any common CNVs associated with the phenotype.

Similar to Section 3.3.1, the approximate population frequency of CNVs were calculated by pooling case and control CNVs together and clustering pooled CNVs into CNVEs [chapter2 method]. 'Common' CNVEs were defined as having a population frequency >1%. This yields 587 common CNVEs (clustered from 775,102 CNVs) out of the total 14,654 CNVEs (clustered from 825,095 CNVs).

Test of CNV-phenotype association can be done either directly using the quantitative measure of copy number, or the integer copy number reflecting the CNV genotype, or indirectly through the genotypes of tagging markers that are highly correlated with the CNV genotypes. As a perfectly correlated SNP could not be found on Affy6 for every common CNVE and the total number of common CNVEs was not prohibitively large, the first approach was adopted. For each common CNVE, I performed a likelihood ratio test for association that models the distribution of quantitative CNV measurements as Gaussian mixtures and controls for potential differential biases between cases and controls, as implemented in the CNVtools package. Due to the complexity and heterogeneity of the measurements of CNVs, the test was repeated 27 times under different combinations of settings, from which the most appropriate result was manually selected (see Section 3.2.5).

Out of 587 common CNVEs, 416 could be tested for association under at least one of the 27 automated settings. After manually curating the clustering, test results could be recovered for another 65 common CNVEs, making a total of 481 testable common CNVEs. Similar to frequently seen SNP GWAS results, the p values of tests at the vast majority of loci approximately followed the distribution expected under the null hypothesis that no association is found. There could be some minor confounding factors (inflation factor $\lambda = 1.03$), such as residual differences in population ancestry, but the effect is very minor and the slight increase of type I error rate is unlikely to affect the very top candidates (Figure 3.8A).

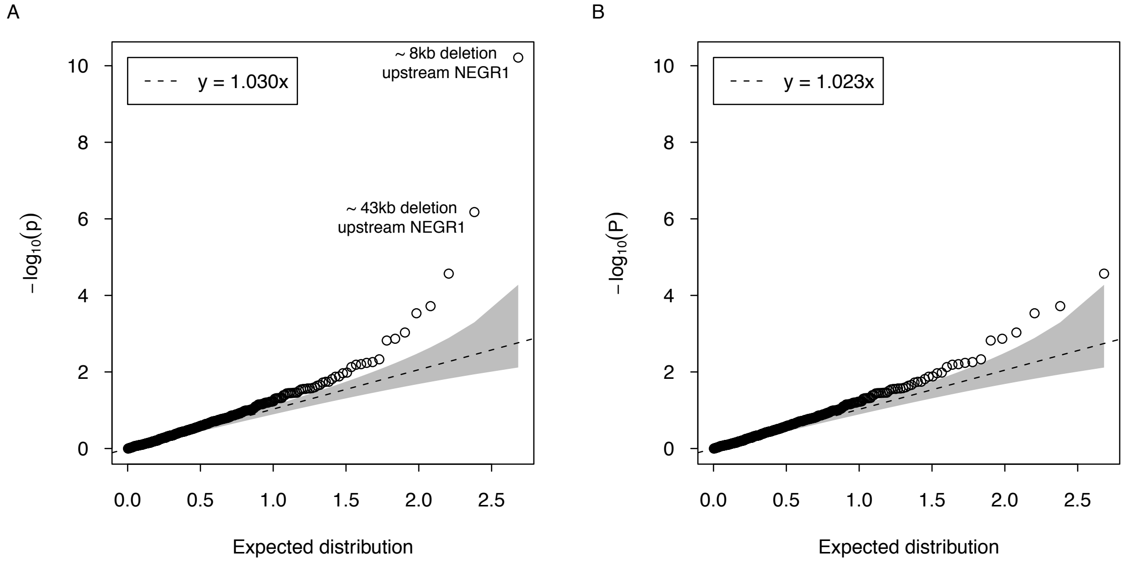

Figure 3.8: Genome-wide association results for common CNVs. (A) Quantile-quantile plot of -$\log_{10}$(p) of all 481 common CNVs. Concentration band represents 95% confidence interval. Inflation factor is represented as the slope of the fitted line. (B) Quantile-quantile plot after removal of the two CNVs upstream of *NEGR1*.

The most and only significant associations came from two deletions upstream of *NEGR1*: a smaller ~8kb deletion (72,528–72,536kb) with inversed association (p = 6.1×10^{-11}) and a larger ~43kb deletion (72,541–72,584kb) with positive association (p = 6.6×10^{-7}) (Figure 3.8A). No other convincing association was observed after their removal (Figure 3.8B).

Both deletions were described in a previous GWAS of BMI in which the larger deletion, but not the smaller one, was reported to associate with increased BMI with a p value of 9.3×10^{-6} by testing using a perfect tagging SNP [72]. The same study also found that the two deletions segregate at the locus on distinct haplotypes in the three HapMap populations, resulting in three alleles: one represented by the reference sequence (denoted as normal), one with the smaller deletion and the one with the larger deletion. The genotypes of the two deletions observed in this study verified this finding (Table 3.8).

Table 3.8: Co-presence of the genotypes of the two deletions

Cases

		Copy number at ∼43kb deletion locus		
		0	1	2
Copy number at ∼8kb deletion locus	0	0	0	19
	1	1	204	67
	2	508	299	54

Controls

		Copy number at ∼43kb deletion locus		
		0	1	2
Copy number at ∼8kb deletion locus	0	0	0	220
	1	4	1360	444
	2	2160	1441	265

As the three alleles are mutually exclusive, the question arises as to whether the two deletion alleles are independently associated with severe, early onset obesity. The frequency of the undeleted allele is approximately the same in cases and controls and is expectedly not associated with the phenotype (OR = 1; 95% CI 0.90–1.1; p = 0.93, two-sided Fisher's exact test). Therefore, a conditional analysis was performed for the larger and the smaller deletion alleles, respectively, by testing the association of one allele conditioned on the genotype of the other. When conditioned on the smaller deletion allele, the association of the larger deletion allele becomes insignificant (OR = 1.09; 95% CI 0.97–1.22; p = 0.16). When conditional on the larger allele, the association of the smaller deletion remains significant (OR = 0.70; 95% CI 0.60–0.82; p = 6.93×10^{-6}). This suggests that the association in this region is largely driven by the protective effect of the ∼8kb deletion allele. A replication study using the Sequenom platform is being undertaken by our collaborators. In this replication experiment the tagging SNPs of the two NEGR1 deletions are being genotyped, along with other putative association signals from the SNP GWAS analysis in large, independent obese and control cohorts.

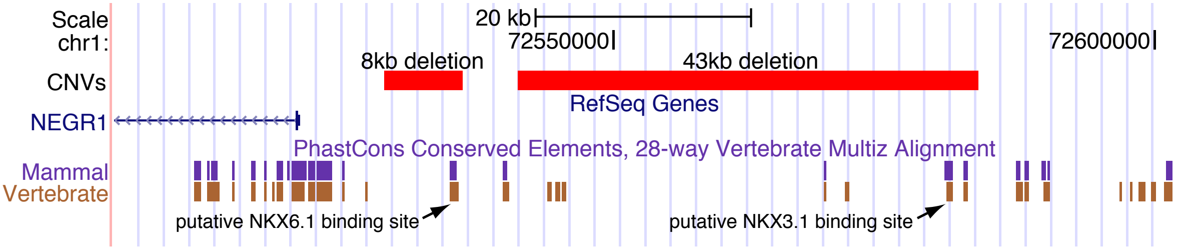
Scale
chr1:
20 kb
72550000
72600000
CNVs
8kb deletion
43kb deletion
RefSeq Genes
NEGR1
PhastCons Conserved Elements, 28-way Vertebrate Multiz Alignment
Mammal
Vertebrate
putative NKX6.1 binding site
putative NKX3.1 binding site

Figure 3.9: The two associated common deletions upstream of *NEGR1*. Plot taken from the UCSC genome browser with deletions denoted by red bars and putative transcription factor binding sites pointed by arrows.

Although the two deletions do not overlap coding sequence, they encompass a few conserved noncoding elements, including binding sites of transcription factor *NKX3.1* (∼43kb deletion) and *NKX6.1* (∼8kb deletion) (Figure 3.9). *NKX6.1* can act as both a potent transcription repressor and a potent transcription activator [99], and is required for the development of pancreatic beta cell [100]. *NKX3.1* is a putative prostate tumor suppressor that is expressed in a largely prostate-specific and androgen-regulated manner [101]. If *NKX3.1* has a trans-regulatory role in the association between the deletions and obesity, given its male specificity, one might expect bias in sex in the association. Indeed, by performing the conditional association analysis in males and females separately, a marginally significant association of the ∼43kb deletion allele was observed in males (OR = 1.21; 95% CI 1.04–1.42; p = 0.012) but not in females (OR = 1; 95% CI 0.86–1.17; p = 1), whereas for the ∼8kb deletion allele, no association was observed in males (OR = 0.81; 95% CI 0.64–1.03; p = 0.087) but the association signal observed in females was very strong (OR = 0.61; 95% CI 0.49–0.75; p = 2.1×10^{-6}) and much stronger than that of the ∼43kb deletion allele in males (see Discussion).

Table 3.9: Allele frequency of the three alleles at 72,528–72,584kb

	Male			Female			Total		
	8kb deletion	43kb deletion	Normal	8kb deletion	43kb deletion	Normal	8kb deletion	43kb deletion	Normal
Case	0.139 (134)*	0.657 (632)	0.204 (196)	0.131 (176)	0.662 (889)	0.206 (277)	0.135 (310)	0.660 (1521)	0.205 (473)
Control	0.182 (1083)	0.599 (3560)	0.218 (1297)	0.199 (1165)	0.610 (3569)	0.190 (1114)	0.191 (2248)	0.605 (7129)	0.205 (2411)

* Numbers in parentheses are counts

3.3.2.3 Rare CNV analysis

Rare CNVs were analyzed with the same strategies described in Section 3.3.1: to identify specific associated loci and assess global CNV burden.

3.3.2.3.1 Specific loci associated with obesity

Genome-wide testing

3,013 rare deletion CNVEs and 2,814 rare duplication CNVEs were subjected to the test of locus-specific enrichment. 462 deletions corresponding to 201 CNVEs observed in 313 cases and 418 duplications corresponding to 180 CNVEs observed in 281 cases were found enriched in cases with a test p value under 0.05. After correcting for multiple hypothesis testing, none of deletion CNVEs and only two duplication CNVEs remained statistically significant. The two duplication CNVEs mapped to regions that encode the variable part of the alpha and gamma chain of T cell receptor, which are likely false associations. CNVs at some of the case-enriched loci previously identified in the earlier analysis of the SCOOP3 samples (Table 3.2) were found in additional cases, such as ones at 5p11, 11q14.1, 16p11.2 and 21q21.2, but failed to reach statistical significance, possibly due to (*i*) differences in case phenotype (different case ascertainment with respect to developmental delay), and (*ii*) inadequate power caused by sharply increased number of tests.

Due to the lack of significant associations, rare case-recurrent genic deletions with a test p value <0.05 and control occurrence ≤ 5 were collected to enrich for pathogenic variants. After manual examination of $\log_2$ ratio profiles, 16 deletions were kept (Table 3.10). Although some of them have been reported to express in brain, their functional relevance remains unclear at this stage.

Table 3.10: Case-enriched recurrent deletions

Loci	Start (kb)	End (kb)	Size (kb)	Type	#Case	#Control	P value	#Overlapped genes
1p21.3	97,934	98,028	94	del	3	0	4.4×10^{-3}	1
2q21.2	133,589	133,691	102	del	2	0	2.7×10^{-2}	1
3p22.1	40,393	40,423	30	del	2	0	2.7×10^{-2}	1
4p12	48,427	48,460	33	del	4	2	8.1×10^{-3}	1
4q24	106,679	106,721	42	del	3	2	3.4×10^{-2}	1
5p13.2	37,497	37,558	61	del	2	0	2.7×10^{-2}	1
6q25.1	150,962	150,982	20	del	3	2	3.4×10^{-2}	1
9p24.2	2,224	2,354	130	del	3	1	1.5×10^{-2}	1*
9p22.2	17,801	17,890	89	del	2	0	2.7×10^{-2}	1*
10q21.3	70,283	70,292	9	del	2	0	2.7×10^{-2}	1
10q21.3	71,013	71,038	25	del	2	0	2.7×10^{-2}	1*
11q22.3	150,962	150,982	20	del	3	3	3.4×10^{-2}	1*
16p12.1	21,725	22,350	625	del	4	5	4.5×10^{-2}	10
16p11.2	28,731	28,951	220	del	5	2	1.8×10^{-3}	10
17p13.2	4,838	4,901	62	del	2	0	2.7×10^{-2}	3
22q11.22	21,328	21,977	649	del	2	0	2.7×10^{-2}	7

* Intronic

Candidate gene testing

No additional rare CNVs emerged from the examination of the genomic windows encompassing and flanking the candidate genes (described above) that are implicated in monogenic forms of obesity or previous discovered GWAS signals associated with BMI.

3.3.2.3.2 Global CNV burden

Previous analysis of the smaller SCOOP3 patient cohort revealed an insignificant enrichment of large rare CNVs in patients with obesity alone (Section 3.3.1.2). With a much large patient cohort and consequently greater power, I investigated if such enrichment exists with the statistical significance assessed using the some permutation method. A significant 1.16 fold enrichment was observed for all CNVs >500kb in size and <1% in frequency. This fold of enrichment is lower than that previously observed in patients with both obesity and developmental delay (1.54 fold) and largely consistent with that observed in patients with obesity alone (1.09 fold). The fold of enrichment in >500kb and <1% deletions is slightly higher than previously observed (1.44 vs 1.29) but still far below that observed in patients with additional developmental delay (3.13). A few previous observations were replicated: (*i*) the enrichment of singleton CNVs is stronger than that of rare recurrent CNVs; (*ii*) the enrichment is stronger in larger events (>500kb) for deletions, but the trend is reversed for duplications; (*iii*) the enrichment of deletions is stronger compared to that of duplications in the range of >500kb, but trend is reversed in the range of 100–200kb. The most significant enrichment is observed in duplications in the range of 100–200kb, which is also consistent with previous observation. Although many tests would lose statistical significance or become only marginally significant after multiple test correction, the consistent observations and the increase in statistical significance suggest the 1.1–1.5 fold increase in CNV burden in patients with severe early onset obesity alone is real (Table 3.11 & 3.12).

Table 3.11: Global CNV burden analysis of >100kb rare CNVs: event type and frequency

Type	Frequency	Case rate	Case/control ratio	P_{MAD}	P_{NCPS}
Losses and gains	All <1%	1.7439	1.1419	0.0000	0.0000
	Single occurrence	0.4002	1.4048	0.0000	0.0000
	Recurrent <0.1%	0.4627	1.1659	0.0017	0.0002
Losses	All <1%	0.5972	1.0983	0.0123	0.0035
	Single occurrence	0.1259	1.2574	0.0072	0.0070
	Recurrent <0.1%	0.1710	1.2085	0.0094	0.0043
Gains	All <1%	1.1467	1.1661	0.0000	0.0000
	Single occurrence	0.2743	1.4846	0.0000	0.0000
	Recurrent <0.1%	0.2917	1.1423	0.0317	0.0080

Table 3.12: Global CNV burden analysis of >100kb rare CNVs: event type and size

Type	Size (kb)	Case rate	Case/control ratio	P_{MAD}	P_{NCPS}
Losses and gains	100–200	0.9731	1.1984	0.0000	0.0000
	200–500	0.5720	1.0504	0.1104	0.0329
	>500	0.1988	1.1658	0.0205	0.0057
Losses	100–200	0.3733	1.1122	0.0165	0.0082
	200–500	0.1762	1.0074	0.4438	0.3395
	>500	0.0477	1.4357	0.0117	0.0076
Gains	100–200	0.5998	1.2590	0.0000	0.0000
	200–500	0.3958	1.0707	0.0953	0.0339
	>500	0.1510	1.1004	0.1446	0.0653

3.4 Discussion

In this chapter, I described the analysis of copy number variants in patients with severe early onset obesity. Under a case-control framework, the role of common CNVs, rare CNVs at specific loci and global burden of rare CNVs were examined. In the initial study of $\sim$300 patients enriched with additional developmental delay and syndromic forms of obesity, I observed a significant two-fold enrichment of $>$500kb and $<$1% deletions in all cases, a stronger three-fold enrichment in cases with both developmental delay and severe early onset obesity, and a insignificant 1.29-fold enrichment in cases with severe early onset obesity only. A heterozygous $\sim$220kb deletion at 16p11.2 encompassing the gene *SH2B1* is identified by both genome-wide and candidate gene approach as a pathogenic variant for the five patients in which the deletion was found, with haploinsufficiency of *SH2B1*, a gene involved in leptin and insulin signaling, being a very likely cause. In the following study of $\sim$1200 patients with severe early onset obesity only, a significant 1.44-fold enrichment of $>$500kb and $<$1% deletions was observed, suggesting that there exists a significant burden of large rare CNVs in patients with obesity alone albeit being weaker than that observed in patients with co-present developmental delay. In the common CNV analysis, a previously reported $\sim$430kb common deletion and an adjacent $\sim$8kb common deletion, both upstream of the gene NEGR1, were found associated with the phenotype. Conditional analysis revealed the $\sim$8kb deletion explains most of the association signal and has a strong sex bias in effect size.

Compared to previous large-scale genome wide association studies of obesity as a common quantitative trait, the two patient cohorts studied here are relatively small, but the patients' phenotype were carefully selected to represent the extremes on the scale of severity. The first patient cohort was intentionally enriched for patients with developmental delay in addition to severe obesity, for the investigations of rare CNVs. This is under the expectation that rare variants each imposing a relatively large effect and leading to a more severe phenotype might account for some of the heritability missed by the common variant model. This study design has proven to be effective at least in this case. The most significant finding of this study, the *SH2B1*-containing deletion actually overlaps a previously reported GWAS signal. The co-presence of both common variants influencing susceptibility to common obesity and more highly penetrant rare CNVs associated with severe early onset form of the disease not only suggests a link in etiology between the two, but also suggest that looking for rare variants near common susceptibility loci may prove to be a fruitful strategy for other common complex disease. Studies of other phenotypes have similarly observed overlap between genes identified using monogenic and GWAS approaches, for example, lipid traits.

In addition to deletions, heterozygous duplications were found at the ~220k minimal overlapping window encompassing *SH2B1* in 9 out of 7,366 controls but none out of 1,309 cases (combining data from both studies). Although this is not a significant observation, it may still hint that extra copies of this part of the genome might be protective against severe early onset obesity. This mirroring of BMI phenotype with dosage of the genomic interval has also been observed at the nearby ~593kb locus in 16p11.2 (29.5–30.1Mb) [102].

The ~593kb 16p11.2 deletion (29.5–30.1Mb) previously associated with autism and mental retardation has recently been suggested to have a causal role in a highly penetrant form of obesity [97]. The deletion was found with significantly higher frequency in cases (9 out of 1,309) relative to controls (4 out of 7,366) in the cohorts here studied. However, 6 of the 9 patients carrying this deletion exhibit development delay or autistic behavior, out of which two also carry the ~220kb *SH2B1*-containing deletion. If removing all cases with developmental delay, the deletion was left in 3 out of 1,152 cases, making it on the verge of (in)significance ($p = 0.057$). Although this does not simply imply a rejection of the role of this deletion in obesity, the established involvement of this deletion in autism and mental retardation does require a more specific study design, such as recruiting non-obese controls with neurodevelopmental phenotypes matching those of cases, to allow disentangling its contribution to obesity.

With a genome-wide association test approach, 652 out of 2,197 rare CNVEs tested in the initial study and 381 out of 5,827 rare CNVEs tested in the second study were found enriched in cases with a p value <0.05. However, the vast majority of these loci did not reach genome wide significance as determined by Bonferroni correction.

The number of significant association signals is even smaller in the second study despite the larger patient cohort. This could be due to the heterogeneity between patients with and without additional developmental delay and the complexity of the genetics of obesity. It may also be due to a drop of power in the second study as: (*i*) it excluded patients with developmental delay, which are enriched for rare CNVs, and (*ii*) the major impact of adding more obesity-only samples was not increased occurrences of existing rare CNVEs, which boosts power as in the case of common CNVEs, but adding a large number of private and extremely rare CNVs at addition loci. Indeed, only 16.8% of the rare CNVEs shared by both studies had increased case occurrences in the second study, whereas 79.5% of the additional CNVEs with at least one case occurrence introduced by the second study were found in that case alone. With a four times larger patient cohort and 2.6 times more tests, the power drops both for individual tests to reach nominal significance and for those passing nominal significance threshold to reach genome-wide significance. This result demonstrates the challenge unique to rare variant studies, wherein increasing sample size is likely to be accompanied with increasing number of tests, which may leads to diminishing returns or even decrease of statistical power, as opposed to common variant studies, wherein increasing sample size is always beneficial as the number of tests is largely unchanged.

The issue of power is linked with the choice of the unit of test and the method of multiple test correction. In this study, each CNVE was chosen as a unit of test and Bonferroni correction was applied under the assumption that all tests are independent. Alternative units of test could be probes or genes (Figure 3.2), each having its advantages and disadvantages. Testing on probes requires no pre-testing procedures such as collapsing CNV calls into CNVEs and the number of tests is fixed regardless of sample size. However, it leads to greatest number of tests of which many are perfectly correlated due to being in the same CNV. I observed that false associations also frequently arise at the border of common CNV calls. Testing on genes does not collapse CNV calls into CNVEs but bins them by genes or genomic windows including certain length of flanking regions of genes, which increases power for small and rare CNVs affecting the same gene. The number of tests is fixed. The number of perfectly correlated tests is reduced compared to testing on probes but still exists when multiple genes are affected by the same large CNV. Spurious association is also likely to arise at the border of common CNV calls. Though current functional studies usually choose to follow genic CNVs, completely ignoring the large proportion of non-genic CNVs still seems undesirable or at least inefficient. Testing on CNVEs, as I did in this study, avoids perfectly correlated tests within the same CNV and spurious associations emerging from the border of common CNV calls. However, it requires the complex pre-step of collapsing CNV calls into CNVEs, which itself is not perfect such as the use of arbitrary call-overlap thresholds. Correlation between tests, though greatly reduced by avoiding multiple tests within the same CNV, still exists as nearby CNVEs can be correlated due to linkage disequilibrium. LD between rare CNVs might generally be weak but is more difficult to assess given the small numbers. Deriving the effective number of independent tests and the proper genome-wide significance threshold for rare CNV analysis is still challenging.

587 common CNVEs with a frequency $>1\%$ were identified from the pooled CNV call set of the second study. This number is considerably smaller than the 1,319 copy number polymorphisms (CNPs) with allele frequency $>1\%$ discovered in the HapMap1 populations using the same array by McCarroll *et al* [34]. However, such differences are expected considering that (*i*) the McCarroll set included more smaller events (median: 7.4kb, IQR: 3.7-17.9kb) that were excluded by the stringent calling and QC pipeline used in this study (median: 31.7kb, IQR: 11.2-90.0kb), (*ii*) the McCarroll set consisted of CNVs found in other populations, especially African population which is known to have higher level of diversity, that could be rare in the UK population, and (*iii*) the size of HapMap1 populations is relatively small (270 in total) in which case accurate estimation of the frequency of less frequent CNVs is difficult.

Population stratification, allele frequency differences between cases and controls due to systematic ancestry differences, could cause spurious association in disease associations. In this study, the proportion of cases having a non-European ancestry was found to be considerably higher than that of controls, and it was partially tackled by excluding the most extreme ethnic outliers from both cases and controls. A minor inflation of the test statistics was still observed ($\lambda = 1.03$), which might be partially accounted for by the remaining ancestral differences between cases and controls. Existing CNV disease association studies rely on either a priori exclusion of ethnic outliers [13], as I did in this study, or on stratified analysis [9], both of which suffer a loss of power. Methods have been developed for SNP GWAS to correct ancestral differences, such as adjusting genotypes and phenotypes individually using the loadings of the principal component that represents the cline of geographical/ancestry distribution [103]. This provides a workaround for common bi-allelic CNVs well tagged by common SNPs. However, similar correction is yet to be incorporated into direct CNV association test that handles untagged CNVs.

The comprehensive association study of common CNVs undertaken by the Wellcome Trust Case Control Consortium reported that common CNVs are unlikely to play a major role in the genetic basis of common diseases and unlikely to account for a substantial proportion of the 'missing heritability' unexplained by SNP GWAS [13]. This seems to hold true in this study of severe early onset obesity. Only two of the 481 tested common CNVs exhibited convincing association and yet both are well tagged by common SNPs and the association of the larger deletion was discovered previously through the tagging SNP [72]. The association of the smaller deletion is a novel finding and the observed association of the larger deletion seems to be driven by this smaller deletion, particularly in females. As both deletions are well tagged by SNPs, existing GWAS data could be used to replicate this result.

The discordance with Willer *et al* [72] finding that it was the $\sim$43kb deletion and not the $\sim$8kb deletion that was associated with BMI might be simply due to technical reasons that the perfect tagging SNP of the $\sim$8kb deletion was not among the tested markers, as the $\sim$8kb deletion appeared to be discovered only after their investigation of the HapMap populations. If that tagging SNP was indeed tested and did not exhibit any association, then it might be attributed to biological differences between the genetic architecture of BMI as a quantitative trait and extreme early onset obesity as a binary trait. As a replication study undertaken by the GIANT consortium that uses the Sequenom platform to genotype tagging SNPs of both deletions in large obese patient cohorts and controls is underway, we shall know the answer very soon.

The sex bias in the association of the deletions upstream of *NEGR1* is intriguing. The data suggest that there is little association with either of the deletions in males, but in females the association is strong and is entirely driven by the smaller deletion. Most sex-specific associations tend to be linked with phenotypes that have biased distribution between the two sexes. However, it is not clear if there is significant obesity-related phenotypic difference between male and female subjects participating the study, at least the study was not designed to introduce such difference. This locus was not among the reported loci that exhibit sex-specific association with waist-hip ratio, a descriptor of body fat distribution, as discovered in a recent GWAS [104], so it seems unlikely to be explained by the difference in body fat distribution between male and female when gaining weight. At molecular level, as *NKX3.1* is regulated by androgen and a conserved putative binding site of *NKX3.1* is found within the larger deletion, the change in its relative position to *NEGR1* at the presence/absence of the smaller deletion might alter the expression of the gene in a sex-specific way. To examine this hypothesis, assays could be designed to monitoring changes in expression of *NEGR1* and other nearby genes on induction of *NKX3.1* in different haplotype backgrounds. Another more complex hypothesis could involve the bifunctional transcription factor *NKX6.1*, of which a conserved putative binding is found within the smaller deletion. In this hypothesis, *NKX6.1* might mask the sex-specific effect of *NKX3.1* by potent activation or repression of *NEGR1* when the binding site is present, and thus regulation by *NKX3.1* is only revealed when the *NKX6.1* binding site is removed by the smaller deletion. To test this hypothesis, experiments could be designed to monitor expression of *NEGR1* and nearby genes on inductions of *NKX3.1* in genetic background wherein the *NKX6.1* binding site within the smaller deletion is point mutated.

CHAPTER 4

CHARACTERIZING AND PREDICTING HAPLOINSUFFICIENCY IN THE HUMAN GENOME

4.1 Introduction

Haploinsufficiency, wherein a single functional copy of a gene is insufficient to maintain the normal phenotype of a diploid organism, is a major cause of human dominant diseases.

Dominance and recessiveness are fundamental concepts of Mendelian genetics. They describe the relationship between a pair of alleles of a gene of a diploid organism with respect to the phenotype they manifest. An allele, A, is dominant to another allele, a, if the corresponding phenotype of Aa is different from aa but indistinguishable from AA. A mutation can be described as dominant or recessive if it is dominant or recessive to the wildtype allele. The majority of observed naturally occurring (deleterious) mutations are recessive. While Fisher explained this as the result of selection for modifier genes that increase the fitness of heterozygotes [105], Wright viewed it as simply a physiological consequence of metabolic pathways [106]. Experimental and theoretical work over the years suggested Wright's explanation is more plausible. Kacser and Burns [107] established an excellent mathematical framework for understanding dominance/recessiveness at the molecular level and they showed that recessiveness emerges naturally from the kinetic properties of multi-enzyme system when most enzymes are far from being saturated.

The dominant mutations and the genes that harbor these mutations, though being the minority, contribute to a disproportionate ∼48% (965/2006) of human autosomal Mendelian disorders with known molecular basis recorded to date [108]. Wilkie categorized the molecular mechanisms of dominance into eight types [109], including haploinsufficiency, increased gene dosage, ectopic or temporally altered expression, increased or constitutive protein activity, dominant negative effect, altered structural protein, toxic protein alterations and new protein function. Among those types, haploinsufficiency is especially interesting, since (*i*) it is a relatively common mechanism for dominant diseases as a variety of mutations can lead to heterozygous loss-of-function; (*ii*) the ascertainment of loss-of-function mutations is relatively easy compared to gain-of-function mutations; (*iii*) the direct impact is solely through dosage reduction, which is easier for functional interpretation than other types of dominant mutation; (*iv*) it can be regarded as a property of a gene as the mutant allele is always defunct irrespective of the specific mutation. From a theoretical perspective, Veitia showed that haploinsufficiency is more likely to occur in systems that require the physical interaction of distinct macromolecules such as transcription regulation and assembly of protein complexes, in which the total output of the system is a sigmoid function of the dosage of each single entity [110]. From a more biological perspective, Wilkie suggested that genes encoding structural proteins are required in large quantities in specific tissues, and that subunits of protein complexes assembled under strict stoichiometry and regulatory proteins working close to a threshold level for different actions are more likely to be haploinsufficient [109]. Examples of these types include type 1 collagen [111], ribosomal proteins [112] and members of the *Hox* gene family [113].

Around three hundred genes have been reported haploinsufficient in human so far and Dang *et al* showed that they are less likely, compared to the rest of the genes, to be located in genomic regions susceptible to structural rearrangements [14]. This is expected, as large genomic deletions, a frequent consequence of structural rearrangements, are a major type of loss-of-function (LOF) mutation. Deletions encompassing the entire length of a gene unambiguously reduce the number of its functional copies. Partial deletions can also cause LOF, if key elements involved in the initialization of transcription, splicing and translation, such as promoter, splicing signals and start codon, are affected. Even if those elements are intact, premature stop codons could be introduced by frame-shifting deletions or simple truncating deletions, which likely subject the transcripts to nonsense-mediated decay, by which these transcripts are digested rather than translated into mutant proteins [114]. Indeed, large deletions have been found to be causal for diverse dominant developmental disorders, which, in turn, has led to the discovery of a number of haploinsufficient genes (HI genes), for example the discovery of the CHARGE syndrome gene, *CHD7* [115].

However, not all LOF mutations are deleterious. It is clear from sequenced genomes [116], exomes [117] and CNV surveys [12] that every genome, including those of apparently healthy individuals studied as controls in disease studies, harbors tens of unambiguous LOF mutations, including large genomic deletions. Some LOF mutations can be even advantageous [118]. Genes deleted in apparently healthy individuals seem not to be haploinsufficient, at least not to the point that carriers of heterozygous LOF mutations in these genes are kept from being recruited as controls for disease studies. Besides these haplosufficient (HS) genes, and the currently known HI genes, the dosage sensitivity of the majority of the genome remains elusive. Previous studies have shown that sets of HI genes, such as genes implicated in dominant diseases, have biased evolutionary and functional properties with respect to the rest of the genome [119–121]. However, there has not been a direct and systematic investigation of differences in properties between known HI genes and haplosufficient (HS) genes and it is unknown which properties are most informative in predicting dosage sensitivity.

With array-based copy number detection and the current generation of sequencing technologies, our ability to discover genetic variants in patients is running far ahead of our ability to interpret their functional impact and there is a pressing need to distinguish between benign and pathogenic variants. Computational methods have been developed to predict the molecular impact of non-synonymous point mutations. Some totally depend on sequence conservation at the site of the mutation, such as SubPSEC [122], Align-GVGD [123] and SIFT [124]. Some also consider structural and biochemical properties of the protein (stability, solubility, active sites, etc), such as SNPs3D [125] and PolyPhen [126]. The output of these algorithms is often a continuous score or a category label indicating how damaging the mutation is to the encoded protein. Although, these outputs have been shown to be useful in identifying pathogenic mutations for Mendelian diseases [127], their power to predict impact on fitness at individual level might still be limited, especially in the case of heterozygous mutation wherein one of the alleles still functions normally, as they do not distinguish between the heterozygous and homozygous genotypes of a variant. Computational tools for predicting the functional impact of large copy number variants are still in their infancy [128]. The problem differs from non-synonymous point mutations in that large CNVs can affect multiple genes as well as non-coding regions simultaneously, and thus their interpretation requires the integration of different functional annotations to maximize the information on all affected entities.

Application of such computational interpretative tools in clinical settings requires careful consideration, as these tools are usually trained on collated sets of known damaging and benign mutations that could well be a biased representation of the true spectrum of causal mutations found in real patients or in the general population. The scores or classifications generated by these computation tools are rarely calibrated to diagnostic outcomes, and only infrequently are the distributions of such scores compared between patients and population controls. Characterizing the distribution of such scores in patient and population cohorts has become more feasible in recent years with the growth in databases of pathogenic variants [129, 130] as well as of variants found in large population surveys [12, 34, 77, 78, 131, 132]. Additionally, pathogenicity scores are often just one of the many different types of evidence that influence diagnostic interpretation and needs to be integrated with the other evidence in a sensible way. Most current genetic diagnostic practices adopt a decision-tree-like procedure [133, 134]. A probabilistic process would be desirable which could give every diagnosis a level of confidence. Goldgar *et al* suggested a naïve Bayesian framework to integrate different, typically uncorrelated, types of information and demonstrated its application to the interpretation of variants of unknown clinical significance in the *BRAC1* and *BRAC2* genes [135].

In the work described in this chapter, I first explored the genomic, functional and evolutionary characteristics of HI genes and then I developed a computational approach to predict which genes might exhibit haploinsufficiency. I then investigated the utility of the gene-based HI predictions to measure pathogenicity of large copy number variants, both deletions and duplications. Finally, I proposed a probabilistic diagnostic framework that integrates population distributions of pathogenicity scores, with additional evidence to generate a level of confidence for the diagnosis of causal CNVs, and potentially other forms of genetic variants.

4.2 Materials and methods

4.2.1 Control data

The controls include a set of 6,000 UK individuals recruited as common controls in GWAS of 13 disease conditions undertaken by Wellcome Trust Case Control Consortium 2 (WTCCC2), of which 3,000 samples are from the 1958 British Birth Cohort and 3,000 samples are from the UK Blood Service Control Group. Another set of 2,421 US control individuals, 1,442 of which have European ancestry and the rest with African-American ancestry, are from a control cohort used in GWAS of Schizophrenia and Bipolar disease undertaken by Genetic Association Information Network (GAIN). Samples were previously genotyped on Affymetrix genome-wide human SNP array 6.0. Affymetrix 6.0 CEL files were obtained from Wellcome Trust Case Control Consortium 2 for WTCCC2 controls and from the Database of Genotype and Phenotype (dbGaP) through accession number phs000017 and phs000021 for GAIN controls.

4.2.2 Asserting of loss of function genes

To identify protein-coding genes disrupted in a LOF manner, CNV calls made by the calling pipeline described in Chapter 2 were compared to gene annotation provided by EnsEMBL [136]. Four scenarios were considered LOF to a protein-coding transcript:

1. deletion of over 50% of coding sequence
2. deletion of the start codon or the first exon
3. deletion-disrupted-splicing
4. deletion-caused frame-shift

A gene was considered LOF if all of its transcripts were LOF. Under these criteria, CNVs were identified in GWAS control individuals with a LOF impact on 2,677 genes. I defined haplosufficient genes as being those observed as LOF genes in two or more GWAS control individuals.

4.2.3 Preparing possible predictor variables

4.2.3.1 Genomic properties

The length of gene, spliced transcript, 3'UTR and coding sequence and the number of exons were calculated on the basis of gene annotation downloaded from EnsEMBL. The number of protein domains was retrieved from EnsEMBL build 50.

4.2.3.2 Evolutionary properties

dN/dS data was downloaded from EnsEMBL. Genomic Evolutionary Rate Profiling (GERP) [137] score was downloaded from EBI. Two summed GERP values, one for coding sequence and the other for promoter region, defined as bases within ±100bp of the transcription start site, were then calculated for all human protein-coding transcripts according to EnsEMBL annotations and summarized by gene using the median values. A third summed GERP value for conserved noncoding elements around genes was calculated as the sum of GERP scores of all bases of annotated conserved noncoding elements within an interval ±50kb of the gene. To derive the list of conserved noncoding elements, I retrieved a list of conserved elements throughout placental mammals from the UCSC genome browser (28-Way

Most Cons track) and removed elements overlapping with exons according to EnsEMBL gene annotation. The number and identity of paralogs were downloaded from EnsEMBL.

4.2.3.3 Functional properties

Gene expression profiles in human were obtained from the GNF Atlas [138]. Total expression levels were normalized across genes and the standard deviation of expression across normal tissue types of each gene was used to indicate its tissue specificity of expression. Genes over-expressed by at least 8 fold in human embryonic stem cells [139], fetal tissues [138] and mouse fetal tissues [140] were collectively treated as genes expressed at embryonic stage. A binary coding was used to represent this property in which genes expressed at embryonic stage were labeled 1 and the rest were labeled 0.

4.2.3.4 Network properties

Two interaction networks were used. One is a binary protein-protein interaction network integrated from a number of sources [141–145]. Proteins were mapped to their coding genes and interactions were not counted repeatedly if multiple proteins were mapped to a single gene. This network included 70,632 interactions among 11,077 genes. The other is a probabilistic gene interaction network (a network of 470,217 links among 16,375 human genes calculated using methods previously described for yeast [146] and worm [147] and derived from 22 publicly available genomics datasets including DNA microarray data, protein-protein interactions, genetic interactions, literature mining, comparative genomics, and orthologous transfer of gene-gene functional associations from fly, worm, and yeast, where the weight of a link is the log likelihood score of the interaction [146]. Measures of centrality (degree, betweenness) and modularity (cluster coefficient) were calculated using MCL [148]. Shortest path distance and sum of weight of interactions [147] were calculated as measures of proximity to a group of ′seed′ genes.

4.2.3.5 Other properties

A list of 300 genes implicated in cancer was downloaded from the COSMIC database [149]. Growth rate of yeast heterozygous deletion strains were from Deutschbauer *et al* [150].

4.2.4 Comparing predictor variables between HI and HS genes

For continuous variables, the two-tailed Mann-Whitney U test was performed to assess if positive (haploinsufficient) and negative (haplosufficient) training data have the same median value for potential predictor variables. For two-class categorical features, Fisher's exact tests were performed. Statistical tests were performed using R (`http://www.r-project.org`).

4.2.5 Feature selection for the predictive model

I assessed different potential sets of predictor variables for input into the predictive model using the following criteria: (*i*) they allow prediction for at least half the genes in the genome, (*ii*) the Spearman correlation ρ^2 between all pairs of predictor variables is less than 0.05, (*iii*) they are drawn from different broad categories (genomic, evolutionary, functional and network) if possible, and (*iv*) achieve best performance in model assessment.

4.2.6 Assessing model performance

The *sensitivity* of the prediction was plotted against $1 - specificity$ and the area under the ROC curve (AUC) [151] was used as quantitative measure of the performance of the model, where $sensitivity = TP/\,(TP+FN)$, and $specificity = TN/\,(TN+FP)$. The other measure used is the Matthews correlation coefficients (MCC) [152], defined as:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

To avoid over-fitting, the sensitivity and specificity were calculated using 10-fold cross-validation. To overcome the variability caused by random partition involved in 10-fold cross-validation, each such assessment was repeated 30 times and the mean values were reported.

4.2.7 Multiple imputation

Multiple imputation was used to fill in ('impute') the missing values for predictor variables incorporated in the model, namely 'dN/dS ratio between human and macaque', 'promoter conservation (GERP)', and 'gene network proximity to HI genes', except for 'embryonic expression' of which the genomic coverage is 100%. Since 'gene network proximity to HI genes' and 'promoter conservation (GERP)' are the top two predictive variables, genes missing both values were removed. To achieve better imputation, I included three additional gene properties, namely 'CDS conservation (GERP)', 'spliced transcript length' and 'gene network betweenness centrality' in the imputation process. Twenty independent imputations of 20 iterations were undertaken. In each iteration, imputation for each predictor variable was in the order of increasing number of missing values using the predictive mean matching method. The computation was done using the R package MICE [153].

4.2.8 Parameter estimation for the Bayesian diagnostic framework

The prior probability of a CNV being causal ($\mathrm{p}(C)$) was estimated as the average number of CNVs found per individual divided by the current diagnostic rate for CNVs. Diagnostic rate and average number of CNVs found per individual were taken from Buysee *et al* [134], which found on average 0.86 deletions and 0.73 duplications per individual and achieved a diagnostic rate of 0.1 using BAC array and Agilent 44K array CGH.

The probability of a causal CNV being rare (population frequency $< 1\%$) ($\mathrm{p}(F|C), F =$ rare) was set at 1. The probability of a causal CNV being *de novo* ($\mathrm{p}(F|C), F =$ *de novo*) was also taken from Buysee *et al* [134] in which 73% of the causal CNVs found were *de novo*. The distribution of pathogenicity scores of *de novo* CNVs in DE-

CIPHER [129] was used to approximate that of causal CNVs. The probability of a causal and rare (or *de novo*) CNV having a pathogenicity score equals to x was taken as the empirical estimation of probability density of the distribution of pathogenicity scores of causal CNVs at x.

The probability of a benign CNV being rare (population frequency $< 1\%$) ($\text{p}(F|\bar{C}), F =$ rare) was estimated as the fraction of WTCCC2 and GAIN control CNVs with a carrier frequency $< 1\%$. The probability of a benign CNV being *de novo* ($\text{p}(F|\bar{C}), F =$ *de novo*) was also taken from Itsara *et al* [131] in which 0.44% of the CNVs found in children of apparently healthy trios were *de novo*. The distribution of pathogenicity scores of benign CNVs was generated using WTCCC2 and GAIN control CNVs after excluding CNVs at known pathogenic loci recorded in DECIPHER. The probability of a benign and rare (or *de novo*) CNV having a given pathogenicity score equals to x was taken as the empirical estimation of probability density of the distribution of pathogenicity scores of benign CNVs with a carrier frequency $< 1\%$ (or with an occurrence of 1, *i.e.* singletons) at x. Since WTCCC2 and GAIN control CNVs were discovered using arrays of considerably higher resolution than the CNVs discovered by Buysee *et al* and the CNVs recorded in DECIPHER, deletions <180kb and duplication <330kb were excluded prior to the above calculation in order to match the number of CNVs discovered per individual.

4.2.9 Text mining through PubMed abstracts

The title and abstract of publications that contain the keyword 'haploinsufficiency' or 'haploinsufficient' were retrieved from PubMed on Aug 2010, using the search term 'haploinsufficient[Title/Abstract] OR haploinsufficiency[Title/Abstract] AND humans[MeSH Terms]'. After cleaning the text, a word frequency table was compiled from all titles and abstracts. A dictionary that maps gene names and synonyms to gene symbols was downloaded from HGNC [154]. For each title and abstract, the sentence containing the keyword 'haploinsufficiency' or 'haploinsufficient' was extracted and parsed by the GENIA tagger [155] to break the sentence into chunks and tag the part-of-speech of each chunk. The chunk immediately before the keyword, the noun chunk in front of a verb and a preposition in front of the keyword were extracted. These chunks were first examined by GENIA tagger to identify the named biomedical entity. If this failed, the noun in the chunk that appeared fewer than 10 times as recorded in the frequency table and contained numbers or capital letters, or followed immediately by ‘gene’, ‘protein’ or ‘transcript’ was kept as potential gene name. These potential gene names and named entities identified by the GENIA tagger were looked up in the gene name dictionary to convert into unique HGNC gene symbols.

4.3 Results

4.3.1 Characteristics of haploinsufficient genes

I first compiled a list of known human HI genes and a catalog of HS genes. Known HI genes were collated from literature [14, 156]. The catalog of HS genes was generated from genes disrupted in a loss-of-function manner in control individuals used in genome-wide association studies by CNVs detected in data from the Affymetrix 6.0 chip (see Methods). I identified 2,676 putative HS genes seen in any control individuals and 1,079 seen in two or more controls (Figure 4.1), and used the latter set in most downstream analyses. Thus the final list of HI and HS genes contains 301 and 1,079 genes respectively.

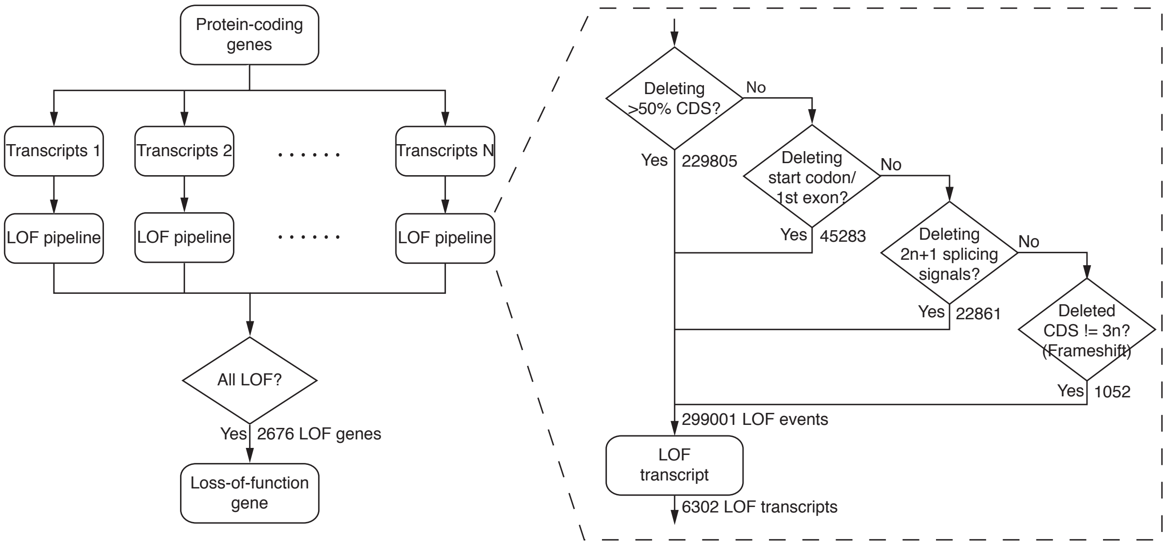

Figure 4.1: Procedure for LOF calling. The flow chart shows the pipeline used to identify LOF genes. A gene with all its transcripts disrupted under any of the four considered LOF scenarios is regarded as LOF. On the right, the numbers under each scenario denotes the number of detected LOF events meeting that criterion. A LOF event is defined as loss of function of one transcript in one individual.

To systematically assess the difference in properties between HI and HS genes, I gathered a large number of annotations describing the evolutionary, functionary and interaction properties of genes (see Methods) and examined the distribution of each individual property in HI and HS genes. I found that HI genes have consistently a more conserved coding sequence (human-macaque dN/dS, p = 3.12×10^{-26}), a less mutable promoter (p $< 1\times10^{-30}$), paralogs with lower sequence similarity (p = 1.84×10^{-9}), a longer spliced transcript (p $<$ 1×10^{-30}), a longer 3′UTR (p = 2.63×10^{-12}), higher expression during early development (p = 1.10×10^{-15}), higher tissue specificity in expression (p = 2.29×10^{-6}), more interaction partners in both a protein-protein interaction network (p $< 1\times10^{-30}$) and a gene interaction network (p $< 1\times10^{-30}$) and higher chances of interacting with other known HI genes (p $<$ 1×10^{-30}) and cancer genes (p $< 1\times10^{-30}$) (Figure 4.2). Interestingly, the growth rate of yeast heterozygous deletion strains does not seem to differ between their HI human homologs and HS human homologs, probably reflecting the vast functional differences between the majority of yeast and human genes, except those involved in highly conserved cellular processes.

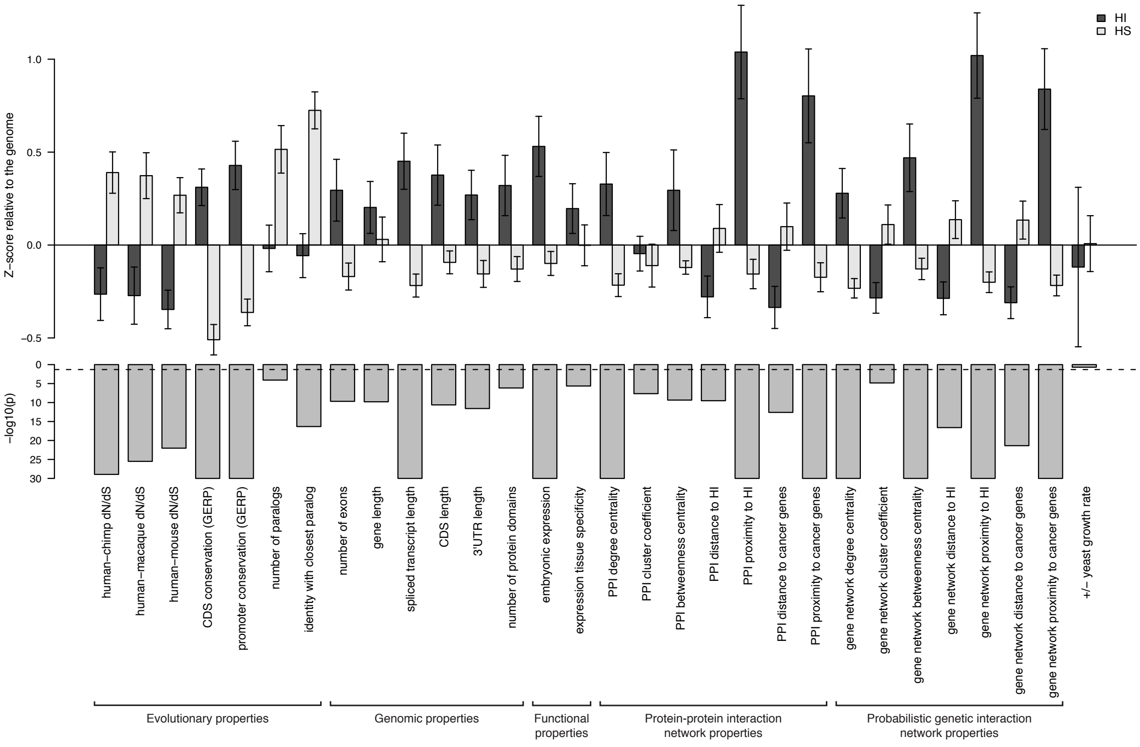

Figure 4.2: Properties that distinguish HI genes from HS genes. The upper part of the figure shows the comparison of the mean of each individual property between HI genes and HS genes. The values are transformed to z-scores relative to the genome average. The error bars represent two times the standard error of the mean. The bars in the middle part present the transformed p value (-log$_{10}$p) of the Mann-Whitney test on each property. The dashed line marks a p value of 0.05.

Table 4.1: Genomic coverage of gene properties

Property	#Genes	Genomic coverage[*]
Human-chimp dN/dS	15,084	79.50%
Human-macaque dN/dS	15,025	79.20%
Human-mouse dN/dS	14,386	75.80%
Coding sequence GERP	17,164	90.50%
Promoter GERP	16,807	88.70%
Number of paralogs	11,066	58.30%
Identity of closest paralog		
Number of exons	17,700	93.30%
Length of gene		
Length of spliced transcript		
Length of coding sequence		
Length of 3'UTR		
Number of domains	14,722	88.50%
Embryonic expression[†]	18,962 (2421)	100% (12.8%)
Tissue specificity of expression	13,950	73.60%
PPI network properties[‡]	11,077	58.40%
Genetic network properties[‡]	14,664	77.30%
+/- Yeast growth rate	3,352	17.70%

[*] Calculated relative to the number of EnsEMBL annotated protein-coding genes that can be uniquely mapped to HGNC symbol.

[†] Since this is a binary factor where every gene is classified as either over-expressed or not in embryo tissue, the coverage is 100%. The number and fraction of genes over-expressed in embryo is listed in parenthesis.

[‡] Including degree, cluster coefficient, betweenness, distance to known HI/cancer genes, proximity to known HI/cancer genes.

4.3.2 Training a model to classify HI and HS genes

The highly significant differences in genomic, evolutionary, functional and network properties between HI and HS genes suggest some combination of these properties may be predictive of haploinsufficiency. I used linear discriminant analysis (LDA) as the supervised classifier, which, given multi-dimensional data and class labels, finds the linear combination of the given dimensions (linear discriminant) that maximizes the inter-class variance. I trained the classifier using various sets of gene properties to obtain a classification model and applied the model to estimate a probability of being HI (p(HI)) for all protein-coding genes in the genome for which all the selected predictor variables were available. Finally, I validated the predictions using external data sets.

The final result is presented below and is followed by discussion of more detailed questions: (*i*) which gene properties should be incorporated (Section 4.3.2.1) ? (*ii*) which training dataset should be used (Section 4.3.2.2) ? (*iii*) does a more sophisticated classifier perform better (Section 4.3.2.3) ? Section 4.3.2.4 presents the validation of prediction. Section 4.3.2.5 described some further improvements of the prediction of which the outcome is not included below as they were undertaken at a later stage.

After assessing various different sets of predictor variables (see Methods, and below) my initial classifier was trained with four predictor variables: dN/dS between human and macaque, promoter conservation, embryonic expression and network proximity to known HI genes. The model was obtained by training on 234 HI genes and 326 HS genes for which the predictor variables were available. All predictor variables were scaled to the same variance before entering LDA so that their contribution can be measured by the coefficients of the resulting linear discriminant. Proximity to known HI genes provided the most predictive power. The model achieved an AUC of 0.81 and a MCC of 0.50 in ten-fold cross-validation (Figure 4.3). I applied the model to estimate a probability of being HI for all 12,443 protein-coding genes in the genome for which all four selected predictor variables were available. The distribution of the predicted p(HI) is clearly bimodal, with a large peak near 0.2 and a much smaller peak at 1 (Figure 4.4 left). The distributions of p(HI) for the HI and HS training sets differ significantly ($p < 1\times10^{-30}$, Mann-Whitney test or Kolmogorov-Smirnov test) (Figure 4.4 right).

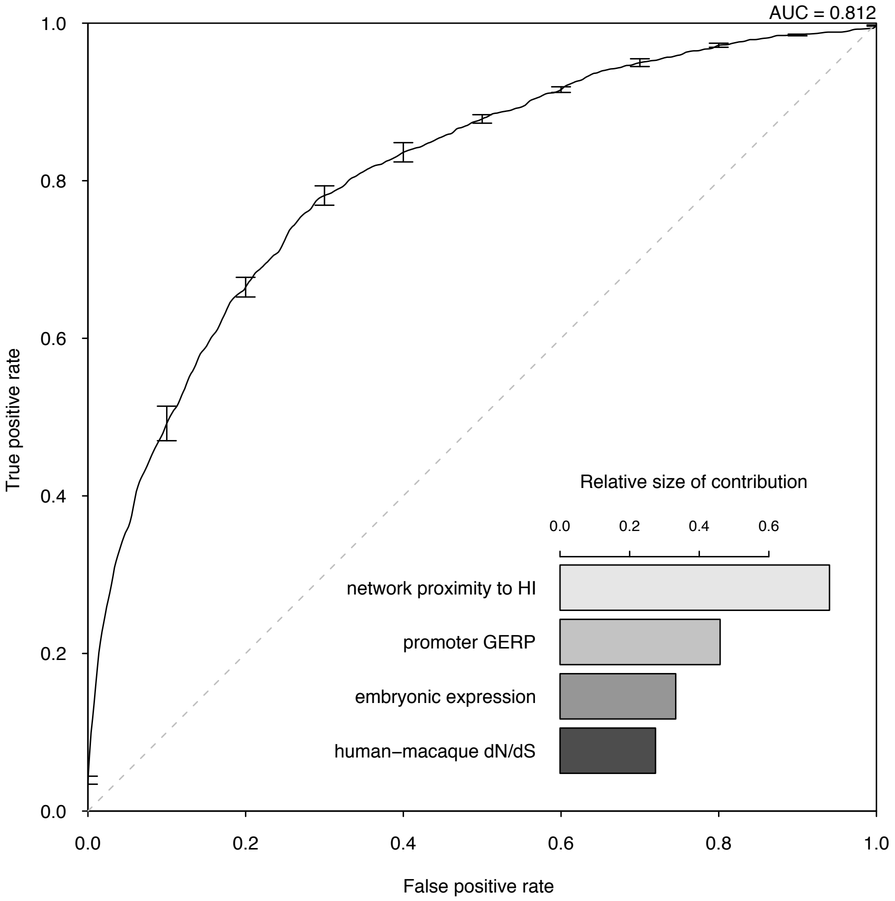

Figure 4.3: Assessment of model performance. The ROC curve demonstrates the performance of the model evaluated by 10-fold cross-validation. The lower right part shows the relative contribution of each predictor variable to the prediction model measured by the absolute value of the scaling factor of each predictor variable constituting the linear discriminant.

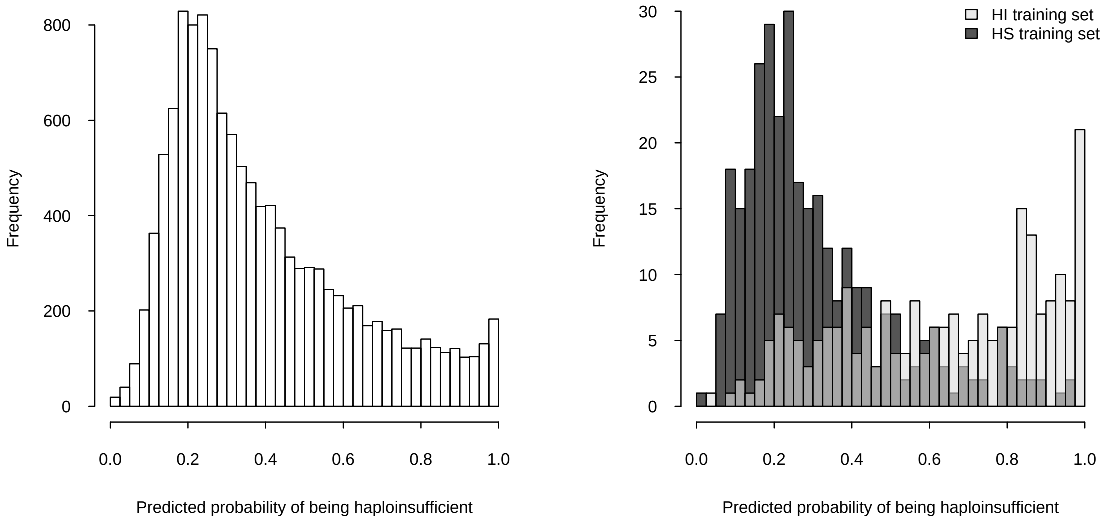

Figure 4.4: Predicted probability of being haploinsufficient. The histogram on the left shows the distribution of the predicted probability of being haploinsufficient (p(HI)) of all 12,443 predictable genes. The histogram on the right shows the distribution of the predicted p(HI) of the HI training set (light grey) and the HS training set (dark grey).

4.3.2.1 Integrating information from multiple 'orthogonal' predictor variables improves classification

To assess the marginal utility of using more than one predictor variable, I trained separate LDA models from the same set of genes (known HI genes plus HS genes) using only one predictor variable at a time and compared the cross-validation performance with using all predictor variables. The latter out-performs models using single predictor variable (max AUC = 0.78 for network proximity to known HI genes whereas the integrated model achieves 0.81) (Figure 4.5), indicating that combining the predictor variables together generated a more predictive model than considering any of the individual predictor variables in isolation.

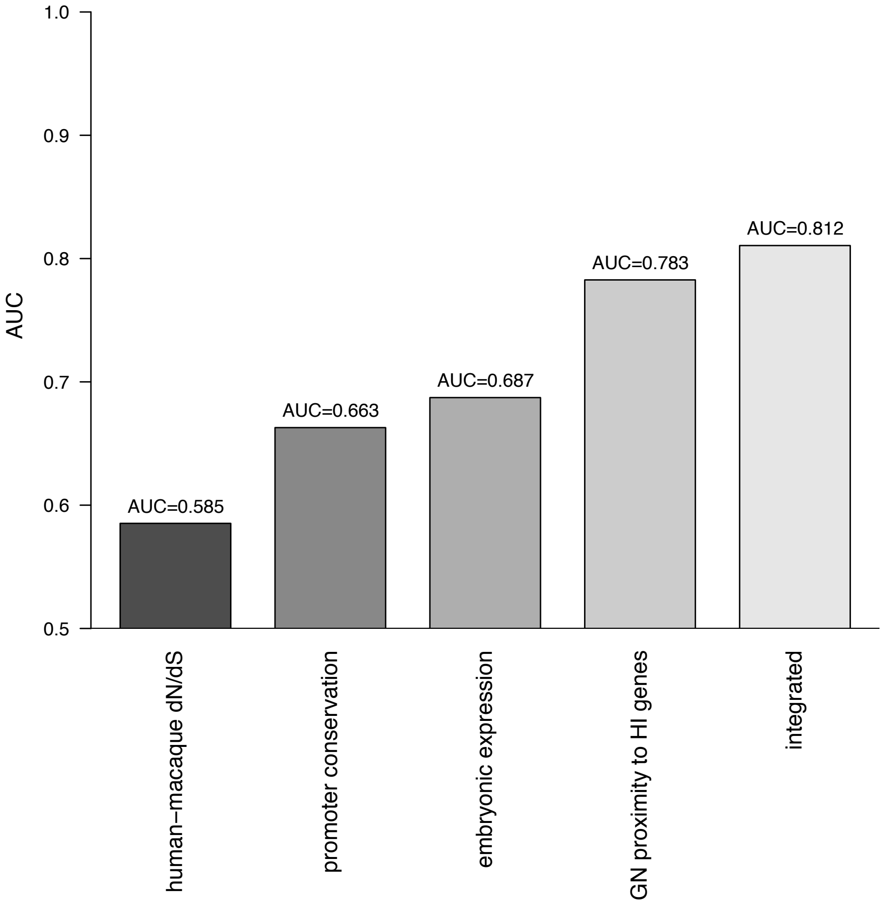

Figure 4.5: Prediction performance of single predictor variable and integrated model. Mean AUC of each model in 10-fold cross-validation repeated 30 times are shown as vertical bars with the actual values label at the top.

Since each gene property annotation is only available for a fraction of genes in the genome (Table 4.1), there is a trade-off between the possible increase in prediction performance by considering more gene properties as predictor variables and the decrease in the coverage of genes one could predict. Therefore, I aimed to select a small number of most predictive properties that are relatively 'orthogonal' in the kind of information they provide (see Methods).

After evaluating a number of possible combinations of predictor variables, which all had similar performance (Figure 4.6), I selected a model comprising of 'dN/dS between human and macaque', 'promoter conservation', 'embryonic expression' and 'network proximity to known HI genes'.

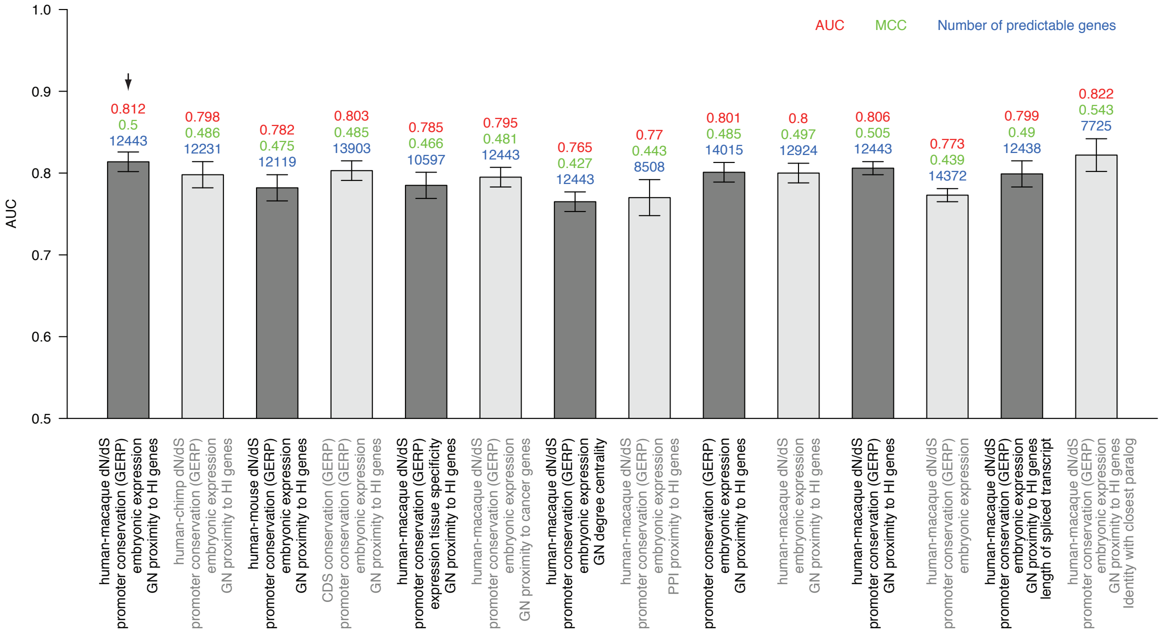

Figure 4.6: Comparison of model performance. The AUCs of each combination of predictor variables in 10-fold cross validation repeated 30 times are shown as vertical bars with error bars represent 2 times standard deviation. The mean AUC (red), mean MCC (green) and the overall gene coverage (blue) are labeled on top of each bar. The bar pointed by the black arrowhead is the chosen combination of predictor variables.

4.3.2.2 Using HS genes as negative training set improves classification

Previous studies [119–121] have compared HI-related gene sets against the rest of the genome to describe their characteristics. I investigated how the choice of negative training set influences the performance of my prediction model. I generated gene sets of different sizes randomly sampled from non-HI genes with complete predictor variable information and compared the cross-validation performance (AUC) resulting from the use of these gene sets as the negative training set to the use of the HS gene set as the negative training set (Figure 4.7). The use of a judiciously selected HS gene set is clearly advantageous.

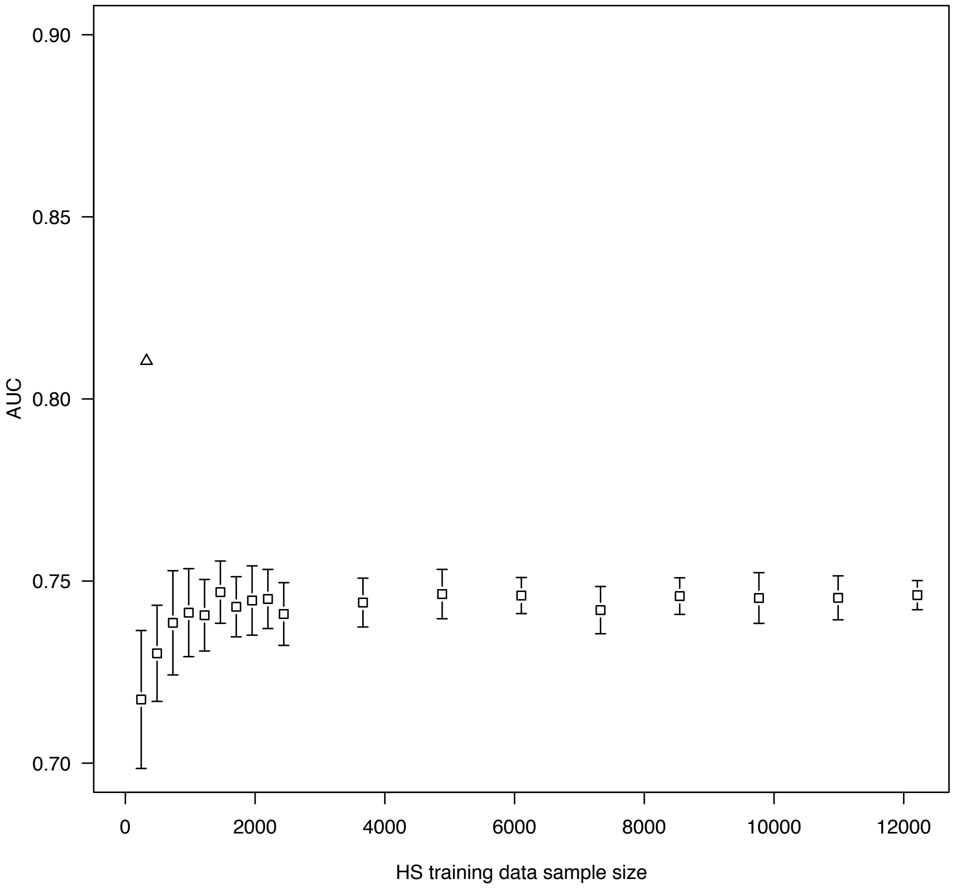

Figure 4.7: Prediction performance of using HS and genome background as negative training set. The plot compares the cross-validation performances resulted from using different gene sets as negative training set. The triangle represents HS gene set generated from CNV data. The squares represent different sizes of random gene sets sampled from the genome after excluding known HI genes. For each size, the gene set was sampled 20 times and the standard deviation of the resulting performances is shown as error bar.

I further investigated if our model performance is sensitive to the CNV discovery and filtering parameters, which determines the stringency of the HS gene set. I examined the influence on cross-validation performance of using different confidence thresholds (Birdseye LOD score) in CNV discovery and population frequency when generating HS gene set. A greater LOD score indicates higher confidence and thus a more stringent CNV set. Similarly, the more frequently a gene is found LOF in apparently healthy individuals, the more likely it is haplosufficient, and thus the negative training set is more stringent. I found that the LOD score threshold has little influence on the model performance, within the range I assessed (Figure 4.8). The use of recurrent LOF genes exhibits an apparent improvement of performance over the use of all LOF genes under most LOD thresholds. Further increase in stringency by requiring higher frequency results in further reduction of the size of negative training set, but little if any increase in performance of the prediction model. Therefore, I adopted the negative training set generated under 'LOD > 10' and 'found in at least two individuals' in further analysis.

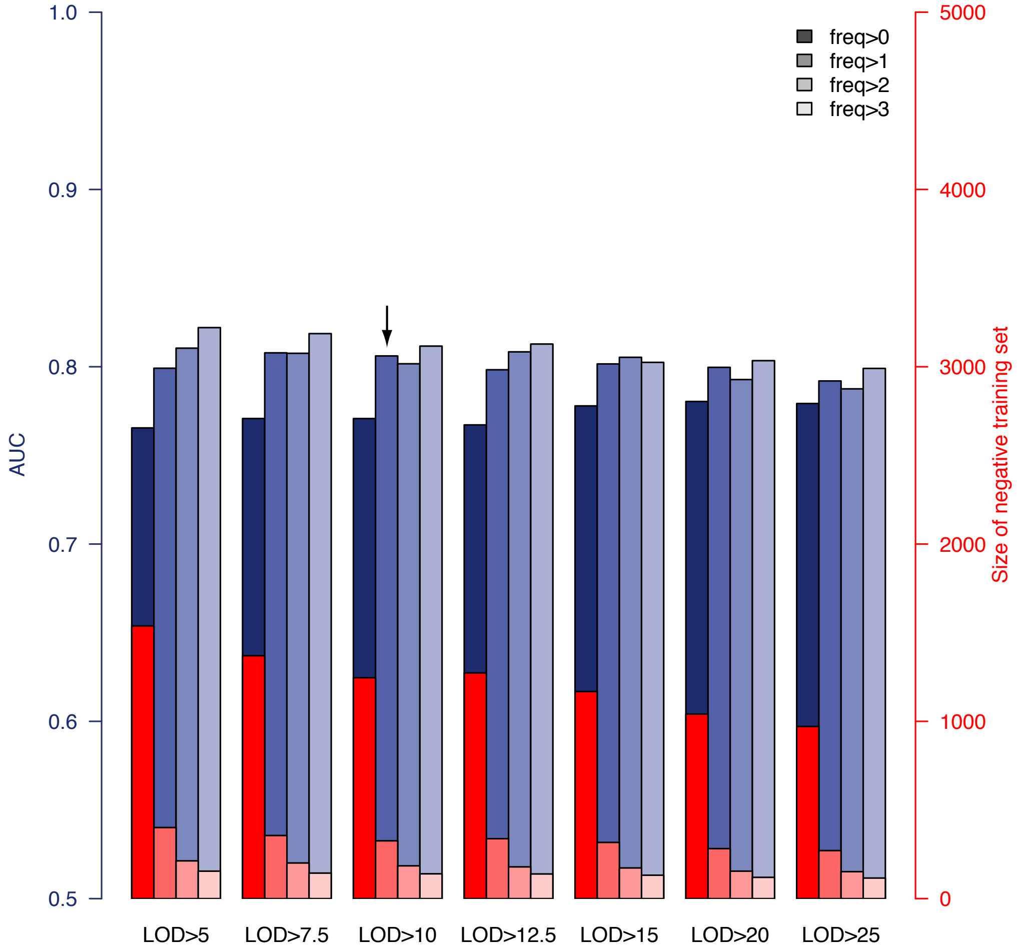

Figure 4.8: Prediction performance under different parameters used in generation of negative training set. The cross-validation performance (AUC) resulted from using negative training sets generated with different parameters are represented by blue vertical bars with axis on the left. The sizes of these negative training sets are represented by red vertical bars with axis on the right. Bars are grouped by the CNV calling parameters, LOD score, and within each group the darkness of coloring represent different frequency threshold used to define HS as shown in the legend. The bar pointed by the black arrowhead represents parameters and corresponding negative training set adopted in further analysis.

4.3.2.3 LDA achieves similar classification performance compared to a more sophisticated classifier

I investigated if the use of support vector machine (SVM), a more sophisticated machine learning method, as classifier would improve prediction performance. An SVM model was trained on the same training set as LDA with optimized parameters (gamma = 0.1, cost = 1) and class weights. The performance was examined by self-validation, leave-one-out cross-validation and 10-fold cross-validation. Despite being more sophisticated and computational expensive, SVM exhibits no appreciable improvement over LDA (Figure 4.9).

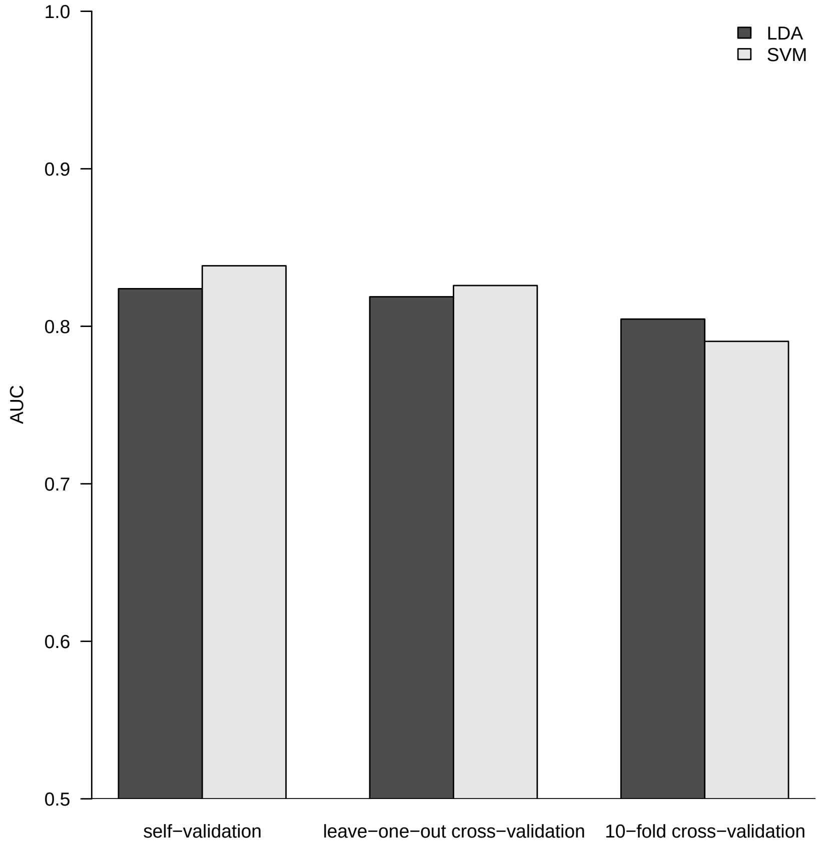

Figure 4.9: Comparing the prediction performance of LDA and SVM. The plot shows the comparison of prediction performance between LDA (dark bar) and SVM (light bar) using three approaches (from left to right): self-validation, leave-one-out cross-validation and 10-fold cross-validation. In the first two comparisons, SVM exhibits only very marginal improvement over LDA, whereas in the third LDA is marginally better.

4.3.2.4 Validating haploinsufficiency predictions using external datasets

It is not possible to assess how well-calibrated the predicted probabilities of being HI are, as the fraction of human genes that exhibit HI is not known. I therefore sought to validate these predictions using indirect approaches that examined the distribution of p(HI) in independent gene sets enriched for HI. As there is no credible estimation of the number of human HI genes, in some of the following validation analyses I arbitrarily labeled the genes in the top 10% of p(HI) as being predicted HI genes. However, the results were robust against this threshold being varied by at least a factor of at least 2.

First, I asked if genes implicated in human dominant diseases were enriched in our predicted HI genes relative to recessive-disease-causing genes. I retrieved 571 and 772 genes implicated in dominant and recessive disease from the OMIM and hOMIM[119] database, respectively, with no information regarding haploinsufficiency (and thus not included in our training data), and compared the distribution of predicted p(HI) against each other. The HI status could be predicted for 392 dominant genes and 606 recessive genes, of which 87 and 39 were predicted as being HI, respectively. This 4.14 fold enrichment of genes predicted to be HI within the dominant disease gene set is highly significant ($p = 4.46\times10^{-13}$, Fisher's exact test). Simply comparing the distribution of p(HI) values for these dominant and recessive genes also shows a highly significant shift towards high p(HI) values in dominant relative to recessive genes ($p = 4.44\times10^{-16}$, Mann-Whitney U test) (Figure 4.10).

Second, I asked if heterozygous knockouts of the orthologs of predicted human HI genes are more likely to cause severe phenotypic abnormalities in mice. For this purpose, I extracted a list of 1,523 mouse genes whose heterozygous knockout cause various abnormal phenotypes from the MGI database, mapped them onto orthologous genes in humans, removed orthologs to genes in our training gene sets and extracted the predicted p(HI) for the remainder. HI status could be predicted for the orthologs of 1,063 of these genes and 260 (24.5%) of them were predicted HI, indicating a 2.45 fold enrichment ($p < 1\times10^{-30}$, Fisher's exact test) (Figure 4.11). If focusing on those genes of which the heterozygous LOF phenotypes involve prenatal lethality (MP:0002080), the fold of enrichment increased to 4.38 ($p = 3.60\times10^{-12}$, Fisher's exact test) (28 predicted as HI out of 64 that could be predicted).

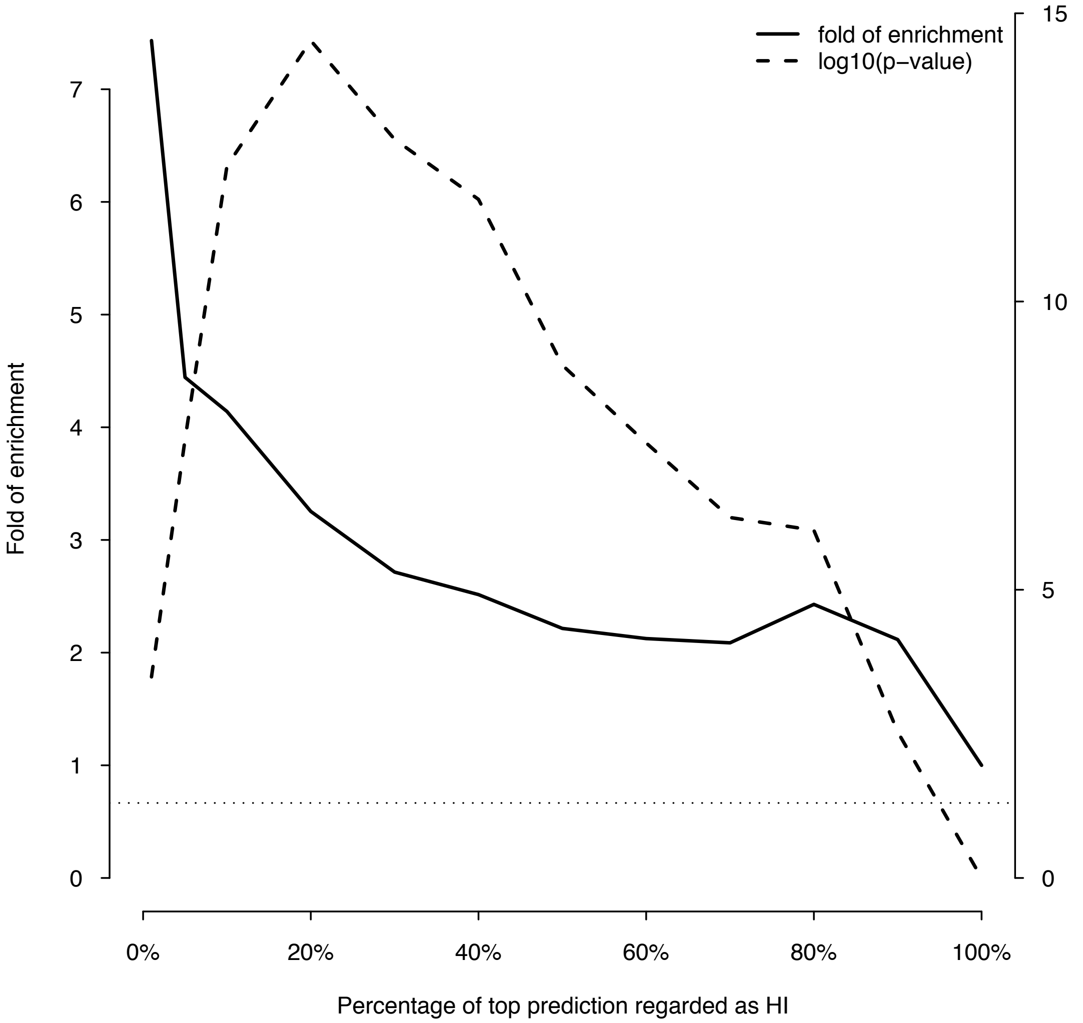

Figure 4.10: Enrichment of predicted HI genes in dominant genes relative to recessive genes. This plot shows the fold of enrichment of predicted HI genes in dominant genes relative to recessive genes (thick solid line) as a function of the proportion of predictions labeled as being haploinsufficient. Also plotted is the transformed p value (-log10p) of the corresponding Fisher's exact test (thick dashed line). The horizontal dashed line marks the p value of 0.05.

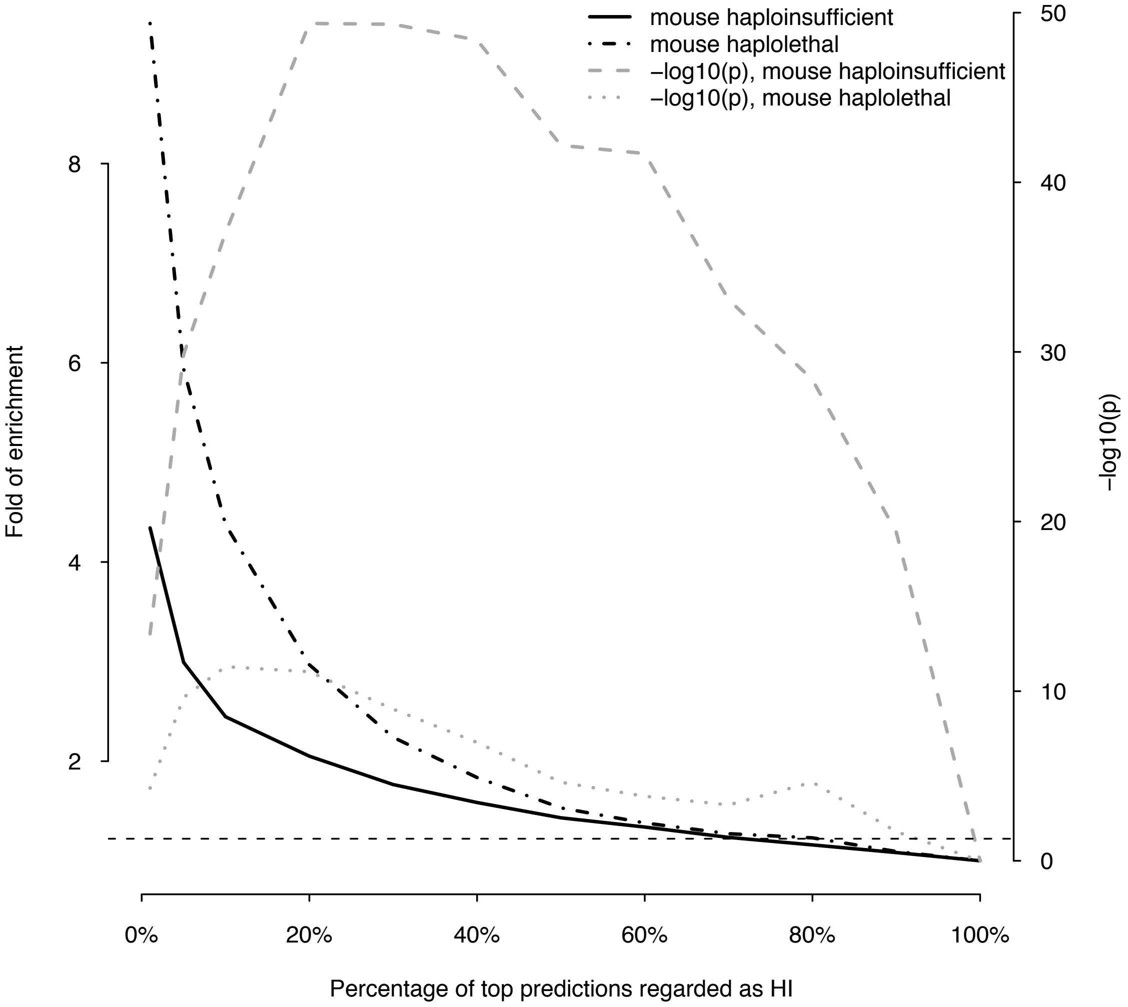

Figure 4.11: Enrichment of predicted HI genes in orthologs of mouse haploinsufficient genes and mouse haplolethal genes. This plot shows the fold of enrichment of predicted HI genes in human orthologs of mouse haploinsufficient genes (black solid line) and mouse haplolethal genes (black dashed line) relative to the genome average as a function of the proportion of predictions labeled as being haploinsufficient. The two lines in grey show the transformed p values of the corresponding Fisher's exact test. The horizontal dashed line marks the p value of 0.05.

4.3.2.5 Improving prediction with expanded training data and improved predictor variables

Having achieved reasonable performance with my initial predictive model of gene haploinsufficiency and shown that neither changing the classifier nor how the HS gene training data are filtered, I explored different potential strategies to improve upon the performance of this predictive model. In this section I describe two, potentially complementary strategies: (*i*) Including new and improved predictor variables into the predictive model, and (*ii*) using improved positive control training data (*i.e.* known HI genes).

4.3.2.5.1 Inclusion of new and improved predictor variables

In the light of the emerging role of conserved noncoding elements in regulation of gene expression, especially of developmental genes known to be dosage sensitive, I investigated several variables that summarize the extent of conserved noncoding sequence within and flanking a gene. I settled on the sum of GERP scores of all bases of conserved non-coding elements within an interval ±50kb of the gene as a candidate predictor variable. This property differs significantly between HI and HS genes ($p = 4.0 \times 10^{-54}$, Mann-Whitney U test).

The coverage of the protein-protein interaction network was also expanded from 11,077 genes and 70,632 interactions to 16,390 genes and 1,240,972 interactions by incorporating data from the STRING database [157]. As a result, the number of genes predictable with the same predictor variables as selected in Section 4.3.2.1 increased to 13,030 (+5%) without imputation or, if using the predictor variables optimized for the updated gene properties as described in Section 4.3.2.5.2, increased to 16,017 (+29%). I also updated the gene property annotations to EnsEMBL 53.

4.3.2.5.2 Improved HI training set through literature mining and manual curation

The known HI genes used as positive training set was initially taken from Dang *et al* and Seidman *et al*, which reflected the current knowledge in Nov 2007. I performed a literature searching on Aug 2010 to include more, newly discovered HI genes. Through text mining of PubMed abstracts (see Methods), 138 genes were added to the HI set, resulting in a combined set of 439 genes. 358 of these genes for which a PubMed abstract is available were manually curated. After curation of the entire set, 40 genes were removed, 55 were labeled as with weak evidence (see Appendix A). 72 genes that are involved in cancer [149] were also removed, since seemingly dominant inheritance could be the result of somatic loss of heterozygosity instead of truly genetic haploinsufficiency in the case of cancer.

To evaluate the expanded and manually curated HI set, the model was re-trained both with and without genes with weak evidence using the updated version of the same predictor variables as Section 4.3.2.1 and the performance were measured by two approaches: (*i*) the cross-validation AUC (AUC_{CV}) and (*ii*) the AUC for classifying pathogenic and benign CNVs using model-prediction-based LOD scores (AUC_{LOD}) . The model trained with the more stringent set exhibited higher cross-validation AUC than the model trained with more relaxed set (0.77 vs 0.75) and the two had the same variant classification AUC (0.98). Whereas the more stringent model achieved the same cross-validation AUC as the model trained on the initial training set after removing cancer genes, both all were noticeably lower than the model trained on the initial training set with the earlier predictor variables (0.81). Therefore, I explored if other combinations of predictor variables perform better with the updated annotations and training set. A comparison of performance statistics is shown in (Table 4.2). Based on both cross-validation AUC and variant classification AUC, I selected the model that incorporates the predictor variables: 'GERP score of conserved non-coding elements', 'median size of spliced transcripts', 'identity to closest paralog' and 'embryonic expression' and 'proximity to other known HI genes in protein-protein interaction network', and I trained this model using the more stringent updated known HI gene set. The new predictive model achieved higher cross-validation AUC (0.86 vs 0.81) and similar variant classification AUC (0.96 vs 0.96) to the un-updated model, while improving prediction coverage without imputation (16,017 vs 12,443). However, when testing if genes found with LOF substitutions and indels in sequenced exomes have lower p(HI) than the genome background as did in Section 4.3.3.4, the difference was less significant despite still being in the same direction (0.13 vs 0.21, $p = 2.2\times10^{-12}$, Mann-Whitney test). The difference was even smaller when comparing p(HI) of genes found with LOF substitutions in a larger exome-sequencing dataset that consisted of $\sim$300 apparently healthy individuals (0.18 vs 0.21, $p = 4.5\times10^{-3}$, Mann-Whitney test). Thus although the cross-validation seems to indicate improved performance from this later model, the comparisons with external datasets of different types, does not back this up.

Table 4.2: Performance comparison of prediction models

HI training set	Predictors	#HI training	#Predictable	AUC_{CV}	AUC_{LOD}
Initial*	CNC_GERP PPI_LLS2HI	237	16,017	0.866	0.915
	CNC_GERP PPI_LLS2HI TRANS_SIZE PARALOG_DIST EARLY_DEV	237	16,017	0.869	0.945
	MACAQUE_DNDS PROMOTER_GERP EARLY_DEV GGI_LLS2HI	237	13,030	0.765	0.97
Expanded	CNC_GERP PPI_LLS2HI	312	16,017	0.864	0.934
	CNC_GERP PPI_LLS2HI TRANS_SIZE PARALOG_DIST EARLY_DEV	312	16,017	0.864	0.964
	MACAQUE_DNDS PROMOTER_GERP EARLY_DEV GGI_LLS2HI	312	13,030	0.765	0.975

* Cancer genes removed

4.3.3 Using HI gene predictions to assess pathogenicity of deletions

4.3.3.1 Defining a genomic-interval-based pathogenicity score

I investigated how my gene-based predictions of haploinsufficiency might be used to discriminate between benign and pathogenic genic deletions. I considered that a natural way to score the probability of a deletion of a genomic interval causing a haploinsufficiency phenotype is to generate a LOD (log-odds) score comparing the probability that none of the genes covered contained in the interval will cause haploinsufficiency with the probability that at least one of the genes will cause haploinsufficiency, as shown schematically in Figure 4.12. This LOD score is calculated using the formula below:

$$LOD = \ln\left(\frac{1 - \prod(1 - \mathrm{p(HI)})}{\prod(1 - \mathrm{p(HI)})}\right)$$

, and assumes that there is no statistical interaction between the genes. Worked examples of this calculation are shown in the figure below. Higher LOD scores indicate deletions are more likely to be pathogenic as a result of haploinsufficiency.

4.3.3.2 Discriminating benign and pathogenic deletions

I then considered how these deletion-based haploinsufficiency scores might be used to assess whether a genic deletion observed in a patient might cause their disease. One way of framing probabilistically this intuitively simple question is to estimate the opposing probability, that the deletion is unrelated to the patient's disease status. This can be equated to the probability of drawing an individual at random from a healthy control population with a deletion at least as pathogenic as the deletion in the patient. This probability can be estimated empirically as the proportion of healthy controls with a genic deletion having the same or greater haploinsufficiency LOD score.

To test this approach, and to avoid circular reasoning, I retained a subset (2,322 GWAS controls used in studies of schizophrenia and bipolar disease) of the 8,458 apparently healthy individuals from which the HS genes in the original training data were derived and generated a new set of p(HI) by training on the reduced HS gene set identified from the rest of apparently healthy individuals using the same method as described in Section 4.3.2. After imputation of predictor variables (see Methods), this new training set contains 287 HI genes

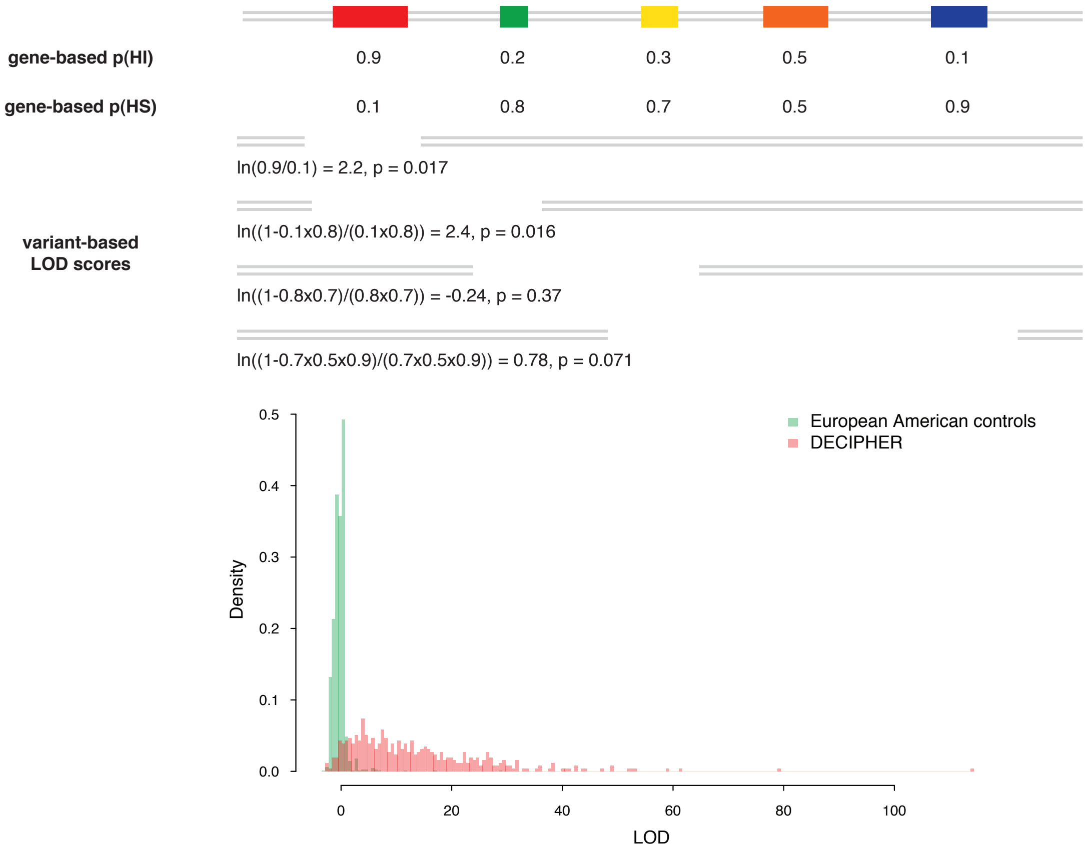

Figure 4.12: Calculation of deletion-based LOD scores and the distribution of LOD score of control individuals and pathogenic *de novo* deletions. The upper portion of the figure is a schematic demonstration of the calculation of the deletion-based LOD score. The contribution of genes with high p(HI) is accordingly weighted in a probabilistic way. The deletion with the largest LOD score in each individual is recorded and their distribution is shown in the lower portion of the figure. The distribution of maximal LOD scores of 2,322 control individuals are shown in green and the distribution of LOD scores of 487 pathogenic *de novo* deletions from DECIPHER are in red. Using the control distribution as the null, the probability a deletion is pathogenic can be assessed.

and 594 HS genes (234 HI genes and 270 HS genes before imputation). The model trained from this reduced training set achieved a similar AUC and MCC in 10-fold cross-validation as the model trained from the original training set (after imputation: AUC = 0.84, MCC = 0.55; before imputation: AUC = 0.81, MCC = 0.50).

The resulting predictions are also highly consistent with the original predictions (correlation between p(HI) is 0.99 both before and after imputation). I used the predictions based on the dataset that includes imputed predictor variables to allow the more reliable assertion of haploinsufficiency of a genomic interval from the vast majority of the genes affected by its deletion (17,456 genes with p(HI) after imputation as opposed to 12,443 before imputation). Based on these predictions I determined the distribution of the maximal deletion haploinsufficiency scores for the retained subset of 2,322 apparently healthy individuals.

To compare this distribution of 'most pathogenic' deletions discovered in apparently healthy individuals with truly pathogenic deletions, I collected 487 *de novo* deletions identified from array-based CNV detection and classified as being putatively pathogenic in the DECIPHER database [129]. I focused exclusively on deletions known to be *de novo* variants, as I infer that their pathogenicity has been ascribed primarily on the basis of their inheritance status, and not their gene content. The distributions of maximal LOD scores in GWAS controls and LOD scores of pathogenic DECIPHER deletions are shown in Figure 4.12. The pathogenic deletions have strikingly significantly higher LOD scores than deletions observed in GWAS controls ($p < 1\times10^{-30}$, Mann-Whitney U test). I observed that for 92% of the pathogenic deletions there was a probability of less than 5% of drawing an individual at random from our control population with a genic deletion of equal or greater LOD score, and for 83% of pathogenic deletions there was a less than 1% probability.

I computed ROC curves to compare three different approaches for discriminating between pathogenic deletions and deletions seen in controls: (*i*) LOD scores, (*ii*) the length of the deletion, and (*iii*) the number of genes in the deletion (Figure 4.13). These ROC curves clearly show that the haploinsufficiency LOD score is the best metric of the three for discriminating between pathogenic deletions in patients and deletions seen in controls.

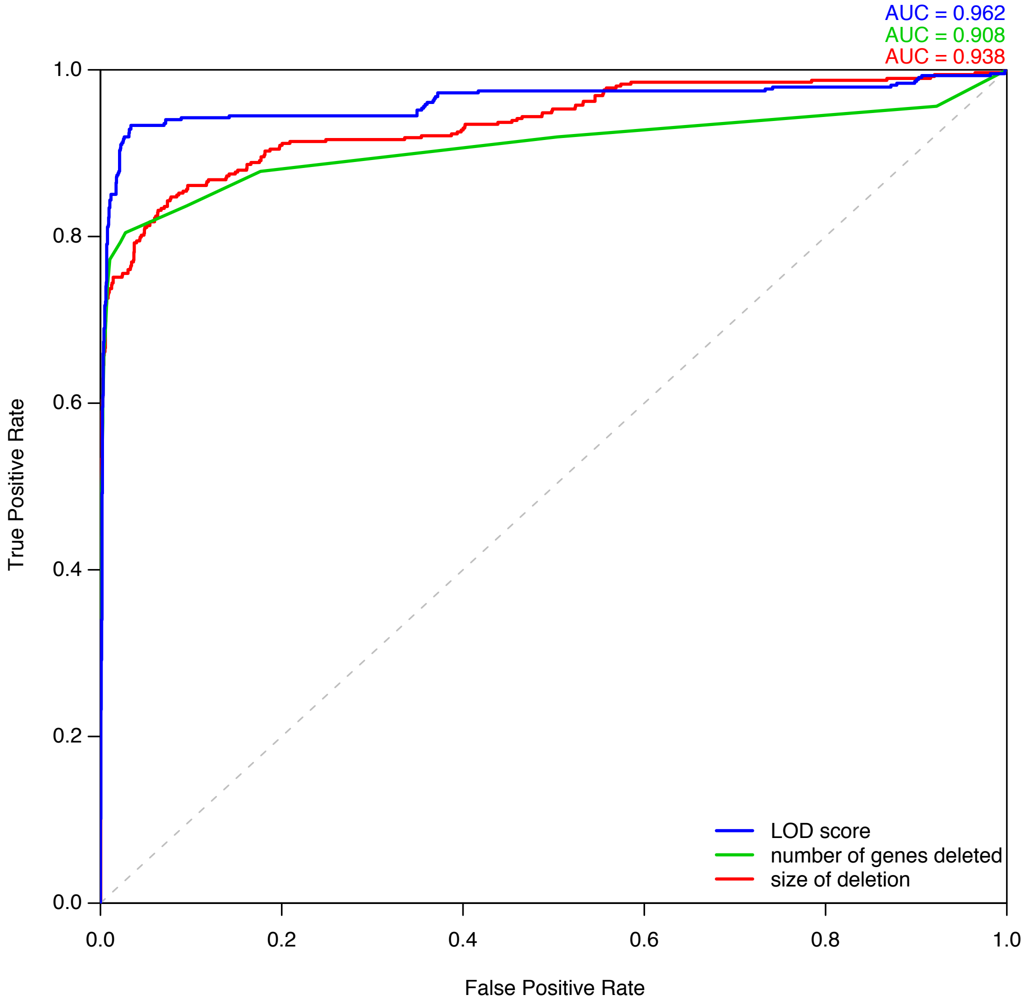

Figure 4.13: Comparison of different metrics for assessing deletion pathogenicity. Three ROC curves represent the performance of three different methods for distinguishing between pathogenic deletions from DECIPHER and the most pathogenic deletions observed in control individuals. The blue curve denotes using LOD score calculated from predicted probability of exhibiting haploinsufficiency as the metric of pathogenicity. The green curve denotes using the number of deleted genes as the metric, in which case the most pathogenic deletion per individual is the one containing greatest number of genes in that individual. The red curve denotes using the size of deletion as the discriminating metric.

I investigated whether the distribution of maximal LOD scores is significantly different between 1,433 European-Americans (EA) and 889 African-Americans (AA) GWAS controls, which, if true, might suggest the necessity of using ethnicity matched population pathogenicity score distributions. I observed that there was not a significant difference in median haploinsufficiency scores in EA and AA populations (p = 0.71, Mann-Whitney U test). The EA controls have a slightly longer tail of more pathogenic deletions (*e.g.* a higher proportion of EA controls have deletions with LOD scores in the top 1% in the pooled distribution, Table 4.3), which is consistent with the previous suggestion that purifying selection is more efficient in African populations due to their larger effective population sizes [158, 159]. However, this difference is again not significant (p = 0.24, Fisher's exact test).

Table 4.3: Population-specific properties of LOF CNVs

Population	Average #(LOF CNV) per individual	Average #(predictable LOF gene) per individual	Average #(LOF gene) in CNV with max LOD per individual	Average of max LOD per individual	Proportion with max LOD $\geq$ 99% of the pooled population
European American	7.41	10.5	2.85	-0.36	1.18%
African American	7.35	10.1	2.74	-0.38	0.79%

4.3.3.3 Extension to duplications

Since the probability of a gene being haploinsufficient partly reflects its general dosage sensitivity, it might be reasonable to expect abnormally increased dosage of at least some HI genes could also be pathogenic, as exemplified by the *PMP22* gene contained in the lim1.5Mb region at 17p11.2 of which duplication causes Charcot-Marie-Tooth syndrome type 1A and deletion causes Hereditary Neuropathy with Liability to Pressure Palsies. Therefore, I investigated if the interval-based haploinsufficiency LOD score could also be applied to classifying the pathogenicity of duplications. All computational procedures were identical to those for deletions, except the slight difference that the LOD scores for duplications were calculated from the p(HI) of genes contained in a genomic interval instead of LOF genes. I again compared the ROC curves of using LOD scores, the length of the duplication, and the number of genes in the duplication (Figure 4.14). The LOD score exhibit similar performance to the size of duplication in discriminating between pathogenic duplications and duplications seen in controls. Both LOD score and size performed better than the number of genes.

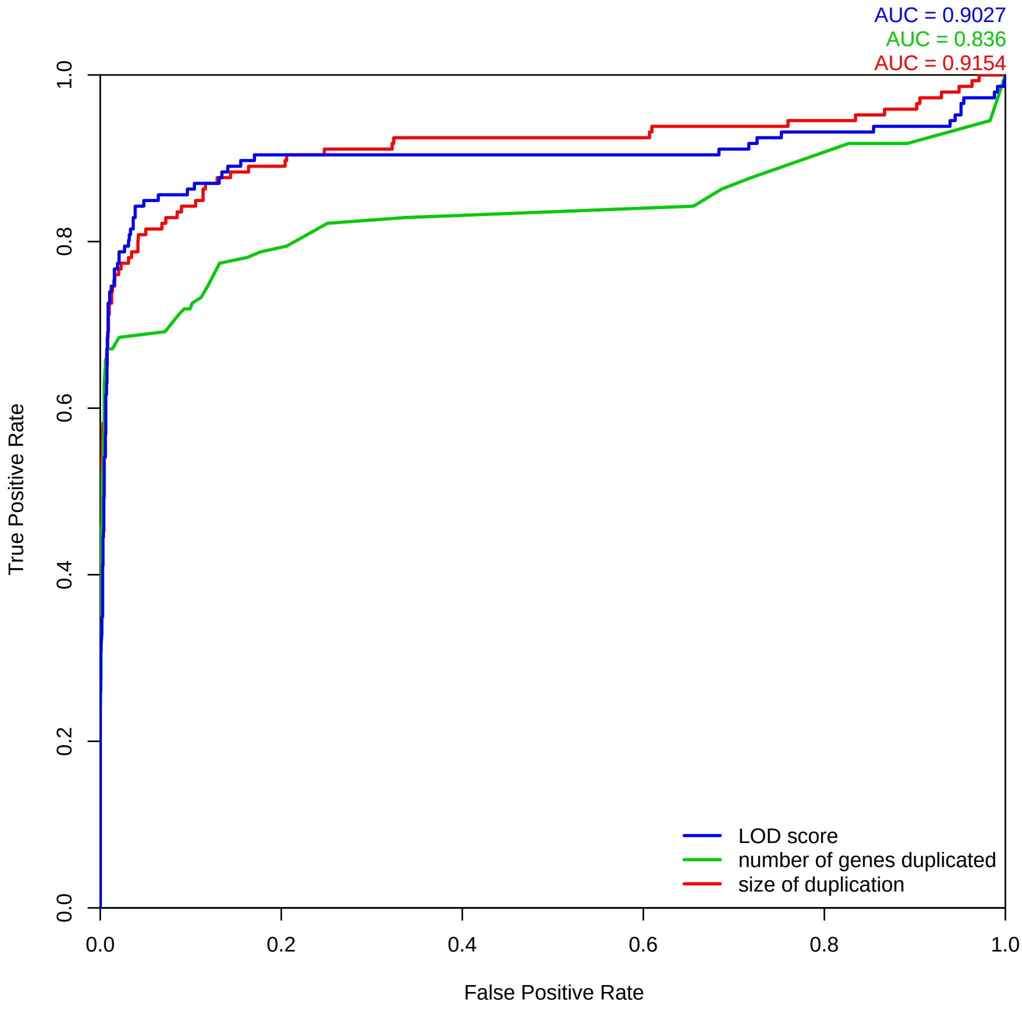

Figure 4.14: Comparison of different metrics for assessing duplication pathogenicity. Three ROC curves represent the performance of three different methods for distinguishing between pathogenic duplications from DECIPHER and the most pathogenic duplications observed in control individuals. The blue curve denotes using LOD score calculated from predicted probability of exhibiting haploinsufficiency as the metric of pathogenicity. The green curve denotes using the number of duplicated genes as the metric, in which case the most pathogenic duplication per individual is the one containing greatest number of genes in that individual. The red curve denotes using the size of duplication as the discriminating metric.

4.3.3.4 Extension to other forms of genetic variation

I investigated whether the gene-based probabilities of haploinsufficiency that I have generated are of general utility across different forms of genetic variation. If this is indeed the case then I should expect that genes harboring loss-of-function substitutions or small indels in apparently healthy individuals should not have a high p(HI). I identified 349 genes as having LOF substitutions and indels in 12 recently sequenced exomes [116], of which I could estimate p(HI) for 176 that were not also in the HS training set (and thus represent a fair set for independent comparisons). These genes are highly significantly enriched among genes with low probabilities of exhibiting haploinsufficiency ($p = 1.06 \times 10^{-20}$ when comparing to the genome, and $p < 1 \times 10^{-30}$ when comparing to known HI genes, Mann-Whitney U test). This result implies that there are not substantial differences between genes that tolerate whole gene deletions and those that tolerate smaller loss-of-function variants.

Moreover, by utilizing a large gene-resequencing dataset that contains 47,576 SNPs found by direct resequencing of 11,404 protein-coding genes in 35 individuals (20 European-Americans (EA) and 15 African-Americans (AA)) [160], I studied the allele frequency spectrum of different types of genic variants with respect to p(HI) of the genes. I hypothesized that genes under stronger negative selection should exhibit an enrichment of rare alleles in their allele frequency spectrum relative to genes under less selective constraint. There are 14,420 nonsynonymous SNPs and 16,213 synonymous SNPs in the dataset found within genes with predicted p(HI). I examined their derived allele frequency (DAF) spectrum as a function of p(HI) of the genes in which they are located (Figure 4.15).

Regardless of population composition, the DAF spectrum of nonsynonymous SNPs are significantly more skewed towards rare variants in gene sets with higher p(HI) than in those with lower p(HI), as assessed by a one-sided Mann-Whitney U test comparing the median of the allele frequency spectrum of nonsynonymous variants in genes with p(HI) in the top 20% with that of nonsynonymous variants in genes with p(HI) in the bottom 80%. The p value for this test in EA was 3.95×10^{-3}, and in AA was 2.85×10^{-7}. As a control, the difference in DAF of synonymous SNPs between high p(HI) genes and low p(HI) genes was not significant (EA $p = 0.127$, AA $p = 0.057$). These results suggest greater selective constraint on genes predicted to exhibit haploinsufficiency.

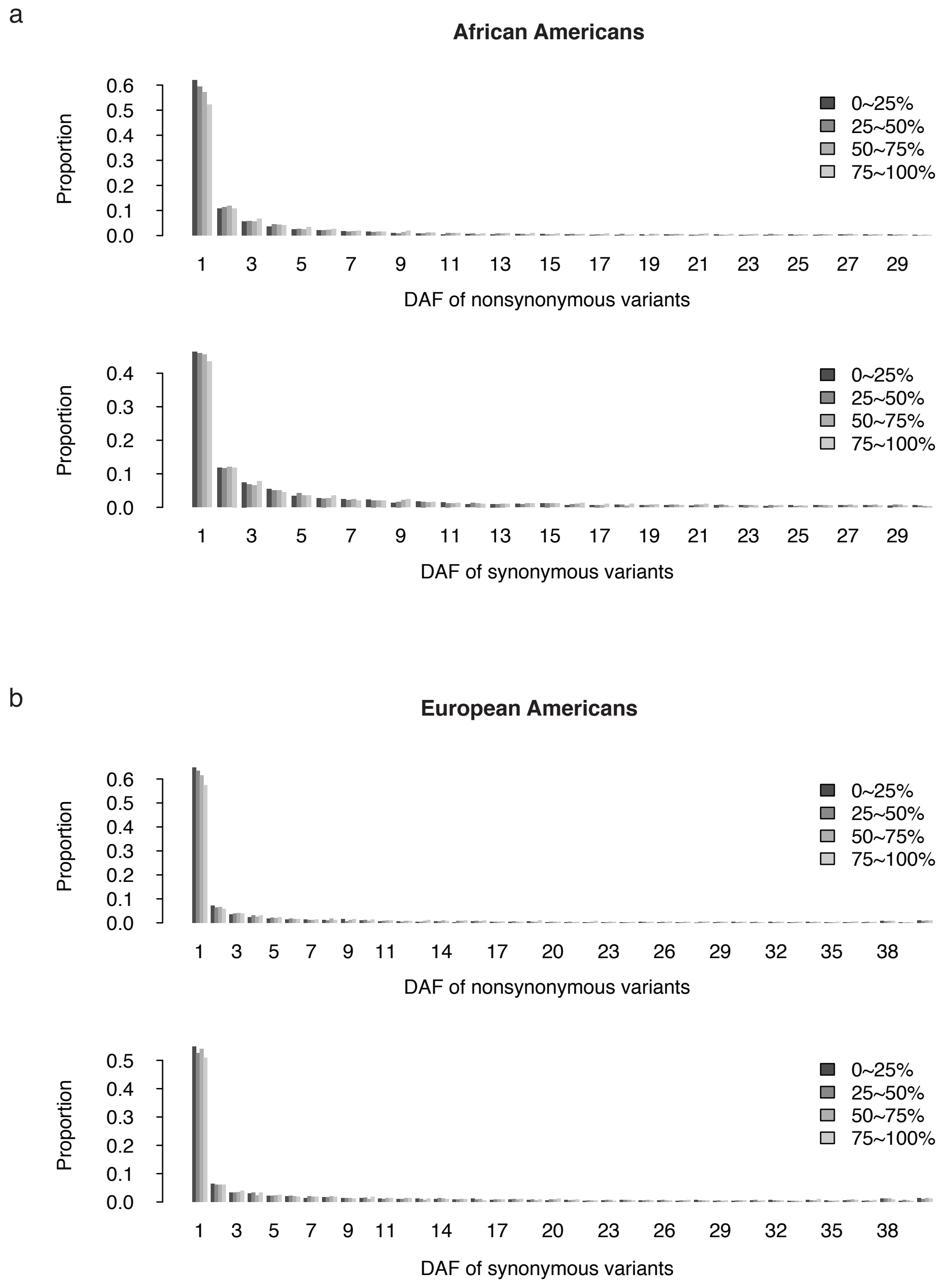

Figure 4.15: Derived allele frequency spectrum of variants in different gene sets. This figure shows the spectrum of derived allele frequency (DAF, represented here as counts of derived allele in the population) of nonsynonymous SNPs and synonymous SNPs discovered by resequencing of human genes in a) 15 African Americans and b) 20 European Americans. In each plot, DAF of variants located in genes of different p(HI) are compared side by side, where bars of decreasing darkness represent quantiles of decreasing p(HI), such that the 0–25% quartile is that with the highest probability of being haploinsufficient.

4.3.4 Probabilistic CNV diagnosis

In Section 4.3.2 and 4.3.3, I demonstrated the usefulness of gene-based p(HI) and its derivative, the interval-based haploinsufficiency LOD score in discriminating between benign and pathogenic deletions by showing that known pathogenic deletions have a LOD score distribution significantly higher than that of even the 'most deleterious' deletions found in apparently healthy individuals. However, in clinical diagnostics the primary question is how likely a variant is pathogenic/causal given all sources of evidence (*e.g.* pathogenic score). This is a typical Bayesian problem of which the answer is affected by both prior belief and evidence. Naturally, I modeled this problem using a Bayesian framework and tried to put it in the context of the general diagnostic process. I applied this framework to CNV diagnostics and examined two frequently encountered scenarios in clinical diagnostics wherein (*i*) the inheritance status of the variant is unknown or (*ii*) the variant is known to have arisen *de novo*.

4.3.4.1 A Bayesian framework for CNV diagnostics

The diagnostic question: 'is this variant, in this patient, sufficient to explain their clinical phenotype?' can be answered by assessing the posterior probability that this variant is causal given all the available evidence, $\mathrm{p}(C|E)$, where C denotes that the variant is causal and E denotes all available evidence. This probability is difficult to measure directly. Instead, the probability to observe such evidence given the variant is causal (and not causal), $\mathrm{p}(E|C)$ (and $\mathrm{p}(E|\bar{C})$), can be estimated directly from medical or population data and can be used to derive $\mathrm{p}(C|E)$ according to the Bayes Rule:

$$\mathrm{p}(C|E) = \frac{\mathrm{p}(C)\mathrm{p}(E|C)}{\mathrm{p}(E)} = \frac{\mathrm{p}(C)\mathrm{p}(E|C)}{\mathrm{p}(C)\mathrm{p}(E|C) + \mathrm{p}(\bar{C})\mathrm{p}(E|\bar{C})}$$

, where $\mathrm{p}(C)$ is the prior probability a variant is causal. Evidence involved in diagnosis of genetic variants includes both dichotomous or categorical conditions and continuous measurements. The former are often used as filters, such as 'overlapping with known disease-causing genes' and 'inherited from similarly affected parents'. The latter can be transformed into filters with defined thresholds, such as the division of common and rare variants based on population frequency thresholds, or used directly as numeric variables, such as pathogenic scores. Therefore, the space of evidence can be split into S, denoting that the variant has a measure of pathogenicity equal to x, and F, representing all other pieces of evidence that can be used as filters. In this way, the posterior probability and the Bayes factor becomes $\mathrm{p}(C|S,F)$ and $\mathrm{p}(S,F|C)$, respectively. The latter can be further expanded to $\mathrm{p}(F|C)\mathrm{p}(S|C,F)$, so that

$$\mathrm{p}(C|S,F) = \frac{\mathrm{p}(C)\mathrm{p}(F|C)\mathrm{p}(S|C,F)}{\mathrm{p}(C)\mathrm{p}(F|C)\mathrm{p}(S|C,F) + \mathrm{p}(\bar{C})\mathrm{p}(F|\bar{C})\mathrm{p}(S|\bar{C},F)}$$

, or in its likelihood ratio form,

$$LR = \frac{\mathrm{p}(C|S,F)}{\mathrm{p}(\bar{C}|S,F)} = \frac{\mathrm{p}(C)\mathrm{p}(F|C)\mathrm{p}(S|C,F)}{\mathrm{p}(\bar{C})\mathrm{p}(F|\bar{C})\mathrm{p}(S|\bar{C},F)}$$

$$\mathrm{p}(C|S,F) = \frac{LR}{1+LR}$$

$\mathrm{p}(F|C)$ (or $\mathrm{p}(F|\bar{C})$) is the probability the variant passes this filter F given the variant is causal (or benign), and $\mathrm{p}(S|C,F)$ (or $p(S|\bar{C},F)$) is the probability of the variant having a measure of pathogenicity equals to x given it is causal (or benign) and passes the filter F.

$\mathrm{p}(F|C)$ can be estimated as the proportion of causal variants discovered in large patient studies that pass the filter, and $\mathrm{p}(S|C,F)$ can be estimated as the proportion of causal variants passing the filter that have a pathogenic measure equal to x. $\mathrm{p}(F|\bar{C})$ and $\mathrm{p}(S|\bar{C},F)$ are best estimated from all benign variants, from both patients and healthy individuals. In practice, benign variants are usually not reported and recorded in patient studies, and depending on the particular filter, F, such information is sometimes not collected for variants found in population-based or control studies (*e.g.* whether a variant is *de novo* or not). Therefore, $\mathrm{p}(F|\bar{C})$ and $\mathrm{p}(S|\bar{C},F)$ often have to be estimated from approximate distributions. Variants found in control individuals should be similar enough to all benign variants provided the sample size of the control cohort is large. For certain filters, the set of variants that pass them may be obtained through proxy properties. After the approximate variant sets are constructed, $\mathrm{p}(F|\bar{C})$ and $\mathrm{p}(S|\bar{C},F)$ can be estimated as for causal variants. With different F, these components need to be estimated from different sets of variants and the posterior probability changes accordingly. Below I consider two categories of possibly causal variant that are frequently encountered in clinical diagnostics: (*i*) the variant can be shown to be rare, but is of known inheritance status, and (*ii*) the variant can be shown to be *de novo*.

Table 4.4: Estimated parameters of the diagnostic framework

Variant type	Size range	F	p(C)	p($F|C$)	p($F|\bar{C}$)
deletion	>180k	rare	0.12	1	0.34
deletion	>180k	*de novo*	0.12	0.73	0.0044
duplication	>330k	rare	0.14	1	0.38
duplication	>330k	*de novo*	0.14	0.73	0.0044

4.3.4.2 The variant is rare, and of unknown inheritance status

Under this scenario, often the only information on the variant is that it is not already known to be pathogenic and is not commonly seen in the population, therefore F denotes the filter that requires variants to be rare as defined by having a population frequency <1%. The estimated value of the parameters: $p(C)$, $p(F|C)$ and $p(F|\bar{C})$ were listed in Table 4.4 (see Methods). I considered either the LOD score or the variant size as the measure of pathogenicity. The distribution of LOD scores and variant sizes for rare casual and benign CNVs, from which $p(S|C, F)$ and $p(S|\bar{C}, F)$ can be calculated, were shown in Figure 4.16–4.19. For both deletions and duplications, the resulting posterior probability $p(C|S, F)$ increases as the LOD score, or the size of the variant, becomes greater. In order to achieve a confidence level of 95%, a rare deletion of unknown inheritance status needs to be larger than 2.1Mb or have a LOD score greater than 7.2, and a rare duplication needs to be larger than 3.2Mb or with a LOD score greater than 15.5.

4.3.4.3 The variant is *de novo*

The *de novo* rate of causal and benign CNVs is even harder to obtain as confirming the *de novo* status would require the genotype information of both the parents and the child, *i.e.* the 'trio', and reaching a reasonable estimate requires genotyping a large number of such trios. There are a few studies that have reported CNV diagnosis in hundreds to more than a thousand patients including parents in which low-resolution array-CGH were used to detect large CNVs and *de novo* status were confirmed where possible [134, 161]. These studies are arguably the best sources from which one can estimate the *de novo* rate of causal CNVs. However, even with this data the number of *de novo* CNV from any one study is

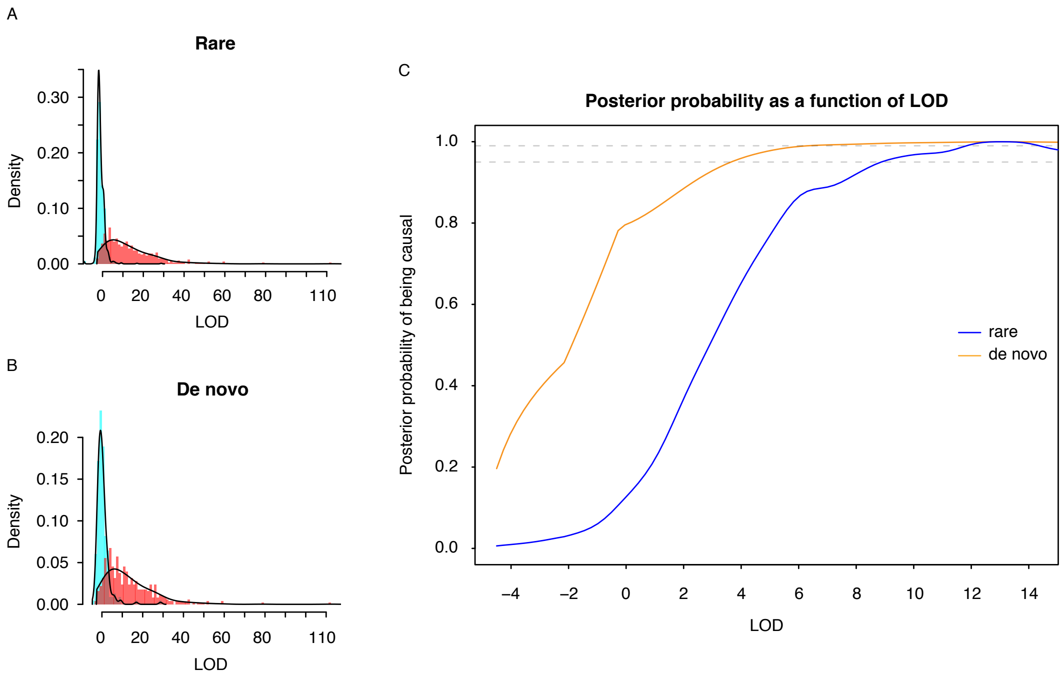

Figure 4.16: The posterior probability of a deletion being causal as a function of pathogenicity score. The distribution of pathogenicity score for causal (red) and benign (green) deletions are shown in A and B. In C, the two horizontal dashed lines represent posterior probabilities of 0.95 and 0.99.

too small to generate a meaningful distribution of measure of pathogenicity. Therefore, the distribution of pathogenicity measures for *de novo* CNVs was approximated using known *de novo* causal CNVs recorded in DECIPHER. Studies reporting *de novo* CNVs discovered in apparently healthy individuals are even scarcer. I took the benign *de novo* rate from Itsara *et al*, which investigated the rate of *de novo* CNVs in 772 transmissions in pedigrees without neurocognitive disease genotyped on median- to high-resolution SNP genotyping arrays and I approximated the distribution of pathogenicity scores for benign *de novo* CNVs using singleton CNVs found in WTCCC2 and GAIN controls.

The estimated values of the parameters are show in Table 4.4 and the causal and benign distributions of measure of pathogenicity are shown in Figure 4.16–4.19. As expected, for both deletions and duplications, the size or the LOD score required for a variant to have a probability of being causal greater than 0.95 is much smaller than that required for a variant of which the inheritance status is unknown. However, being *de novo* alone does not guarantee pathogenicity as the probability of being a causal variant is still not convincingly high when the variant is small (0.8 at size = 500kb) or with very low LOD score (0.7 at LOD = -2).

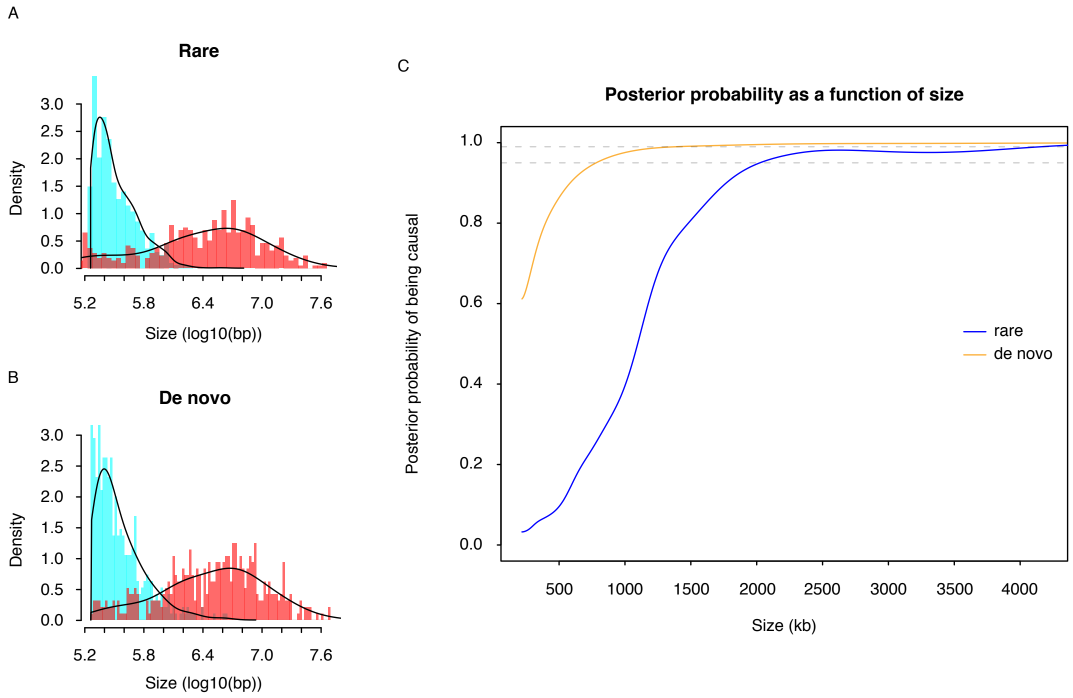

Figure 4.17: The posterior probability of a deletion being causal as a function of size. The distribution of pathogenicity score for causal (red) and benign (green) deletions are shown in A and B. In C, the two horizontal dashed lines represent posterior probabilities of 0.95 and 0.99.

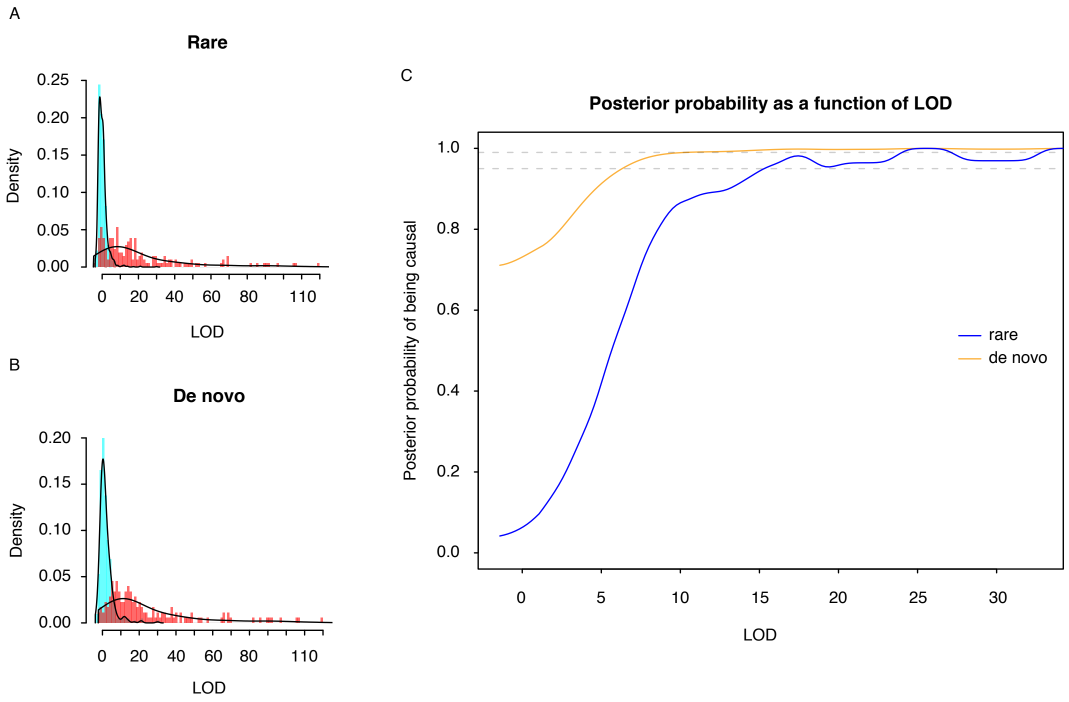

Figure 4.18: The posterior probability of a duplication being causal as a function of pathogenicity score. The distribution of pathogenicity score for causal (red) and benign (green) duplications are shown in A and B. In C, the two horizontal dashed lines represent posterior probabilities of 0.95 and 0.99.

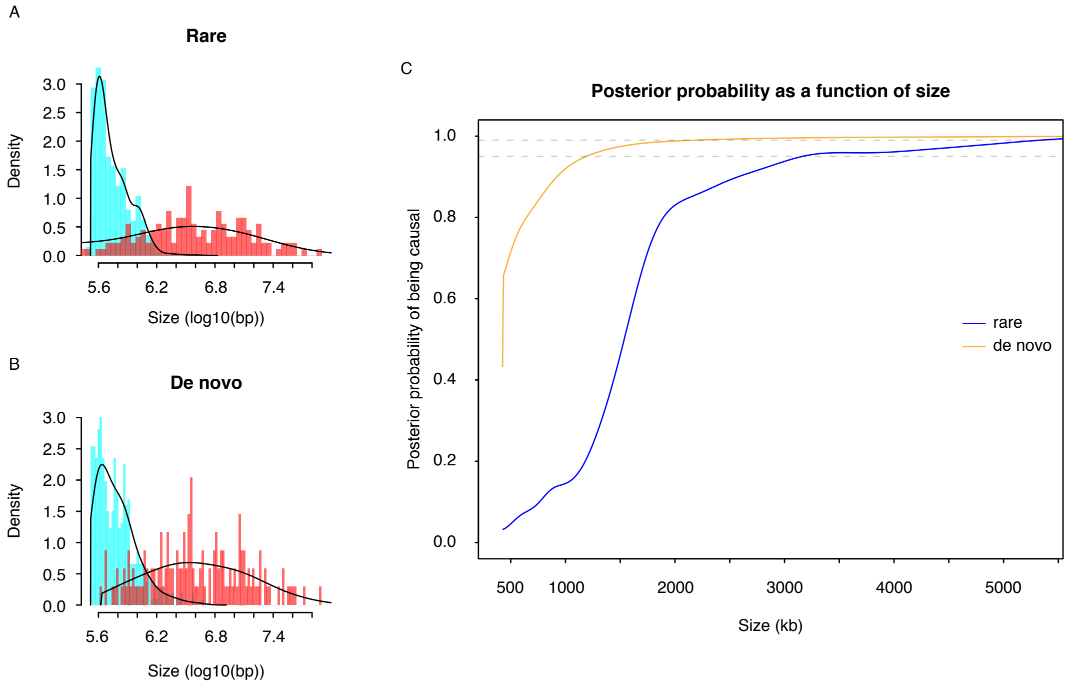

Figure 4.19: The posterior probability of a deletion being causal as a function of size. The distribution of pathogenicity score for causal (red) and benign (green) deletions are shown in A and B. In C, the two horizontal dashed lines represent posterior probabilities of 0.95 and 0.99.

4.4 Discussion

In this chapter, I described the collection of human HI genes and HS genes, their differences in genomic, evolutionary, functional and network properties, and a computational method that distinguishes the two and predicts the probability of exhibiting haploinsufficiency for human protein-coding genes of unknown dosage sensitivity. A measure of pathogenicity for large genic copy number variants was developed on the basis of the HI predictions. A probabilistic diagnostic framework was designed to transform evidence of pathogenicity of a patient variant into confidence of diagnosis by taking into account the population variance of that measure of pathogenicity.

The traditional view that recessiveness is the norm of deleterious mutations is supported by earlier mutagenesis screen of model organisms [162]. In human, the $\sim$300 known HI genes only account for $\sim$1.5% of the protein-coding genome. However, haploinsufficiency, like most concepts in Mendelian genetics, is a qualitative, rather than quantitative, description based on a phenotype-specific definition of insufficiency, Insensitive or incomplete phenotyping or diagnosis could lead to underestimation of the proportion of the genome that is actually dosage sensitive. In genetics studies of model organisms, it is common that only the most prominent phenotypic consequence of a mutation or traits that are in relation with certain prior expectation are examined and reported. Abnormalities that are subtle and require specially designed tests to reveal or occur in completely unexpected tissues or cells can often be overlooked. Even in human, wherein measurements of physiological and morphological abnormalities is thought to be much more sensitive and thorough, complete phenotyping is never guaranteed. For example, the mutant allele of the gene *GJB2*, which is causal for the most frequent form of recessive congenital hearing loss, was recently found responsible for increased epidermal thickness in a dominant or semi-dominant manner [163, 164]. Thickened epiderm is obviously a less prominent trait that could not be detected without skin ultrasonography or similar technologies. In this chapter, the definition of haploinsufficiency has focused on severe clinical phenotypes (broadly-defined) as sufficiency relates to being qualified to be recruited as an apparently healthy control in a study of common disease susceptibility. With more complete phenotyping and hence a more stringent definition of sufficiency, the haploinsufficient/dosage-sensitive proportion of the genome might grow larger. In addition, most early work of Fisher, Wright and others that emphasized the dominance of the wildtype allele focused on metabolic enzymes. We now know that metabolic enzymes are less likely to be haploinsufficient whereas transcription factors, structural proteins and subunits of protein complexes are more likely to be haploinsufficient due to the kinetic properties of the respective molecular system in which they function [107, 165, 166]. As transcription factors alone account 5–10% of the human protein-coding genome [167], the currently $\sim$300 known human HI genes is likely just a tip of the iceberg.

Not surprisingly, the known HI genes were found to be larger in size, which is a general characteristic of disease genes [168, 169], though it might be attributed to ascertainment bias, as, all things being equal, it is easier to find multiple families with causal mutations in the same gene if the gene is larger. HI genes were found to be more conserved in their coding sequence than HS genes, which is consistent with previous comparison between dominant and recessive disease genes [119]. In addition, the promoter sequences of HI genes are more conserved as well, which might suggest transcription regulation of these genes, as a part of dosage control mechanism, is under greater purifying selection, although this needs to be confirmed by human variation data. HI genes were found to have fewer paralogs and/or paralogs with lower sequence similarity than HS genes. This is consistent with a yeast study [150] which reported that HI genes tend not to have paralogs and suggested having a close paralog may provide a buffer against the effects of haploinsufficiency, but contradicts another report by Kondrashov *et al* [120] that found human dominant disease genes tend to have more paralogs than recessive disease genes and argued that such is the result of positive selection. However, the latter finding is not strictly comparable to this study, since homozygous LOF mutation of recessive disease genes can cause severe phenotypic defects and are hence under selection and less likely to be found in large genomic deletions, from which the HS gene set used in this study are collected. Indeed, there is a significant underrepresentation ($p = 0.0023$) of recessive disease genes in the HS gene set. The strong enrichment of olfactory receptor genes in the HS set (13% compared to 2% genome-wide, $p < 2.2\times10^{-16}$) could also affect the result. With respect to their spatiotemporal expression patterns, HI genes are more tissue specific and active during early development, which is expected since many of the haploinsufficient transcription factors play vital and tissue specific role in early developmental processes such as patterning, morphogenesis and organ development [156]. As for network properties, HI genes are found to be more central and closer to one another. The latter may support the view that haploinsufficiency tend to occur in certain molecular systems (early-development-related signaling and transcription regulation pathways, protein complexes), but may also be confounded by the ascertainment bias that search for novel disease genes tends to follow interaction partners of known disease genes.

The prediction of HI was implemented by training a statistical classifier on known HI and HS genes using gene properties that best distinguish the two as predictor variables. This is not a strictly mechanism-based approach, but an approach that exploits the correlation between haploinsufficiency and other gene properties. Though the performance of the prediction, as assessed by cross-validation using the training data, is moderately good (AUC = 0.81 without imputation of predictors and 0.84 with imputation; when requiring 80% sensitivity, the version without imputation has 70% specificity and version with imputation has 75% specificity), it is better than using any single gene property alone and has been validated to be able to prioritize potential real genes. Proximity to other known HI genes within gene or protein networks was found to be the most predictive property of which the contribution to performance cannot be fully explained by sequence conservation, tissue-specificity of expression, or other gene properties. Incomplete coverage of all genes in the genome by gene-gene and protein-protein networks is therefore also the major factor limiting the genome coverage of these predictions. The predictions should be substantially improved in both accuracy and coverage with the future generation of more complete and accurate human genetic interaction networks.

Although haploinsufficiency can be regarded as the property of a single gene, phenotypic manifestation of any genetic mutation, including heterozygous LOF mutation, is, strict speaking, the output of a perturbed multi-layer system of interacting molecules, cells and organs. Consequently, dosage sensitivity of a gene could vary across different genetic backgrounds. For example, heterozygous deletion of *Tbx5* causes embryonic lethality in 129S mice, but produces viable mice on B6 background [170]. In humans, patients carrying a second large CNV in addition to the micro-deletion at 16p12.1 exhibit much severer developmental delay than those having the 16p12.1 micro-deletion alone [91]. Therefore, the ideal prediction of haploinsufficiency should come from a system biology approach that models all interacting genes and biochemical reactions in a cell mathematically similar to that of Kacers and Burns [107] in which haploinsufficiency could be determined by numeric simulation and single component sensitivity analysis.

The measure of pathogenicity of a CNV was defined as the log of odds that at least one affected gene is haploinsufficient. As the likelihoods of being haploinsufficient of individual genes are combined in such a probabilistic way, its application is not limited to individual genomic intervals. For example, one can measure the genome-wide pathogenic burden of an individual by calculating the odds that at least one gene is haploinsufficient out of all genes affected by any CNV, or other LOF variants, in this individual's genome. However, there are also obvious caveats of such measure. First, the measure can only be applied to CNVs affecting protein-coding genes for which a prediction is available. Second, potential functions of intergenic sequences are ignored. Third, the effects of each gene are assumed to be independent. To tackle the first two limitations, one could consider features that are not bound to genes, such as the density of repeat elements or the number of conserved non-coding elements. However, these properties need to be combined with the likelihoods of genes exhibiting haploinsufficiency in a meaningful way. For the third caveat, the idea solution would again be a system biology approach as described above, substituting the single-component sensitivity analysis with a multiple-component sensitivity analysis.

The probabilistic diagnostic framework provides a natural way to integrate both qualitative and quantitative measures of pathogenicity and produces quantified confidence of diagnosis by considering the population variance of the quantitative measure of pathogenicity. Being a Bayesian method, it has the advantage of not naively assuming that different measures are independent, but at the same time it requires knowledge of the conditional distribution of the quantitative measure of pathogenicity, which is not always readily available. In its application to rare and *de novo* CNVs in Section 4.3.4, the patient and control distribution of pathogenicity score under the condition that the CNVs are *de novo* were unavailable and were substituted with approximated distributions. Another problem, which is common for all Bayesian inferences, is the requirement of a proper prior. The prior probability of a variant being causal can be affected by a number of factors, for example, the specific type of disease and the filters or tests having been applied before the application of this framework. As for CNVs, since different CNV discovery platforms vary vastly in their sensitivity and resolution, which could have profound impact on the population distribution of the quantitative measure of pathogenicity, the prior should be estimated from the same or similar platform that the population distribution of pathogenicity scores is generated.

Previously, *de novo* CNVs discovered in patients were highly likely to be diagnosed as being a causal variant in clinical practice. As early CNV discovery technologies, such as cytogenetic methods and low-resolution array CGH could only find very large events, those diagnoses might largely hold correct. However, in recent years, with improved CNV discovery technology and accumulated CNV datasets, it is known that *de novo* CNVs, especially smaller ones, arise at an appreciable rate (estimates ranging from 1×10^{-2} to 3×10^{-2} CNVs per haploid genome per generation [89, 131, 171]) in healthy individuals. Therefore, there is growing recommendation for not relying solely on the *de novo* status in the interpretation of variant causality [133, 172]. My application of this diagnostic framework to *de novo* CNVs not only supported this view, but also provides a quantitative level of confidence as a function of the size of the variant or its pathogenicity score. However, these quantitative values should be interpreted with caution at this stage, and are not mature enough for clinical implantation, for several reasons. First, as the distribution of CNVs and functional sequences is uneven across the genome whereas the size or the pathogenicity score of CNVs are locus-independent measures. In addition, these results are highly dependent on the CNV discovery platform and the prior. Furthermore, the use of approximate conditional distributions of pathogenicity scores has introduced additional uncertainty. With the increasing application of array CGH, high-resolution genotyping array and medical sequencing, and hence ascertainment of a more complete spectrum of variants in patient genomes, this diagnostic framework is expected to produce a more accurate estimation of confidence to aid the diagnosis of novel, rare variants for which detailed locus-specific information is unavailable.

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 genome-wide 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 sequence-based 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.

REFERENCES

[1] Hurles, M. E., Dermitzakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends Genet* **24**, 238–45 (2008).

[2] Lupski, J. R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* **1**, e49 (2005).

[3] Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–53 (2007).

[4] Millar, J. K. *et al.* Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum Mol Genet* **9**, 1415–23 (2000).

[5] Nakata, K. *et al.* DISC1 splice variants are upregulated in schizophrenia and associated with risk polymorphisms. *Proc Natl Acad Sci U S A* **106**, 15873–8 (2009).

[6] Millar, J. K. *et al.* DISC1 and PDE4B are interacting genetic factors in schizophrenia that regulate cAMP signaling. *Science* **310**, 1187–91 (2005).

[7] Shen, S. *et al.* Schizophrenia-related neural and behavioral phenotypes in transgenic mice expressing truncated Disc1. *J Neurosci* **28**, 10893–904 (2008).

[8] Kirov, G. *et al.* De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry* (2011).

[9] Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–72 (2010).

[10] Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat Genet* **43**, 838–46 (2011).

[11] International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–41 (2008).

[12] Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–12 (2010).

[13] Wellcome Trust Case Control Consortium *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–20 (2010).

[14] Dang, V. T., Kassahn, K. S., Marcos, A. E. & Ragan, M. A. Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. *Eur J Hum Genet* **16**, 1350–7 (2008).

[15] Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**, 1256–60 (2007).

[16] Johnson, M. E. *et al.* Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–9 (2001).

[17] Zhang, J., Zhang, Y.-p. & Rosenberg, H. F. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat Genet* **30**, 411–5 (2002).

[18] Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40**, 1253–60 (2008).

[19] Affymetrix. Affymetrix genome-wide human SNP array 6.0 data sheet. `http://media.affymetrix.com/support/technical/datasheets/genomewide_snp6_datasheet.pdf` (2009).

[20] Affymetrix. Manuals of Affymtrix Power Tools. `http://media.affymetrix.com/support/developer/powertools/changelog/index.html` (2009).

[21] Price, T. S. *et al.* SW-ARRAY: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic Acids Res* **33**, 3455–64 (2005).

[22] Veltman, J. A. *et al.* High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am J Hum Genet* **70**, 1269–76 (2002).

[23] Wang, P., Kim, Y., Pollack, J., Narasimhan, B. & Tibshirani, R. A method for calling gains and losses in array CGH data. *Biostatistics* **6**, 45–58 (2005).

[24] Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–72 (2004).

[25] Jong, K., Marchiori, E., Meijer, G., Vaart, A. V. D. & Ylstra, B. Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics* **20**, 3636–7 (2004).

[26] Hupé, P., Stransky, N., Thiery, J.-P., Radvanyi, F. & Barillot, E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413–22 (2004).

[27] Pique-Regi, R. *et al.* Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* **24**, 309–18 (2008).

[28] Fridlyand, J., Snijders, A., Pinkel, D., Albertson, D. & Jain, A. Hidden Markov models approach to the analysis of array CGH data. *Journal of Multivariate Analysis* **90**, 132–153 (2004).

[29] Marioni, J. C., Thorne, N. P. & Tavaré, S. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics* **22**, 1144–6 (2006).

[30] Guha, S., Li, Y. & Neuberg, D. Bayesian hidden Markov Modeling of array CGH data. *Journal of the American Statistical Association* **103**, 485–497 (2008).

[31] Shah, S. P. *et al.* Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics* **22**, e431–9 (2006).

[32] Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* **35**, 2013–25 (2007).

[33] Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665–74 (2007).

[34] McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**, 1166–74 (2008).

[35] Marioni, J. C. *et al.* Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol* **8**, R228 (2007).

[36] The HDF5 group. Hierarchical data format version 5. `http://www.hdfgroup.org/HDF5` (2009).

[37] Alted, F., Vilata, I. *et al.* PyTables: hierarchical datasets in Python. `http://www.pytables.org` (2002).

[38] Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–54 (2006).

[39] Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* **40**, 575–83 (2008).

[40] Zöllner, S. CopyMap: localization and calling of copy number variation by joint analysis of hybridization data from multiple individuals. *Bioinformatics* **26**, 2776–7 (2010).

[41] Williams, N. M. *et al.* Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *Lancet* **376**, 1401–8 (2010).

[42] Whitby, H. *et al.* Benign copy number changes in clinical cytogenetic diagnostics by array CGH. *Cytogenet Genome Res* **123**, 94–101 (2008).

[43] Bell, C. G., Walley, A. J. & Froguel, P. The genetics of human obesity. *Nat Rev Genet* **6**, 221–34 (2005).

[44] Rennie, K. L. & Jebb, S. A. Prevalence of obesity in Great Britain. *Obes Rev* **6**, 11–2 (2005).

[45] Allison, D. B. *et al.* The heritability of body mass index among an international sample of monozygotic twins reared apart. *Int J Obes Relat Metab Disord* **20**, 501–6 (1996).

[46] Maes, H. H., Neale, M. C. & Eaves, L. J. Genetic and environmental factors in relative body weight and human adiposity. *Behav Genet* **27**, 325–51 (1997).

[47] Stunkard, A. J., Harris, J. R., Pedersen, N. L. & McClearn, G. E. The body-mass index of twins who have been reared apart. *N Engl J Med* **322**, 1483–7 (1990).

[48] Chen, H. *et al.* Evidence that the diabetes gene encodes the leptin receptor: identification of a mutation in the leptin receptor gene in db/db mice. *Cell* **84**, 491–5 (1996).

[49] Huszar, D. *et al.* Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell* **88**, 131–41 (1997).

[50] Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**, 425–32 (1994).

[51] Cummings, D. E. & Schwartz, M. W. Genetics and pathophysiology of human obesity. *Annu Rev Med* **54**, 453–71 (2003).

[52] Walley, A. J., Asher, J. E. & Froguel, P. The genetic contribution to non-syndromic human obesity. *Nat Rev Genet* **10**, 431–42 (2009).

[53] Clément, K. *et al.* A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* **392**, 398–401 (1998).

[54] Krude, H. *et al.* Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. *Nat Genet* **19**, 155–7 (1998).

[55] Montague, C. T. *et al.* Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* **387**, 903–8 (1997).

[56] Farooqi, I. S. & O'Rahilly, S. Monogenic obesity in humans. *Annu Rev Med* **56**, 443–58 (2005).

[57] Farooqi, I. S. *et al.* Beneficial effects of leptin on obesity, T cell hyporesponsiveness, and neuroendocrine/metabolic dysfunction of human congenital leptin deficiency. *J Clin Invest* **110**, 1093–103 (2002).

[58] Goldstone, A. P. Prader-Willi syndrome: advances in genetics, pathophysiology and treatment. *Trends Endocrinol Metab* **15**, 12–20 (2004).

[59] O'Rahilly, S. & Farooqi, I. S. Human obesity: a heritable neurobehavioral disorder that is highly sensitive to environmental conditions. *Diabetes* **57**, 2905–10 (2008).

[60] Farooqi, I. S. *et al.* Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. *N Engl J Med* **348**, 1085–95 (2003).

[61] Comuzzie, A. G. *et al.* A major quantitative trait locus determining serum leptin levels and fat mass is located on human chromosome 2. *Nat Genet* **15**, 273–6 (1997).

[62] Mitchell, B. D. *et al.* A quantitative trait locus influencing BMI maps to the region of the beta-3 adrenergic receptor. *Diabetes* **48**, 1863–7 (1999).

[63] Kissebah, A. H. *et al.* Quantitative trait loci on chromosomes 3 and 17 influence phenotypes of the metabolic syndrome. *Proc Natl Acad Sci U S A* **97**, 14478–83 (2000).

[64] Walder, K., Hanson, R. L., Kobes, S., Knowler, W. C. & Ravussin, E. An autosomal genomic scan for loci linked to plasma leptin concentration in Pima Indians. *Int J Obes Relat Metab Disord* **24**, 559–65 (2000).

[65] Dong, C. *et al.* Interacting genetic loci on chromosomes 20 and 10 influence extreme human obesity. *Am J Hum Genet* **72**, 115–24 (2003).

[66] Hunt, S. C. *et al.* Linkage of body mass index to chromosome 20 in Utah pedigrees. *Hum Genet* **109**, 279–85 (2001).

[67] Frayling, T. M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–94 (2007).

[68] Dina, C. *et al.* Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat Genet* **39**, 724–6 (2007).

[69] Hinney, A. *et al.* Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *PLoS One* **2**, e1361 (2007).

[70] Scuteri, A. *et al.* Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* **3**, e115 (2007).

[71] Loos, R. J. F. *et al.* Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* **40**, 768–75 (2008).

[72] Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* **41**, 25–34 (2009).

[73] Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42**, 937–48 (2010).

[74] McDermid, H. E. & Morrow, B. E. Genomic disorders on 22q11. *Am J Hum Genet* **70**, 1077–88 (2002).

[75] Boerkoel, C. F., Inoue, K., Reiter, L. T., Warner, L. E. & Lupski, J. R. Molecular mechanisms for CMT1A duplication and HNPP deletion. *Ann N Y Acad Sci* **883**, 22–35 (1999).

[76] Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38**, 75–81 (2006).

[77] Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949–51 (2004).

[78] Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–8 (2004).

[79] Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78–88 (2005).

[80] Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727–32 (2005).

[81] McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* **40**, 1107–12 (2008).

[82] Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–40 (2005).

[83] McKinney, C. *et al.* Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on susceptibility to rheumatoid arthritis. *Ann Rheum Dis* **67**, 409–13 (2008).

[84] McCarthy, S. E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* **41**, 1223–7 (2009).

[85] Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–6 (2008).

[86] Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–43 (2008).

[87] Glessner, J. T. *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459**, 569–73 (2009).

[88] Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* **82**, 477–88 (2008).

[89] Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–9 (2007).

[90] Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* **358**, 667–75 (2008).

[91] Girirajan, S. *et al.* A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet* **42**, 203–9 (2010).

[92] Bochukova, E. G. *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**, 666–70 (2010).

[93] Barnes, C. *et al.* A robust statistical method for case-control association testing with copy number variation. *Nat Genet* **40**, 1245–52 (2008).

[94] Raychaudhuri, S. *et al.* Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet* **6** (2010).

[95] Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–50 (2005).

[96] Rankinen, T. *et al.* The human obesity gene map: the 2005 update. *Obesity (Silver Spring)* **14**, 529–644 (2006).

[97] Walters, R. G. *et al.* A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* **463**, 671–5 (2010).

[98] Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat Genet* **36**, 512–7 (2004).

[99] Iype, T. *et al.* The transcriptional repressor Nkx6.1 also functions as a deoxyribonucleic acid context-dependent transcriptional activator during pancreatic beta-cell differentiation: evidence for feedback activation of the nkx6.1 gene by Nkx6.1. *Mol Endocrinol* **18**, 1363–75 (2004).

[100] Schisler, J. C. *et al.* Stimulation of human and rat islet beta-cell proliferation with retention of function by the homeodomain transcription factor Nkx6.1. *Mol Cell Biol* **28**, 3465–76 (2008).

[101] Thomas, M. A., Preece, D. M. & Bentel, J. M. Androgen regulation of the prostatic tumour suppressor NKX3.1 is mediated by its 3′ untranslated region. *Biochem J* **425**, 575–83 (2010).

[102] Jacquemont, S. *et al.* Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature* **478**, 97–102 (2011).

[103] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–9 (2006).

[104] Heid, I. M. *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet* **42**, 949–60 (2010).

[105] Fisher, R. A. *The genetical theory of natural selection* (The Clarendon press, Oxford, 1930).

[106] Wright, S. *Evolution and the genetics of populations: a treatise* (University of Chicago Press, Chicago, 1968). URL `http://www.loc.gov/catdir/description/uchi051/67025533.html`.

[107] Kacser, H. & Burns, J. A. The molecular basis of dominance. *Genetics* **97**, 639–66 (1981).

[108] Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res* **37**, D793–6 (2009).

[109] Wilkie, A. O. The molecular basis of genetic dominance. *J Med Genet* **31**, 89–98 (1994).

[110] Veitia, R. A. Exploring the etiology of haploinsufficiency. *Bioessays* **24**, 175–84 (2002).

[111] Willing, M. C. *et al.* Osteogenesis imperfecta type I: molecular heterogeneity for COL1A1 null alleles of type I collagen. *Am J Hum Genet* **55**, 638–47 (1994).

[112] Ebert, B. L. *et al.* Identification of RPS14 as a 5q- syndrome gene by RNA interference screen. *Nature* **451**, 335–9 (2008).

[113] Devriendt, K. *et al.* Haploinsufficiency of the HOXA gene cluster, in a patient with hand-foot-genital syndrome, velopharyngeal insufficiency, and persistent patent Ductus botalli. *Am J Hum Genet* **65**, 249–51 (1999).

[114] Maquat, L. E. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat Rev Mol Cell Biol* **5**, 89–99 (2004).

[115] Vissers, L. E. L. M. *et al.* Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat Genet* **36**, 955–7 (2004).

[116] Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–6 (2009).

[117] Ng, P. C. *et al.* Genetic variation in an individual human exome. *PLoS Genet* **4**, e1000160 (2008).

[118] Xue, Y. *et al.* Adaptive evolution of UGT2B17 copy-number variation. *Am J Hum Genet* **83**, 337–46 (2008).

[119] Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr Biol* **18**, 883–9 (2008).

[120] Kondrashov, F. A. & Koonin, E. V. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet* **20**, 287–90 (2004).

[121] Nguyen, D.-Q., Webber, C. & Ponting, C. P. Bias of selection on human copy-number variants. *PLoS Genet* **2**, e20 (2006).

[122] Thomas, P. D. & Kejariwal, A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci U S A* **101**, 15398–403 (2004).

[123] Tavtigian, S. V., Byrnes, G. B., Goldgar, D. E. & Thomas, A. Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. *Hum Mutat* **29**, 1342–54 (2008).

[124] Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073–81 (2009).

[125] Yue, P., Melamud, E. & Moult, J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* **7**, 166 (2006).

[126] Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30**, 3894–900 (2002).

[127] Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* **7**, 61–80 (2006).

[128] Hehir-Kwa, J. Y. *et al.* Accurate distinction of pathogenic from benign CNVs in mental retardation. *PLoS Comput Biol* **6**, e1000752 (2010).

[129] Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* **84**, 524–33 (2009).

[130] Stenson, P. D. *et al.* The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum Genomics* **4**, 69–72 (2009).

[131] Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* **84**, 148–61 (2009).

[132] Zhang, J., Feuk, L., Duggan, G. E., Khaja, R. & Scherer, S. W. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet Genome Res* **115**, 205–14 (2006).

[133] Breckpot, J. *et al.* Challenges of interpreting copy number variation in syndromic and non-syndromic congenital heart defects. *Cytogenet Genome Res* **135**, 251–9 (2011).

[134] Buysse, K. *et al.* Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. *Eur J Med Genet* **52**, 398–403 (2009).

[135] Goldgar, D. E. *et al.* Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. *Am J Hum Genet* **75**, 535–44 (2004).

[136] Hubbard, T. J. P. *et al.* Ensembl 2009. *Nucleic Acids Res* **37**, D690–7 (2009).

[137] Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901–13 (2005).

[138] Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062–7 (2004).

[139] Assou, S. *et al.* A meta-analysis of human embryonic stem cells transcriptome integrated into a web-based expression atlas. *Stem Cells* **25**, 961–73 (2007).

[140] Smith, C. M. *et al.* The mouse Gene Expression Database (GXD): 2007 update. *Nucleic Acids Res* **35**, D618–23 (2007).

[141] Brown, K. R. & Jurisica, I. Online predicted human interaction database. *Bioinformatics* **21**, 2076–82 (2005).

[142] Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res* **35**, D572–4 (2007).

[143] Keshava Prasad, T. S. *et al.* Human Protein Reference Database–2009 update. *Nucleic Acids Res* **37**, D767–72 (2009).

[144] Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–8 (2005).

[145] Vastrik, I. *et al.* Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* **8**, R39 (2007).

[146] Lee, I., Li, Z. & Marcotte, E. M. An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae. *PLoS One* **2**, e988 (2007).

[147] Lee, I. *et al.* A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nat Genet* **40**, 181–8 (2008).

[148] Van Dongen, S. Graph clustering via a discrete uncoupling process. *Siam Journal On Matrix Analysis and Applications* **30**, 121–141 (2008).

[149] Forbes, S. *et al.* COSMIC 2005. *Br J Cancer* **94**, 318–22 (2006).

[150] Deutschbauer, A. M. *et al.* Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169**, 1915–25 (2005).

[151] Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874 (2006).

[152] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16**, 412–424 (2000).

[153] Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Software* **45**, 1–67 (2011).

[154] Seal, R. L., Gordon, S. M., Lush, M. J., Wright, M. W. & Bruford, E. A. genenames.org: the HGNC resources in 2011. *Nucleic Acids Res* **39**, D514–9 (2011).

[155] Tsuruoka, Y. *et al.* Developing a robust part-of-speech tagger for biomedical text. *Advances In Informatics, Proceedings* **3746**, 382–392 (2005).

[156] Seidman, J. G. & Seidman, C. Transcription factor haploinsufficiency: when half a loaf is not enough. *J Clin Invest* **109**, 451–5 (2002).

[157] Jensen, L. J. *et al.* STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37**, D412–6 (2009).

[158] Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–7 (2005).

[159] Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994–7 (2008).

[160] Boyko, A. R. *et al.* Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4**, e1000083 (2008).

[161] Pickering, D. L. *et al.* Array-based comparative genomic hybridization analysis of 1176 consecutive clinical genetics investigations. *Genet Med* **10**, 262–6 (2008).

[162] Hrabé de Angelis, M. H. *et al.* Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat Genet* **25**, 444–7 (2000).

[163] D'Adamo, P. *et al.* Does epidermal thickening explain GJB2 high carrier frequency and heterozygote advantage? *Eur J Hum Genet* **17**, 284–6 (2009).

[164] Guastalla, P. *et al.* Detection of epidermal thickening in GJB2 carriers with epidermal US. *Radiology* **251**, 280–6 (2009).

[165] Hurst, L. D. & Randerson, J. P. Dosage, deletions and dominance: simple models of the evolution of gene expression. *J Theor Biol* **205**, 641–7 (2000).

[166] Orr, H. A. A test of Fisher's theory of dominance. *Proc Natl Acad Sci U S A* **88**, 11413–5 (1991).

[167] Magrane, M. & Uniprot Consortium. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, bar009 (2011).

[168] López-Bigas, N. & Ouzounis, C. A. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res* **32**, 3108–14 (2004).

[169] Smith, N. G. C. & Eyre-Walker, A. Human disease genes: patterns and predictions. *Gene* **318**, 169–75 (2003).

[170] Bruneau, B. G. *et al.* A murine model of holt-oram syndrome defines roles of the t-box transcription factor tbx5 in cardiogenesis and disease. *Cell* **106**, 709–21 (2001).

[171] Xu, B. *et al.* Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* **40**, 880–5 (2008).

[172] Vermeesch, J. R., Balikova, I., Schrander-Stumpel, C., Fryns, J.-P. & Devriendt, K. The causality of de novo copy number variants is overestimated. *Eur J Hum Genet* **19**, 1112–3 (2011).

APPENDIX A

TABLE OF MANUALLY CURATED HI GENES

Class definition:

1. Severe dominant developmental disorder, smaller critical region, genes disrupted in multiple patients or in a single patient with strong additional evidence.
2. Severe dominant developmental disorder, genes disrupted in a (large) deletion interval in a single patient, with (or without) other evidence.
3. Recessive disorder, or multigenic diseases resulting from compound mutations, or cancer which requires LOF or compound heterozygosity.
4. Evidence failed to support haploinsufficiency.

Gene	Class	Gene	Class	Gene	Class	Gene	Class
ACVRL1	1	AFF3	1	ALX4	1	ANKRD11	1
APC	1	ATM	1	ATP1A2	1	ATP2A2	1
ATP2C1	1	BDNF	1	BMPR2	1	BRCA2	1
BUB1B	1	CAMTA1	1	CBFB1	1	CD2AP	1
CDKN2A	1	CDX2	1	CHD2	1	CHD7	1
CHRNA7	1	CNTNAP2	1	COL1A1	1	COL3A1	1
COL5A1	1	CREEBP	1	CYP11A1	1	DMPK	1

continued on next page

continued from previous page

Gene	Class	Gene	Class	Gene	Class	Gene	Class
DSPP	1	DYRK1A	1	EHMT1	1	ELN	1
ELOVL4	1	EXT2	1	EYA1	1	FBN1	1
FGF10	1	FGF3	1	FGF8	1	FGFR1	1
FGFR2	1	FLG	1	FOXC1	1	FOXE3	1
FOXF1	1	FOXG1	1	FOXL2	1	FOXP2	1
GATA4	1	GDF5	1	GLI3	1	GNB1L	1
GNRH1	1	GPR98	1	GRN	1	GTF2I	1
GTF2IRD1	1	HMGA2	1	HNF1B	1	HOXD13	1
IGF1R	1	INSR	1	IRF6	1	ITPR1	1
JAG1	1	KCNAB2	1	KRIT1	1	KRT14	1
LEMD3	1	LMX1B	1	MBD5	1	MBP	1
MEF2C	1	MSX1	1	MSX2	1	NF1	1
NFIA	1	NFIB	1	NFKB1	1	NKX2-1	1
NKX2-6	1	NOTCH1	1	NR2F2	1	NR5A1	1
NSD1	1	OPA1	1	PAG1	1	PAX2	1
PAX3	1	PAX5	1	PAX6	1	PBX1	1
PHOX2B	1	PITX2	1	PKD1	1	PRKAR1A	1
PRKD2	1	PTCH1	1	PTEN	1	RAI1	1
RALGAPA1	1	ROR2	1	RPS14	1	RPS19	1
RUNX2	1	SALL1	1	SCN1A	1	SERPINA6	1
SGCE	1	SHH	1	SHOX	1	SLC2A1	1
SMAD5	1	SOCS1	1	SOX10	1	SOX2	1
SOX9	1	SRGAP3	1	STK11	1	STXBP1	1
SUMO1	1	TAB2	1	TBX3	1	TBX5	1
TBX6	1	TECTA	1	TIMM23	1	TTF1	1
TWIST1	1	UFD1L	1	WT1	1	ZEB2	1
ABCA3	2	ADAR	2	ANXA7	2	APAF1	2
ARFGAP1	2	ATR	2	ATXN1	2	BAZ1B	2
BLM	2	BMP4	2	BRCA1	2	BUB3	2
C10orf11	2	CASK	2	CDC73	2	CDKL3	2
CDKN1B	2	CDKN1C	2	CELF2	2	CHEK1	2
CHRNA4	2	CNTN4	2	COL6A1	2	COMT	2
COPS3	2	CSH1	2	CTCF	2	CYFIP1	2
DFFB	2	DLL4	2	DMRT1	2	DOCK1	2
DSG1	2	EFNB2	2	EGR1	2	EME1	2

continued on next page

continued from previous page

Gene	Class	Gene	Class	Gene	Class	Gene	Class
ENG	2	ERBB4	2	ETV6	2	FECH	2
FLCN	2	FLII	2	FOXC2	2	FOXL1	2
FOXO3	2	FZD4	2	GATA3	2	GDNF	2
GNAS	2	GPC1	2	GPR35	2	HFE	2
HIC1	2	HIRA	2	HNF1A	2	ID2	2
IGF1	2	IGFBP3	2	ITGB6	2	KCNQ1	2
KCNQ2	2	KHDRBS1	2	KIAA2022	2	KLF4	2
KLF6	2	LETM1	2	MAGOH	2	MAP3K4	2
MC4R	2	MED15	2	MITF	2	MLL	2
MNX1	2	MSH2	2	MSH6	2	MUS81	2
MYCN	2	MYF6	2	MYH9	2	NBN	2
NCF1	2	NF2	2	NFRKB	2	NKX2-5	2
NKX3-1	2	NOG	2	NPAS3	2	NPM1	2
NR2F1	2	NUP98	2	P2RY12	2	PAFAH1B1	2
PAX9	2	PCGF2	2	PDX1	2	PML	2
PRM2	2	QKI	2	RAD50	2	RAD51L1	2
RAE1	2	RB1	2	REEP1	2	RET	2
RNF135	2	RUNX1	2	SALL3	2	SCN2A	2
SCN5A	2	SEMA5A	2	SHANK3	2	SIM1	2
SIX6	2	SLC1A2	2	SLC4A11	2	SMARCB1	2
STK25	2	TACR1	2	TBX1	2	TCOF1	2
TERT	2	TNXB	2	TOB	2	TP53	2
TP73	2	TSC2	2	TUBGCP5	2	WHSC1	2
WNT2B	2	YWHAG	2	ZIC2	2	ADCY9	3
ARFGAP3	3	CADM1	3	CAV1	3	CFH	3
CHN2	3	COL11A1	3	COL1A2	3	CYP2A6	3
DGKD	3	DMRT2	3	DNAJC3	3	EXOC6B	3
FADD	3	FAS	3	FAT1	3	FGFR3	3
FKBP6	3	FOXO1	3	GGCX	3	GPD2	3
GPSM2	3	H2AX	3	IFI44	3	IKZF1	3
IL3RA	3	IRF8	3	KCNRG	3	KLF2	3
LIMK1	3	MAP4K2	3	NGF	3	NR5A2	3
NTNG2	3	PACRG	3	PAPSS2	3	PAX1	3
PINK1	3	PTHLH	3	RHOBTB1	3	RNF139	3
RPS4X	3	SERPIND1	3	SHFM1	3	SLC26A4	3

continued on next page

continued from previous page

Gene	Class	Gene	Class	Gene	Class	Gene	Class
SMAD4	3	SPI1	3	SPINK5	3	SRL	3
SSSCA1	3	SUFU	3	TNS3	3	UTP6	3
VAV3	3	WRD	3	ANGPT2	4	ATP2B2	4
COPS5	4	CRX	4	CTCFL	4	ENOSF1	4
FBXW7	4	FEN1	4	GPR172A	4	GSK3B	4
HAND2	4	HOXA1	4	IGFALS	4	MCL	4
MUTYH	4	MYOC	4	NR3C2	4	P2RY8	4
PARK7	4	PPARG	4	PRG2	4	PRODH	4
PROKR2	4	PXMP2	4	RELA	4	RORA	4
SH3PXD2A	4	SMN2	4	SMS	4	SOCS2	4
SOX18	4	SPAST	4	STAT5A	4	STK19	4
SUB1	4	TAS2R38	4	TCF7L2	4	TGIF1	4
TRPS1	4	XRCC5	4				