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

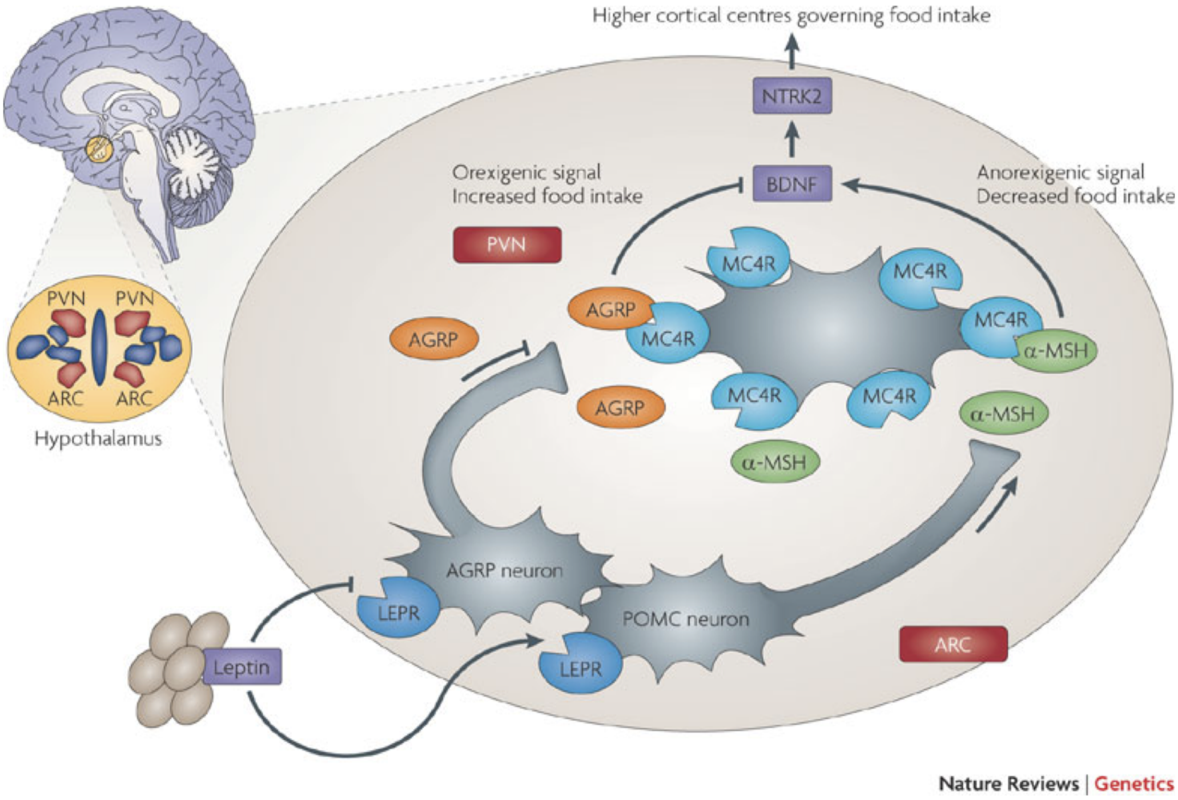

Figure 3.1: The leptin-melanocortin pathway. ARC: arcuate nucleus; PVN: paraventricular nucleus. Figure taken from Walley *et al* [52].

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).

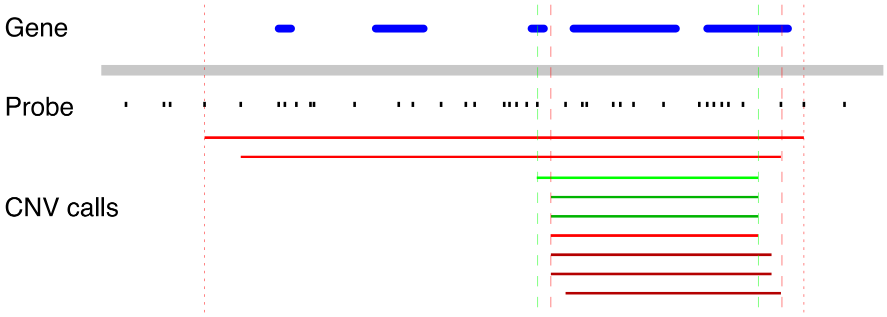

Figure 3.2: Illustration of the unit of test. Red horizontal lines represent deletions and green horizontal lines represent duplications. Control CNVs are in darker colors. Green dashed lines mark the CNVE clustered from duplications. Red dashed lines mark the CNVE clustered from the smaller deletions. 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

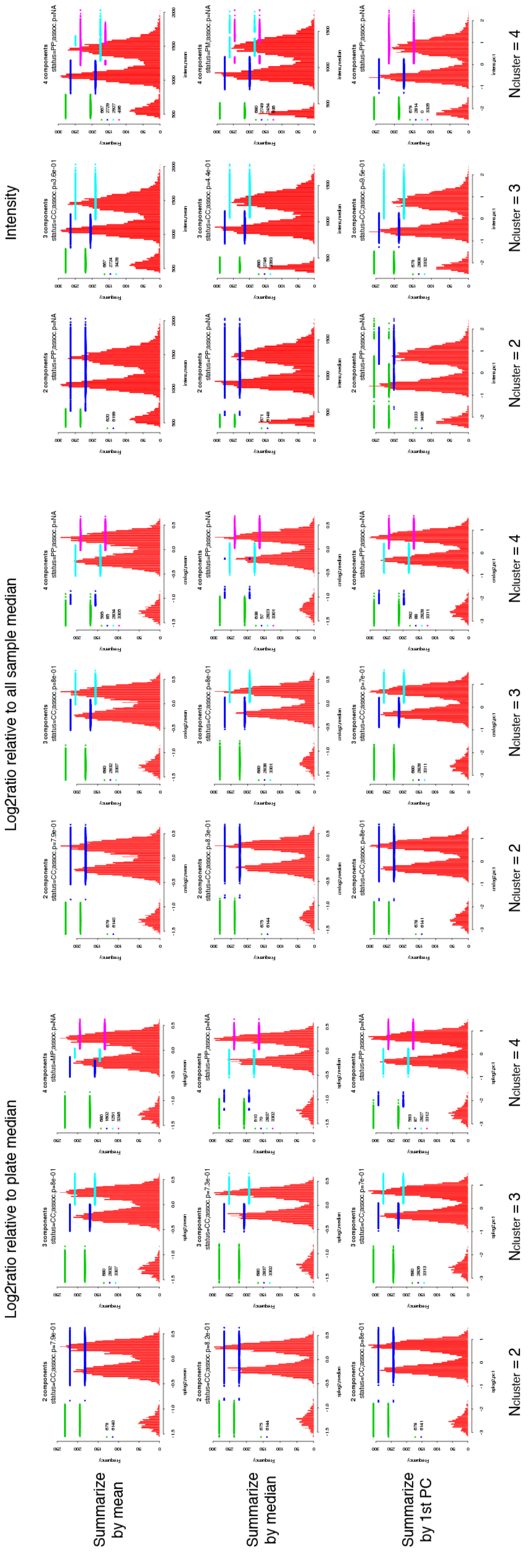

Figure 3.3: An example of test results of one common CNV loci generated by CNVtools under combinations of probe measurement, summarization method and preset number of genotype clusters. These results were subjected to manual examination.

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.

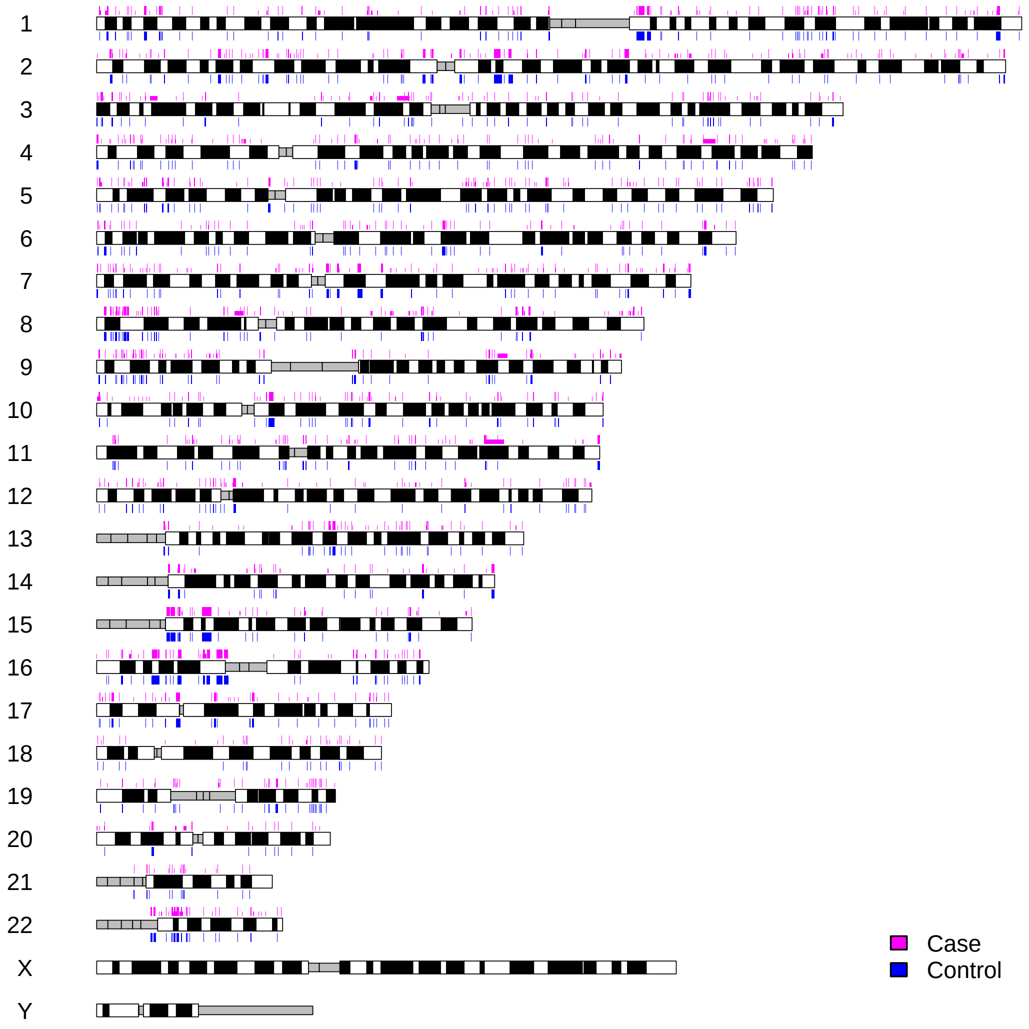

Figure 3.4: Overview of rare CNVs. The lengths of the colored rectangles represent the size of CNVs whereas the heights distinguish recurrent CNVs and singletons by which the former is taller than the latter. No CNV is displayed on chromosome X and Y since only autosomal calls were kept.

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	$1.4{\times}10^{-3}$	1
3q28	193,437	193,452	16	del	2	0	$1.4{\times}10^{-3}$	1
6p12.1	52,875	52,892	17	del	2	0	$1.4{\times}10^{-3}$	1
8q24.3	143,250	143,600	350	dup	2	0	$1.4{\times}10^{-3}$	2
9q31.1	106,401	106,407	5	dup	2	0	$1.4{\times}10^{-3}$	1
10p15.3	432	877	445	dup	2	0	$1.4{\times}10^{-3}$	3
10q11.23	52,980	52,985	5	del	2	0	$1.4{\times}10^{-3}$	1
11q13.4	71,980	72,107	126	dup	2	0	$1.4{\times}10^{-3}$	2
22q13.33	49,246	49,349	103	dup	2	0	$1.4{\times}10^{-3}$	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 ∼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 ∼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.

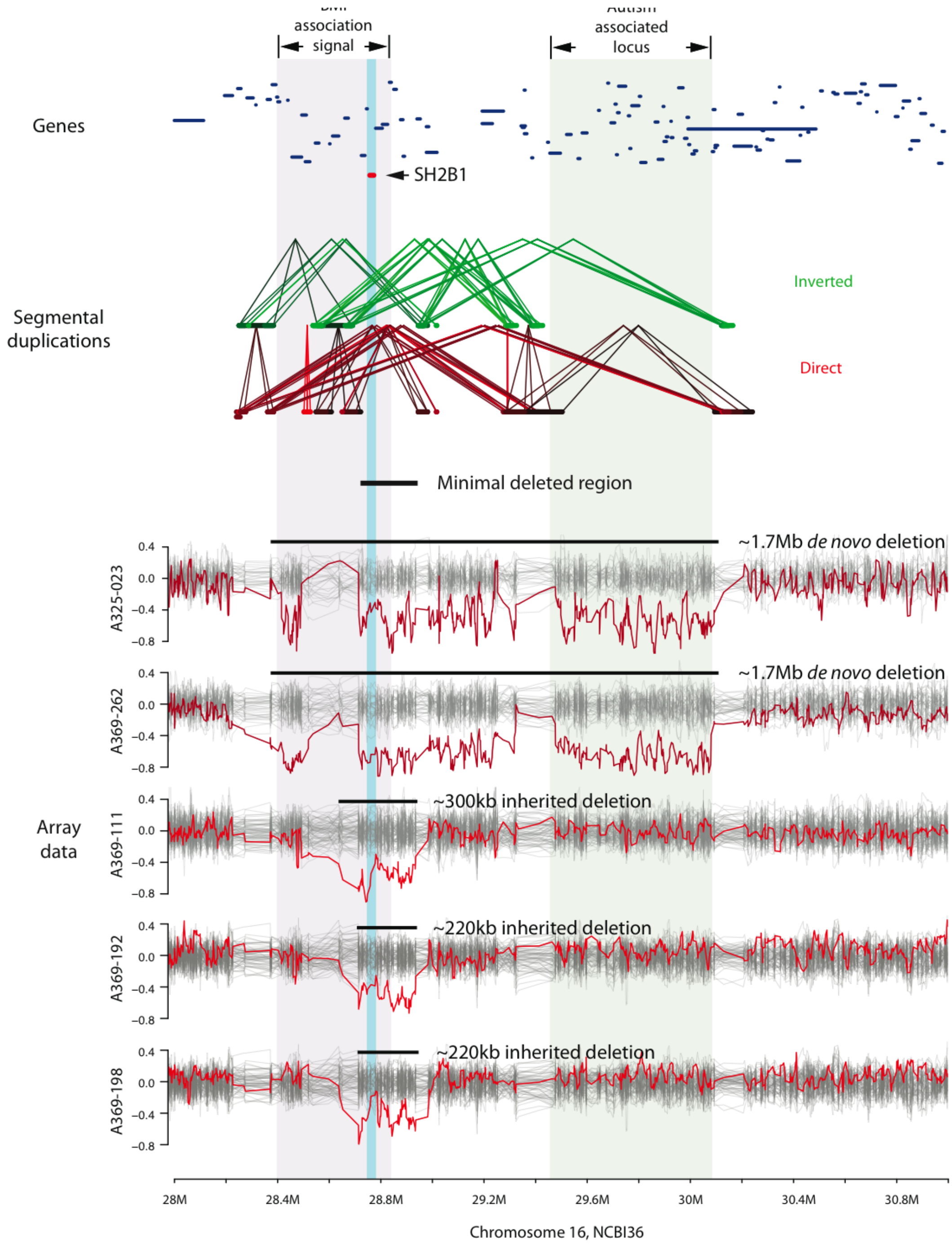

Figure 3.5: Deletions at 16p11.2 overlapping *SH2B1*. Affymetrix 6.0 array data for five patients with deletions at 16p11.2 is shown. Log_2 ratios of the five samples are highlighted in dark red with other samples in the same genotyping plate shown in grey. Annotation of the segmental duplications was 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.

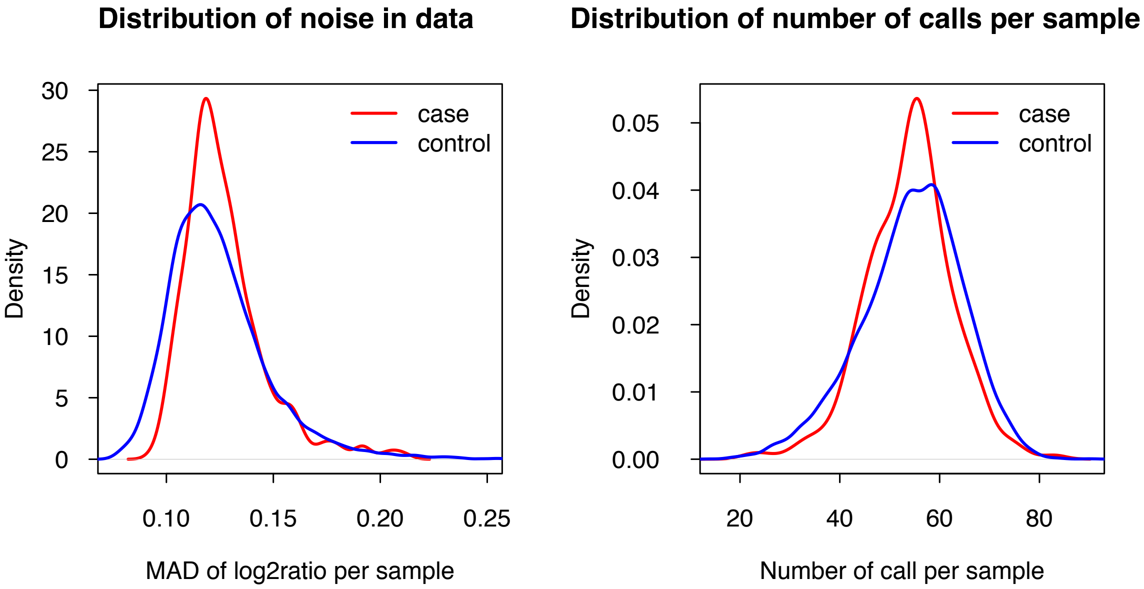

Figure 3.6: Comparison of level of noise and number of CNV calls per sample between cases and controls. Case samples have higher level of noise than controls ($p = 1.2 \times 10^{-3}$, Mann-Whitney test), leading to slightly greater number of calls per sample, though such difference is insignificant (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 ($p = 5\times10^{-4}$, Fisher's exact test). A stronger three-fold enrichment is observed in cases with developmental delay in addition to severe early onset obesity ($p = 3\times10^{-4}$), 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).

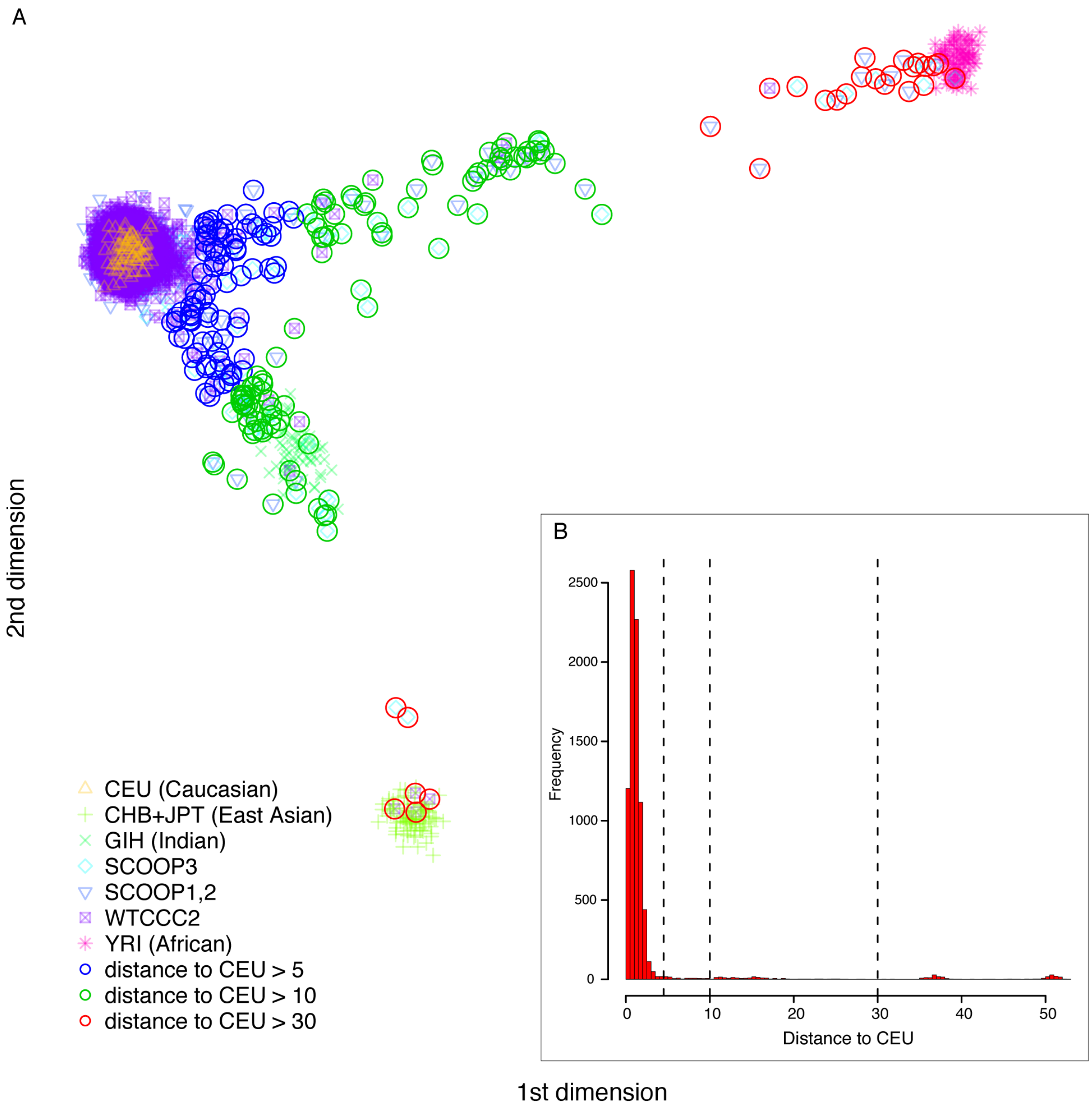

Figure 3.7: Identifying ethnic outliers based on SNP genotypes and MDS projection. (A) Each small symbol represents a sample. The European, East Asian and African populations are well separated and serve as reference points for samples of unknown ethnicity. As a positive control, the Indian population 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).

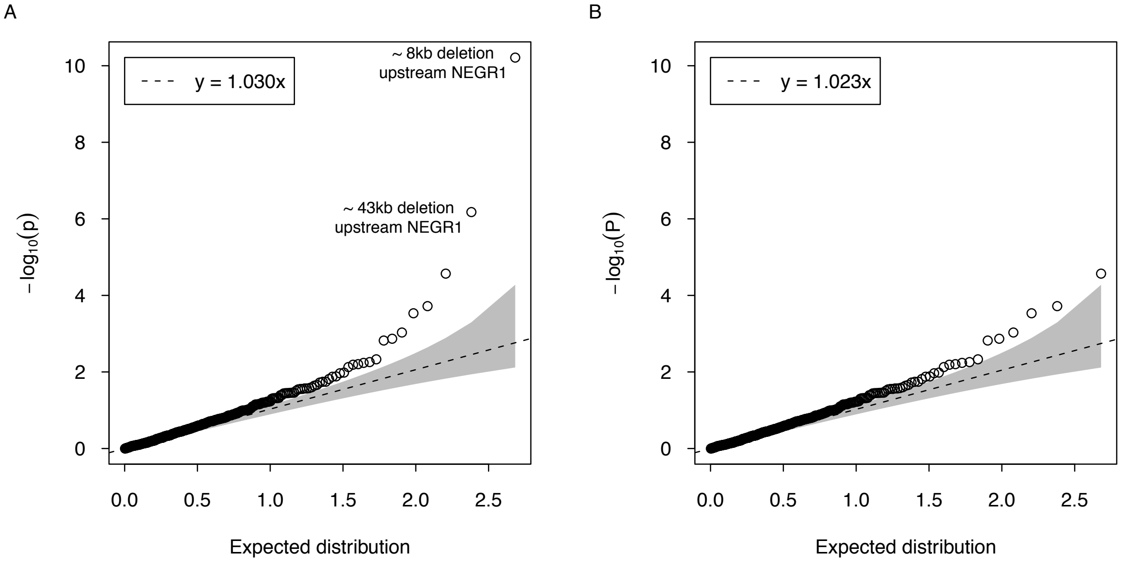

Figure 3.8: Genome-wide association results for common CNVs. (A) Quantile-quantile plot of -log$_{10}$(p) of all 481 common CNVs. Concentration band represents 95% confidence interval. Inflation factor is represented as the slope of the fitted line. (B) Quantile-quantile plot after removal of the two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

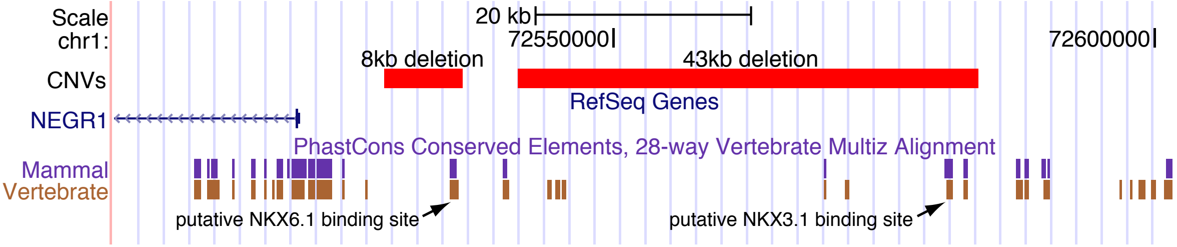
Scale
chr1:
20 kb
72550000
72600000
CNVs
8kb deletion
43kb deletion
RefSeq Genes
NEGR1
PhastCons Conserved Elements, 28-way Vertebrate Multiz Alignment
Mammal
Vertebrate
putative NKX6.1 binding site
putative NKX3.1 binding site

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 ~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 ~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 ~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 ~430kb common deletion and an adjacent ~8kb common deletion, both upstream of the gene NEGR1, were found associated with the phenotype. Conditional analysis revealed the ~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 (λ = 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.