Exploring the Role of Low-Frequency and Structural Genetic Variation in Human Complex Traits

Submitted by

Marcus Aelred Tuke to the University of Exeter Medical School

as a thesis for the degree of

Doctor of Philosophy in Medical Studies

in July 2016

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ……………………………………………………………..

Abstract

Quantitative traits and disease risk in humans are affected by both genetic and environmental factors. Using genome-wide association studies (GWAS) over the past decade, researchers have been successful in finding common genetic polymorphisms that explain a proportion of the variation in many common phenotypes. Despite these significant leaps forward in our understanding, the heritable components of many traits remain largely unaccounted for. A number of explanations as to the “missing heritability” of complex traits and disease risk have been postulated. This thesis addresses some of the unexplained potential sources of heritable trait variation and explores two of its potential causes: low frequency and structural genetic variation.

Chapter 1 provides a background to GWAS, what we have learned from them, discusses the different mechanisms of heritability and reviews the potential explanations for “missing heritability” in complex traits. The chapter then describes low frequency and structural genetic variation and how they fit into the spectrum of genetic variation.

Chapter 2 describes a study that tests the extent to which low frequency association signals can be discovered through low pass whole genome sequencing when using well-powered gene expression and biomarker phenotypes as model traits. The study then compares these association signals to 1000 Genomes based imputation in the same individuals.

Chapter 3 uses methods to detect the structural forms of the human amylase locus with whole-genome sequencing data. The study detects and validates multi-allelic copy number within this region and finds a lack of evidence of a previous association between structural variation of the amylase locus and obesity and body mass index.

Chapter 4 scans for rare copy-number variation (CNV) using SNP microarray data from over 120 thousand individuals at 69 sites that were previously identified as being associated with developmental delay. The chapter aims to refine their prevalence in the general population and attempts to understand their relationship with developmental delay and complex traits.

Chapter 5 aims to detect large deletions and duplications genome-wide using SNP microarray data in a sample of over 120 thousand individuals where we have power to detect rare copy number events. I used novel approaches to test their association with 204 clinically relevant complex traits to determine their role in the heritability of complex traits.

Chapter 6 discusses the findings from the previous chapters within this thesis. I then continue by describing some limitations of this work and explore the potential further directions for future work in this area of study.

Table of Contents

Acknowledgements

I would first like to thank my supervisors Mike Weedon and Tim Frayling - their expert guidance, advice and encouragement has permitted me to continually exceed my own expectations as a researcher and has transformed me into a considerably more adept critical thinker. Although they would be the first to admit that they are rarely in agreement with one another - especially during our group meetings - they both epitomise a high standard of scientific rigour that I've yet to see elsewhere in the field.

I would also like to thank my colleagues in the complex trait genetics group in Exeter. Thank you to Andy Wood for his mentorship in the early days of my PhD and whose opinion and advice I value and respect greatly. Thank you also to Anna Murray, Kate Ruth, Hanieh Yaghootkar, Sam Jones, Rob Beaumont, Rachel Freathy and Jess Tyrrell of the Complex Trait Genetics group in Exeter for their numerous discussions and invaluable insights throughout my PhD. I would also like to thank Konrad Paszkiewicz who has been instrumental in my professional development and has given me numerous opportunities to refine my teaching skills during my studies.

I would like to thank my family for their unwavering love and support. I thank my wife Kelly, the love of my life whom I married during my PhD - without her seemingly superhuman patience and affection I would not have come this far and I therefore dedicate this thesis to her. I also thank my mother Janice and my father Mark who always taught me to work hard, strive to always improve myself and stand up for what I believe in – I could not have asked for better parents. I thank many of my

friends, especially Ben for being the best man at my wedding and for being a consistently great friend throughout the years.

Specific acknowledgments for chapters in this thesis for work completed by other individuals are as follows.

Chapters 2 and 3

Samples were collected as part of the InCHIANTI longitudinal population study. The InCHIANTI study is supported by contract funding from the U.S National Institute on Aging (NIA), and the research was supported in part by the Intramural Research Programme, NIA, and National Institute of Health (NIH).

We thank Michelle Mao and colleagues at the Beijing Genomics Institute (BGI), Beijing, China who were involved in the initial Illumina HiSeq 2000 sequencing of the samples used for analysis. We thank Andrew Wood for carrying out quality control and association analyses and co-writing of the manuscript in Chapter 2. Andrew Wood is supported by the European Research Council grant: SZ-50371-GLUCOSEGENES-FP7-IDEASERC. We thank Dena Hernandez, Michael Nalls, Sean Chong, Matthew Moore, Raphael Gibbs and Andrew Singleton at the National Institute on Aging (NIA), National Institutes of Health for the processing of the Illumina Human-HT12 gene-expression microarrays and Illumina HiSeq targeted sequencing data. We thank Jack Guralnik at the Department of Epidemiology and Public Health, University of Maryland School of Medicine, USA and Stefania Bandinelli, Sanitaria di Firenze, Florence, Italy, and all others involved in sample collection and processing. We thank Hanieh Yaghootkar in the Genetics of Complex

Traits group in Exeter for gene-expression probe remapping and quality control. We also than Anna Murray and David Melzer of the University of Exeter Medical School, Exeter and Luigi Ferruci at the National Institute on Aging, Harbor Hospital, USA for allowing access to the InCHIANTI study for analyses.

We thank the T2D-GENES consortium for whole-genome sequencing in the San Antonio Family Heart Study. Sequence data were generated by the T2D-GENES consortium with support from NIH/NIDDK U01's DK085501, DK085524, DK085526, DK085545 and DK085584. Funding to pay the Open Access publication charges for Chapter 2 was provided by the Wellcome Trust.

We thank Christina Usher of the Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA for carrying out droplet digital PCR validation of the *AMY1* and *AMY2* genes in a subset of InCHIANTI individuals. We also thank Robert Handsaker, Steven McCarroll and Christina Usher of the Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA for their expert feedback and for providing replication whole-genome sequence association data for Chapter 3.

Chapters 4 and 5

Research was conducted using the UK Biobank resource (www.ukbiobank.ac.uk). We thank Jessica Tyrrell of the Complex Trait Genetics group, Exeter for residualizing and inverse normalising the phenotypes analysed in these chapters. We thank Samuel Jones and Robin Beaumont of the Complex Trait Genetics group in Exeter for UK Biobank data transfer and quality control. We also thank Andrew Wood, Katherine Ruth, Hanieh Yaghootkar, Michael Weedon and Timothy Frayling of the Complex Trait Genetics group, Exeter for their contributions to study design, data processing, statistical analyses and interpretation.

We thank Aurélien Macé and Zoltán Kutalik of the Statistical Genetics group at the University of Lausanne for allowing us to use their software used to derive the CNV quality scores in chapter 5.

Author's Declaration

I was involved in the study design, analysis and writing for all chapters and publications that form the basis of my thesis. Each chapter includes work carried out by other analysts, but for each chapter I took a lead role. My specific contributions are given below.

In **chapter 2** I carried out processing of the whole-genome sequence data, and performed subsequent variant calling, refinement, imputation and quality control. I assisted with downstream analysis of the data and co-wrote the manuscript.

In **chapter 3** I processed raw data and called multi-allelic copy number variation in the InCHIANTI cohort. I carried out association analyses, concordance checks with droplet digital PCR data, power calculations and genetic risk score association comparisons. I wrote the chapter and co-wrote the publication that the chapter was adapted from **-** specifically the sections referring to the InCHIANTI cohort.

In **chapters 4 and 5** I wrote the method used to normalise bespoke SNP microarray probeset intensity data in the 120,286 UK Biobank individuals. I called copy number variation both for candidate CNV regions and genome-wide deletions and duplications. I carried out all downstream processing and association analyses with the exception of the CNV quality scores utilised in chapter 5. I wrote both chapters.

CHAPTER 1

Introduction

1.1 Genome-wide association studies have identified thousands of common variants associated with polygenic traits

Genome-wide association studies (GWAS) have identified thousands of genetic factors contributing to polygenic diseases and traits over the past decade (**Fig. 1**). These studies have traditionally focussed on the principle that the heritable component of common disease is largely explained by common genetic variation with a frequency of greater than 5% in the population. There are now 65 loci where a common genetic variant is associated with Type 2 Diabetes [1] for example, and 50 variants associated with measures of quantitative glycaemic traits [2]. These associations are providing new insights to the biology of polygenic disease. The association of a common variant in the *FTO* gene that predisposes to childhood and adult obesity is an example of a successful genome-wide association based study [3]. Here, a genome-wide association study compared 500,000 common single-nucleotide polymorphisms (SNPs) between 1,924 U.K. Type 2 Diabetes patients and 2,938 U.K. population controls and identified a strong association at the *FTO* locus. It was shown that the *FTO* association was specifically driven by an effect on fat mass – with the Type 2 Diabetes increasing allele adding an average of ~1.5kg fat weight. Before GWAS, *FTO* was a poorly characterised gene with no indication that it was involved in adiposity regulation. Subsequently, a vast amount of effort has gone into determining how the *FTO* gene affects adiposity, with some work suggesting a primary effect on basal metabolic rate [4], adipocyte browning [5] and other work suggesting an effect on appetite [6]. There is as yet no definitive answer, but it is a good example of how GWAS has opened up a whole new area of study into the biology of complex disease.

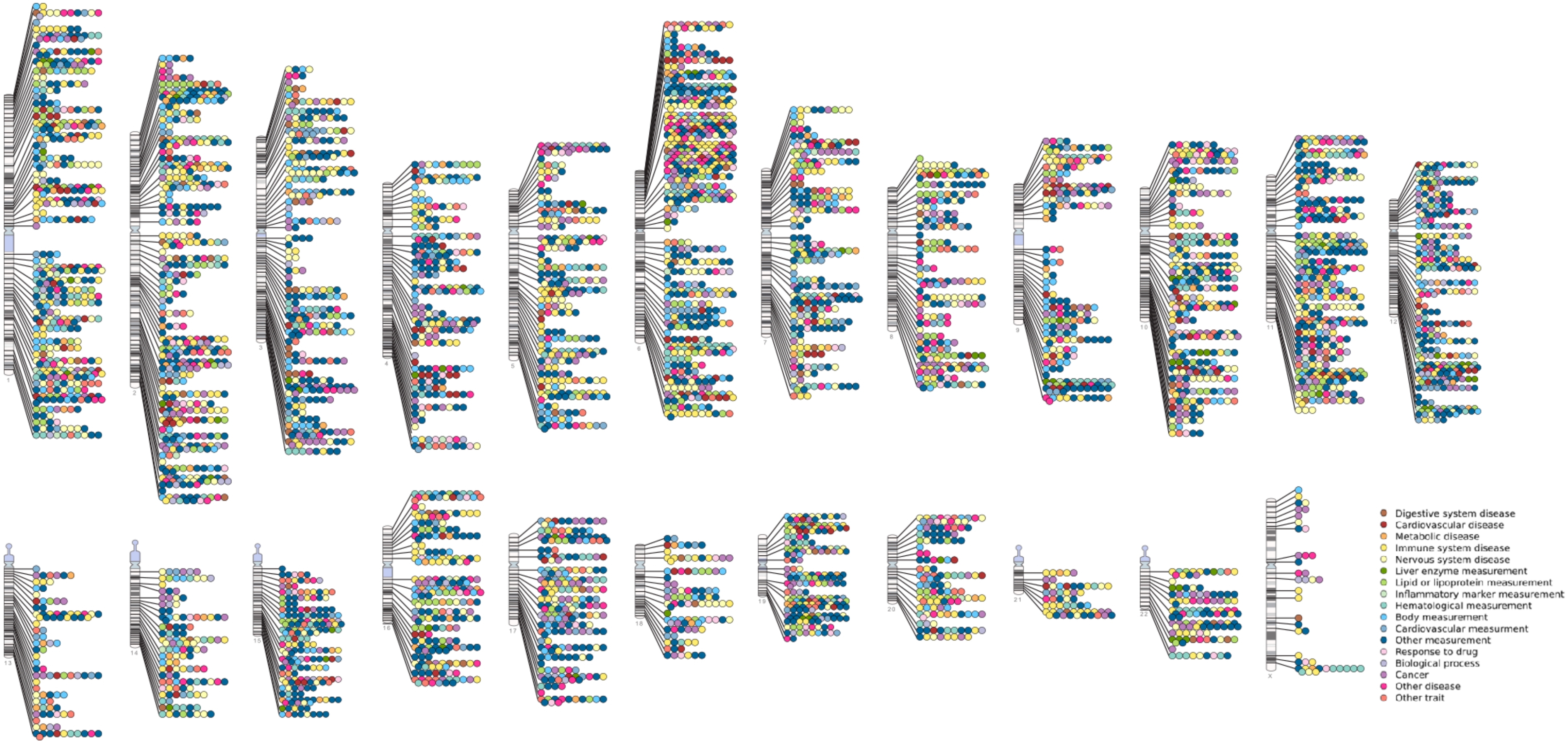


Figure 1. Genome-wide ideogram showing all GWAS results to date. Taken from the GWAS catalog at http://www.genome.gov in May 2016.

1.2 Common variants identified from GWAS account for a fraction of the heritability of most complex traits

It has been possible to estimate ‘narrow-sense’ heritability (h^2) – the heritability of polygenic traits under an additive model using genetic information [7]. Despite the major leaps in progress, the associated loci identified from GWAS still account for a relatively small fraction of h^2 in most polygenic traits studied. The 65 Type 2 Diabetes loci, for example, account only for ~10% of the genetic component of Type 2 Diabetes risk [8]. Human height, a model polygenic trait is a classic example of how known genetic variation explains heritability. A GWAS meta-analysis of 253,288 individuals has revealed that 697 genetic variants in 423 loci influence final adult height [9]. However, these variants only explain ~25% of the total narrow-sense heritability of height. There is currently much interest in explaining the remaining, or “missing” heritability of complex diseases and traits.

1.3 Explaining “missing heritability”

There are several hypotheses as to how so-called missing heritability could be accounted for [10-12]. Some initial theories have hypothesised the original heritability estimates using twin/family studies are inflated, possibly by factors contributing to broad sense heritability (H^2), that is, non-additive genetic components [11]. Another reason for inflation of h^2 estimates is that phenotypes may have varying diagnoses or definitions. A group of individuals classified with a given disease for example may have different biological mechanisms leading to the same disease, rendering cases to have a possible misclassification, although traits that are empirically measured such as adult height, BMI or circulating biomarkers are less likely to have this problem. Assuming a given trait is well characterised, the remaining causes of missing heritability from a discovery standpoint can be classified into two sets: (a) hiding heritability which can be described as non-variant-type causes of missing heritability such as, high polygenicity, allelic heterogeneity or departure from the additive model, or (b) still-missing heritability, which represents the types of genetic variant that have not been captured using conventional GWAS methods [12] (**Fig. 2**).

1.3.1 'Hiding' heritability

Hiding heritability is the term given to the common genetic variation used in GWAS that is still not accounted for as a heritable component because of the way these variants are analysed and interpreted. When discussing h^2 only, a further exploration of heritability can be in the form of allelic heterogeneity or higher levels of polygenicity whereby smaller effect genetic variants that we previously did not have power to detect are contributors. Hiding heritability may also be confounded by H^2, heritability explained by a departure from the additive model which includes non-additive effects, interactions between genes and other genes (epistasis), and interactions between genes and their environment. These concepts are discussed in further detail below.

1.3.1.1 Highly polygenic effects

A highly polygenic hypothesis stipulates that thousands of genetic variants contribute to quantitative traits and common disease. Each of the thousands of genetic variants may exhibit a very small effect on the trait or disease risk (**Fig. 3**). It is however, very difficult to detect these variants without increasing sample size dramatically into hundreds of thousands of samples. The aforementioned height meta-analysis of 253,288 individuals gives a glimpse into a more comprehensive genetic architecture of complex traits whereby several hundred loci have been implicated [9]. Future studies in multiple traits with very large sample sizes will help uncover the levels of polygenicity of complex traits and common disease risk.

1.3.1.2 Allelic Heterogeneity

The presence of multiple disease variants at a single locus is known as allelic heterogeneity. Few studies have accounted for allelic heterogeneity at known loci and the strongest signal is generally labelled as the index SNP for that locus with all other variants assumed to be in linkage disequilibrium (LD) or not associated with the trait. Conditional analyses have shown that when repeating the analysis whilst using the given index SNP as a covariate, other separate signals can sometimes arise [13]. Allelic heterogeneity has shown to reveal more complex patterns of association and therefore explain additional trait heritability.

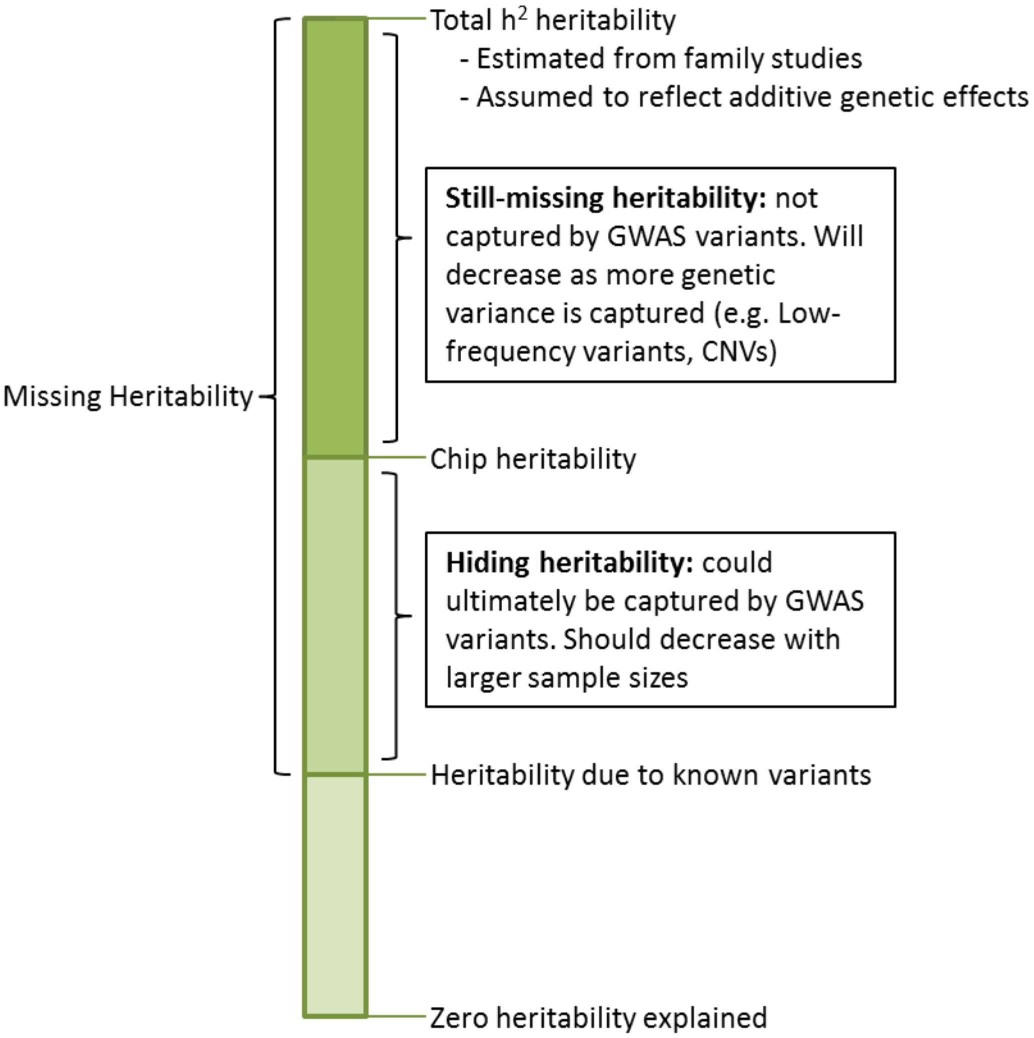

Total h2 heritability
- Estimated from family studies
- Assumed to reflect additive genetic effects
Still-missing heritability: not captured by GWAS variants. Will decrease as more genetic variance is captured (e.g. Low-frequency variants, CNVs)
Missing Heritability
Chip heritability
Hiding heritability: could ultimately be captured by GWAS variants. Should decrease with larger sample sizes
Heritability due to known variants
Zero heritability explained

Figure 2. An explanation of 'Hiding' heritability and 'Still missing' heritability and their contribution to overall narrow-sense heritability (h^2). Adapted from Witte et al. 2014

1.3.1.3 Deviation from the additive model and genetic interaction

Genome wide association studies are based on an additive effects model such that each trait increasing allele exhibits a linear increase or decrease in levels of a quantitative trait or risk of disease. Many GWAS to date have not investigated non-additive effects playing a role in complex traits and common diseases. Dominant and recessive models test for interactions between different alleles at the same locus whereas gene-gene interaction analyses test the effect of one variant being dependent on the genotype of another at a separate locus [14]. Additionally, gene-environment interaction is where the different alleles of a genotype respond to an environment in different ways [15]. These analyses are generally underpowered and are often confounded by additional genetic variants that were not accounted for [16].

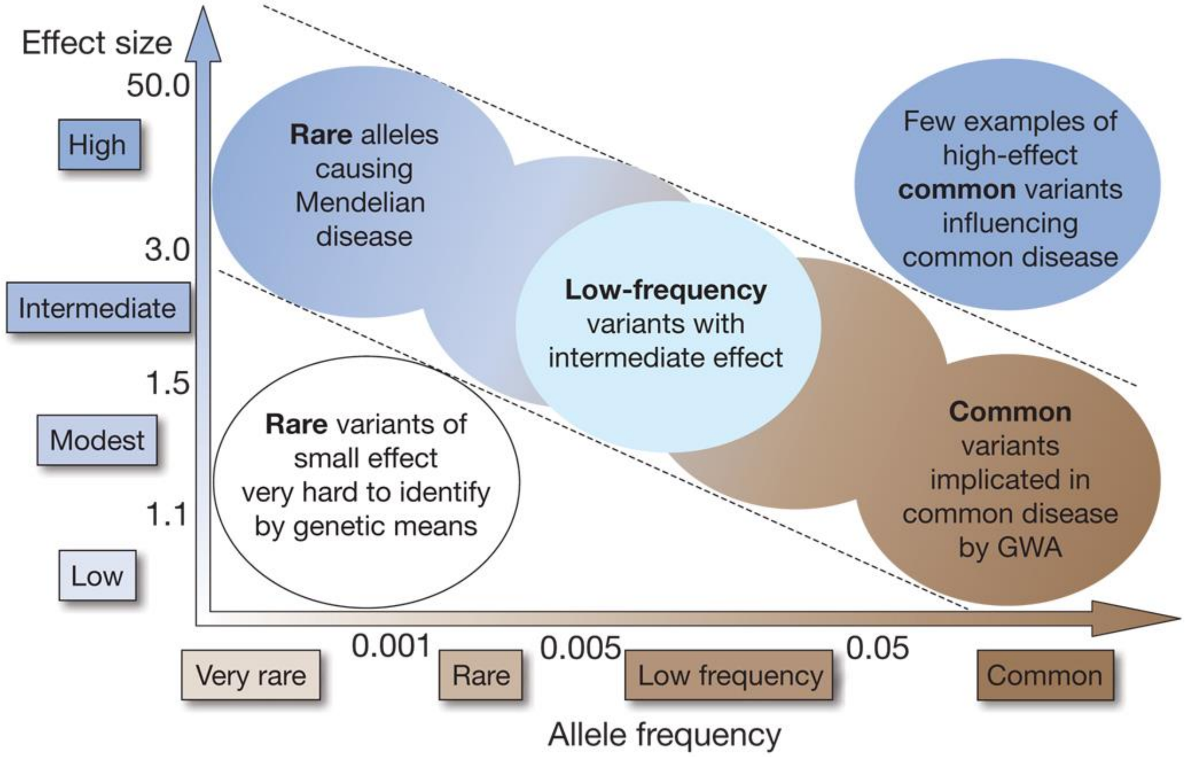

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Figure 3. A visualisation of the allele frequency vs effect size variant definition spectrum. Taken from Manolio et al. 2009.

1.3.2 ‘Still-missing’ heritability

The current GWAS approach has focussed on common SNPs (minor allele frequency > 5%). This has been because of technological limitations. The most feasible way to genotype a large number of individuals in a study has been to use a genotyping microarray, but these ‘SNP chips’ can only reliably genotype common SNPs. An unknown proportion of the missing heritability of complex traits is therefore thought be derived from low-frequency and rare variants (minor allele frequency < 5%) and structural variation.

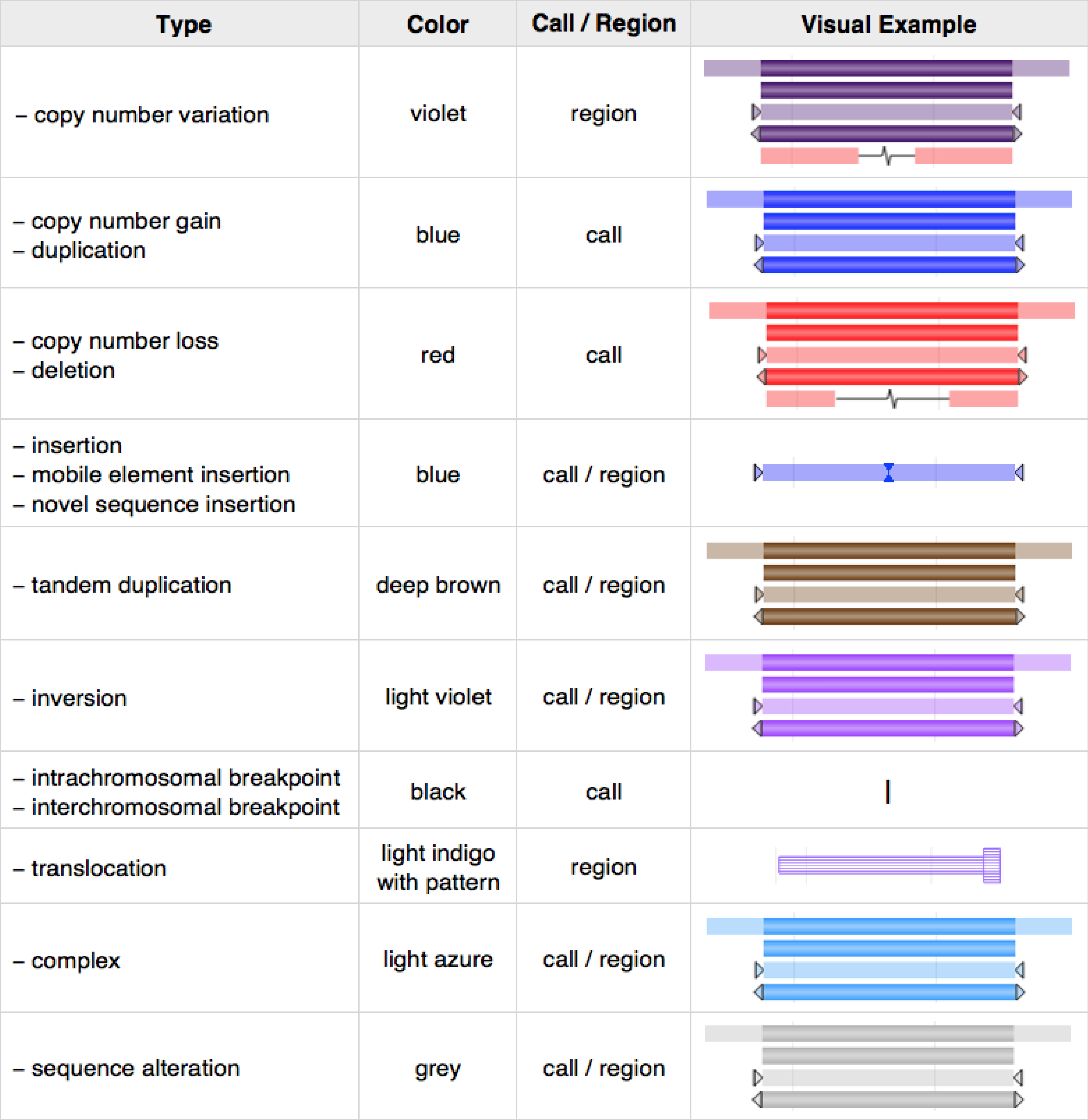
1.3.2.1 Low frequency and rare variation

Detecting low frequency (allele frequency between 1% and 5%) and rare (<1%) variation has been very challenging until recently. Technological advances have led to the development of whole genome sequencing (WGS), and it is now possible to sequence significant numbers of human genomes with an acceptable cost, accuracy and speed [17]. Whole genome sequencing allows many types of variants to be comprehensively detected and there is initial evidence to show that all of these types of variation are likely to be important in explaining missing heritability. A type 1 diabetes re-sequencing study of 480 cases and 480 controls for example, discovered four rare variants of *IFIH1* (interferon induced with helicase C domain 1) that substantially lowered risk of type 1 diabetes [18]. This was an important finding because it highlighted a potential viral origin of type 1 diabetes.

An alternative and more cost and time effective approach is to utilise the 1000 Genomes reference panel to statistically impute the majority of low-frequency genetic variants into the population being analysed [19]. The 1000 genomes project aimed to capture greater than 99% of genetic variants present in all human populations at a frequency of >1%. This approach relies on the rare and low-frequency variants tagging neighbouring SNPs either on the SNP chip itself or in the reference panel. Novel rare or low-frequency, large-effect signals may not be in strong linkage disequilibrium with the common variants and would not therefore be well imputed whereas WGS could empirically detect a potential untagged signal. Until now, it has been difficult to determine how much more powerful whole-genome sequencing will be over using imputation reference panels to capture novel association signals.

1.3.2.2 Structural variation

Structural variation (SV) is a term that covers genomic imbalances (DNA deletions, insertions and duplications) also known as copy number variation (CNV), or insertions/deletions (Indels). These terms are interchangeable but it is generally accepted that CNVs affect loci greater than 1Kb in size whereas Indels are smaller than 1Kb. SVs are also used to describe DNA inversions, translocations and other complex types of polymorphism (**Fig. 4**). Regardless of their type, it is widely believed that SVs become rarer as they become larger (**Fig. 5**) as larger SVs have the potential to be more pathogenic and therefore undergo negative selection [20]. Findings so far have indicated that larger SVs exhibit a stronger phenotypic effect with many smaller SVs being thought to have a benign effect [21]. There are several potential mechanisms in which an SV can alter the function of a gene and therefore influence trait effects or disease risk, these include dosage imbalance and compound effects with intersecting variants (**Fig. 6**) [22]. Larger SVs usually intersect with more genes, so it would be rational to hypothesise that they have a larger and more frequent effect on complex traits and disease risk, however more analyses need to be carried out to characterise the full spectrum of structural variation before we can fully understand its effects on complex traits and disease risk.



Type	Color	Call / Region	Visual Example
– copy number variation	violet	region	
– copy number gain – duplication	blue	call	
– copy number loss – deletion	red	call	
– insertion – mobile element insertion – novel sequence insertion	blue	call / region	
– tandem duplication	deep brown	call / region	
– inversion	light violet	call / region	
– intrachromosomal breakpoint – interchromosomal breakpoint	black	call	
– translocation	light indigo with pattern	region	
– complex	light azure	call / region	
– sequence alteration	grey	call / region	

Figure 4. The various types of structural variation (SV) ranging from simple deletions or duplications to more complex inversions and translocations. Some regions of the genome will contain a compound of these variants. Taken from the dbVar overview of SVs at http://www.ncbi.nlm.nih.gov/dbvar

Structural variants can be detected genome-wide in SNP chips in regions that harbour a high density of SNP probes, and more comprehensively using WGS [19]. Significant computational challenges are still yet to be overcome in calling SVs genome-wide, but nevertheless different forms of SV have been associated with complex traits and disease risk. For example, a large, rare deletion of chromosome 16p11.12 has been shown to cause obesity, whereas the duplication of this region has been shown to be associated with lower obesity [23,24], and it is hypothesised that a dosage imbalance of one or more genes in this region give rise to phenotypic variation. A smaller common CNV at the *NEGR1* locus has also been found to be associated with severe early-onset obesity [25]. In addition to complex trait association findings, a number of large (≥50Kb) and rare deletion and duplication events have been found to be associated with developmental delay. These rare copy number variants have been found to be highly pleiotropic, whereby a single genetic signal influences multiple seemingly unrelated traits, and of variable penetrance [20,26], giving insight into the pathogenicity of large, rare and negatively selected copy number variants.

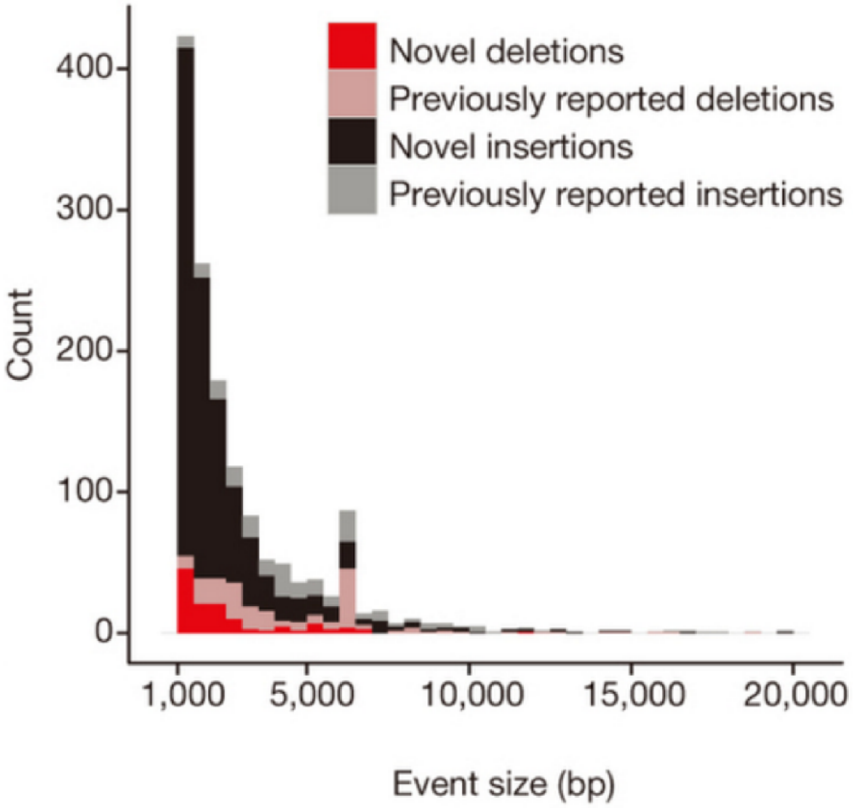
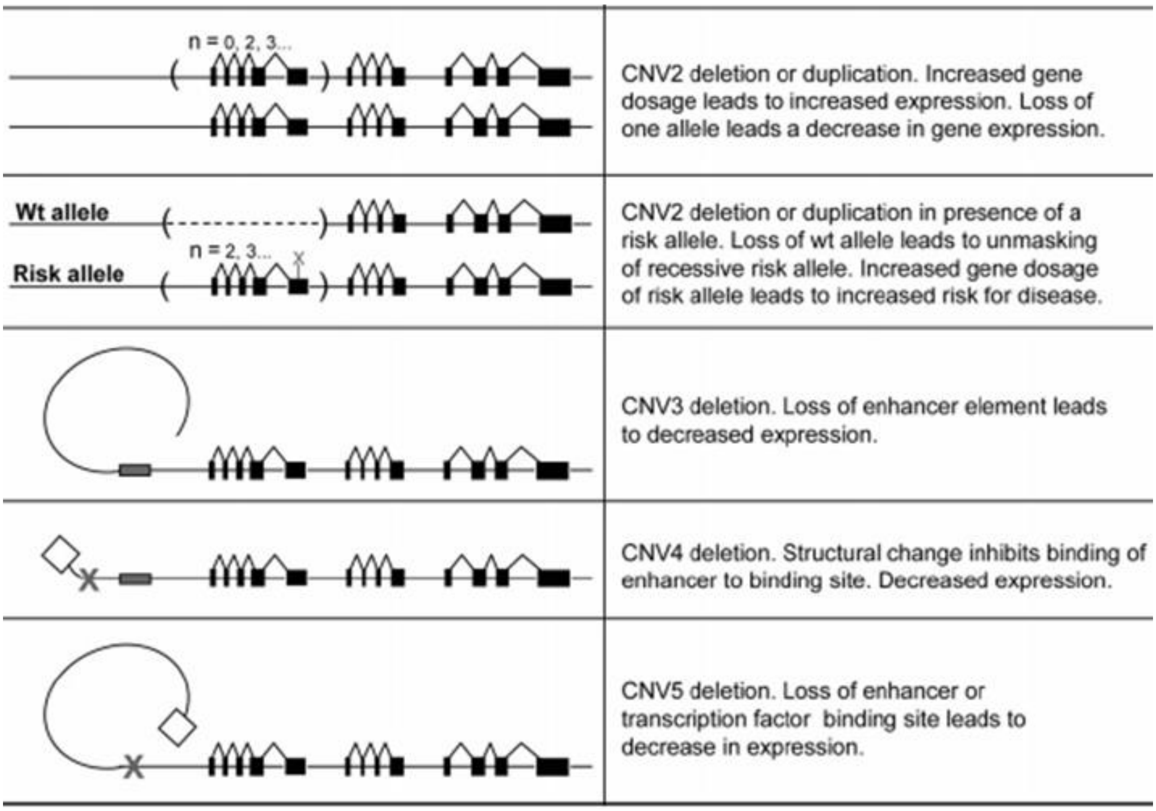


Figure 5. The frequency of structural variation in the human genome when divided into 100bp event size bins. Taken from Chaisson et al 2015.

In addition to single structural events, many regions of the genome are structurally complex and include many types of SV acting together to produce a complex multi-haplotype variant. The 17q12.31 region for example, contains a megabase-long inversion and various combinations of flanking duplications and deletions. The inversion itself was shown to be positively selected in Europeans, specifically; its presence was advantageous and was therefore swept across the population as indicated by an increased number of offspring in inversion carriers [27]. Another example of a complex region is the *AMY1* locus, a gene that transcribes the Amylase enzyme that is secreted in the salivary glands and converts dietary starches into sugar [28]. There are numerous variations of the number of copies of this gene, with duplications positively selected for populations that have evolved with diets rich in starch [29]. Much work is being carried out to further understand and characterise complex regions such as these examples and their impact on complex traits and disease risk. Further understanding the role of all forms of structural variation in human disease and complex traits remains challenging, but methods are rapidly progressing to allow a more complete understanding of these elusive parts of the genome.



CNV2 deletion or duplication. Increased gene dosage leads to increased expression. Loss of one allele leads a decrease in gene expression.

CNV2 deletion or duplication in presence of a risk allele. Loss of wt allele leads to unmasking of recessive risk allele. Increased gene dosage of risk allele leads to increased risk for disease.

CNV3 deletion. Loss of enhancer element leads to decreased expression.

CNV4 deletion. Structural change inhibits binding of enhancer to binding site. Decreased expression.

CNV5 deletion. Loss of enhancer or transcription factor binding site leads to decrease in expression.

Figure 6. Five examples of how a CNV can potentially affect the biological mechanisms of complex traits and/or disease through unmasking recessive point mutations or by effects on gene dosage. Taken from Feuk et al 2006.

1.4 Aim of thesis

The aim of this thesis is to utilise both next generation sequencing, cohorts with larger sample sizes, and post-GWAS-era analytical techniques to test if we can explain more of the genetic component of complex traits and disease risk than from GWAS alone.

In chapter 2 we process and analyse low pass whole genome sequence data in several hundred individuals to see if there are novel statistical associations between low frequency SNPs/Indels and well-powered *cis*-eQTLs and biomarker phenotypes. We then look at how these associations compared when repeating the same analyses in SNP chip data imputed with 1000 genomes reference genotypes.

In chapter 3 we attempt to more clearly define, validate and replicate multi-allelic copy number of the *AMY1* gene using whole-genome sequencing data with two separate computational detection methods. Copy number at the *AMY1* locus has been thought to be associated with BMI and obesity in previous studies using microarray data and quantitative PCR [30,31]. We want to see if this *AMY1* copy number association can be replicated using alternative approaches.

In chapter 4 we use custom built SNP array data in ~120,000 individuals to further characterise large (≥50Kb) pathogenic copy number variants associated with developmental delay [26]. These CNVs have a phenotypic impact when detected as both deletions and duplications, with a higher prevalence in cases compared to controls. We want to see if (a) these CNVs are still pathogenic when multiplying the size of the control population several times and (b) how these pathogenic CNVs are associated with complex surrogate measures of developmental delay and other complex traits in a large control population.

Finally, in chapter 5 we investigate the effect of large deletions and duplications genome-wide in ~120,000 individuals. Large deletions and duplications have previously been too rare to be detected in previous GWAS cohorts, here we treat deletions and duplications separately and attempt to account for breakpoint heterogeneity between individuals. We test for association between deletions and duplications adjusted for a novel quality score metric against ~200 continuous traits to see if further large, rare deletions and duplications play a key role in our understanding the heritable component of complex traits where we have sample sizes large enough to detect them.

1.5 References

1. Morris AP (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet advance online publication.
2. Scott RA, Lagou V, Welch RP, Wheeler E, Montasser ME, et al. (2012) Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet advance online publication.
3. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889-894.
4. Hubacek JA, Pikhart H, Peasey A, Kubinova R, Bobak M (2011) FTO variant, energy intake, physical activity and basal metabolic rate in Caucasians. The HAPIEE study. Physiol Res 60: 175-183.
5. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, et al. (2015) FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med 373: 895-907.
6. Karra E, O'Daly OG, Choudhury AI, Yousseif A, Millership S, et al. (2013) A link between FTO, ghrelin, and impaired brain food-cue responsivity. J Clin Invest 123: 3539-3551.
7. Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 9: 255-266.

8. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44: 981-990.

9. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46: 1173-1186.

10. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356-369.

11. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.

12. Witte JS, Visscher PM, Wray NR (2014) The contribution of genetic variants to disease depends on the ruler. Nat Rev Genet 15: 765-776.

13. Wood AR, Hernandez DG, Nalls MA, Yaghootkar H, Gibbs JR, et al. (2011) Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. Hum Mol Genet 20: 4082-4092.

14. Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, et al. (2014) Detection and replication of epistasis influencing transcription in humans. Nature 508: 249-253.

15. Hunter DJ (2005) Gene-environment interactions in human diseases. Nat Rev Genet 6: 287-298.

16. Wood AR, Tuke MA, Nalls MA, Hernandez DG, Bandinelli S, et al. (2014) Another explanation for apparent epistasis. Nature 514: E3-5.

17. Mardis ER (2011) A decade's perspective on DNA sequencing technology. Nature 470: 198-203.

18. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324: 387-389.

19. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

20. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, et al. (2011) A copy number variation morbidity map of developmental delay. Nat Genet 43: 838-846.

21. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 42: D986-992.

22. Feuk L, Marshall CR, Wintle RF, Scherer SW (2006) Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 15 Spec No 1: R57-66.

23. Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, et al. (2010) A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature 463: 671-675.

24. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, et al. (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron 72: 245-256.

25. Wheeler E, Huang N, Bochukova EG, Keogh JM, Lindsay S, et al. (2013) Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nat Genet 45: 513-517.

26. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, et al. (2014) Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet 46: 1063-1071.

27. Boettger LM, Handsaker RE, Zody MC, McCarroll SA (2012) Structural haplotypes and recent evolution of the human 17q21.31 region. Nat Genet 44: 881-885.

28. Groot PC (1989) The human [alpha]-amylase multigene family consists of haplotypes with variable numbers of genes. Genomics 5: 29-42.

29. Perry GH (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39: 1256-1260.

30. Falchi M (2014) Low copy number of the salivary amylase gene predisposes to obesity. Nat Genet 46: 492-497.

31. Mejia-Benitez MA, Bonnefond A, Yengo L, Huyvaert M, Dechaume A, et al. (2015) Beneficial effect of a high number of copies of salivary amylase AMY1 gene on obesity risk in Mexican children. Diabetologia 58: 290-294.

CHAPTER 2

Whole genome sequencing to understand the genetic architecture of common gene expression and biomarker phenotypes

Wood, A.R.*, Tuke, M.A.*, Nalls, M., Hernandez, D., Gibbs, J.R., Lin, H., Xu, C.S., Li, Q., Shen, J., Jun, G. et al. (2015) Whole-genome sequencing to understand the genetic architecture of common gene expression and biomarker phenotypes. Human molecular genetics, 24, 1504-1512.

2.1 Abstract

Initial results from sequencing studies suggest that there are relatively few low frequency (<5%) variants associated with large effects on common phenotypes. We performed low pass whole genome sequencing in 680 individuals from the InCHIANTI study to test two primary hypotheses: i) that sequencing would detect single low frequency - large effect variants that explained similar amounts of phenotypic variance as single common variants, and ii) that some common variant associations could be explained by low frequency variants. We tested two sets of disease-related common phenotypes for which we had statistical power to detect large numbers of common variant - common phenotype associations – 11,132 *cis*-gene expression traits in 450 individuals and 93 circulating biomarkers in all 680 individuals. From a total of 11,657,229 high quality variants of which 6,129,221 and 5,528,008 were common and low frequency (<5%) respectively, low frequency - large effect associations comprised 7% of detectable *cis*-gene expression traits (89 of 1,314 *cis*-eQTLs at $P<1x10^{-06}$ (FDR ~5%)) and 1 of 8 biomarker associations at $P<8x10^{-10}$. Very few (30 of 1,232; 2%) common variant associations were fully explained by low frequency variants. Our data show that whole genome sequencing can identify low frequency variants undetected by genotyping based approaches when sample sizes are sufficiently large to detect substantial numbers of common variant associations, and that common variant associations are rarely explained by single low frequency variants of large effect.

2.2 Introduction

Initial results from sequencing studies suggest that there are relatively few low frequency (minor allele frequency <5%) variants associated with large effects on common human phenotypes [1-5]. However, few of these sequencing experiments have used sample sizes similar to those required to identify most common variant-phenotype associations [1,4]. Still fewer sequencing studies have examined the whole genome, instead most have focused on exomes [5] or targeted sets of genes [1,6] or have focused on population genetics rather than phenotype associations [7]. The frequency of heterozygote alleles in a population is 2pq, where ‘p’ is equal to dominant allele frequency and ‘q’ recessive allele frequency, the proportion of phenotype variance explained is a function of allele frequency and effect size (approximated as beta^2 x (2pq)). Limitations in sample size means that many current sequencing studies are powered to detect only those single low frequency variants that explain substantially more phenotypic variance than single common SNPs. In this study we define “low frequency - large effect” as a variant that has a minor allele frequency less than 5% but that has a sufficiently large per-allele effect on a phenotype that it explains a similar proportion of phenotypic variance as common variants detectable in the same sample size.

Current whole genome and exome sequencing based studies are aiming to answer several questions of relevance to common disease and quantitative phenotypes. First, how many low frequency and rare variant associations can we reasonably expect to identify and in what sample sizes? Second, are common variant – common phenotype associations driven by low frequency associations? Third, could we do just as well by imputing genotypes from the 1000 Genomes Project and other reference panels? These questions are important for a number of reasons. First, it is known that most human genetic variation is low frequency and rare but there is considerable debate as to how best to identify which of these variants are associated with common phenotypes. Studies include whole exome and gene targeted approaches [5,6] low pass sequencing in unrelated individuals [4] and high pass sequencing in families. Second, if low frequency and rare variants are responsible for many common variant - common phenotype associations it will likely implicate a different set of causal genes and regulatory elements for follow up. Finally, few studies have tested the power of imputation from reference panels to identify low frequency association signals and this approach could be the most efficient way of studying lower frequency effects in large sample sizes, as recently shown by a deCODE study [8].

To help answer these questions we performed low pass (median 7X) whole genome sequencing in 680 individuals from the population based InCHIANTI study (**Table 1, Fig. S1 and Table S1, Materials and Methods**). We selected two sets of common phenotypes for which the InCHIANTI study provided sufficient statistical power to detect large numbers of common variant - common phenotype associations – 11,132 whole blood based *cis*-gene expression traits in 450 individuals, and 93 circulating biomarkers in 673 individuals. Previous micro-array based GWAS of 1,200 InCHIANTI individuals detected 1,298 *cis*-eQTLs and 30 circulating biomarker associations [9,10]. In addition to providing good power to detect common variant associations, these phenotypes are highly relevant to human disease. Multiple studies have shown that common variant disease associations are enriched for variants affecting gene expression in *cis* and *trans* [11-16] and our biomarkers included many of public health importance: vitamins A and D; cholesterol; magnesium, calcium and potassium ions; inflammatory markers and circulating proteins associated with metabolic disease (including leptin and adiponectin).

Table 1. Basic characteristics of the 680 InCHIANTI individuals selected for sequencing at baseline.

Characteristic	**Mean (range) or %**
Age (years)	64.2 (23-90)
Sex (% male)	44.9%
BMI	27.2 (18.1-46.6)
Current Smokers (% case)	21.3%
History of hypertension (% case)	33.8%
History of diabetes (% case)	8.7%
History of myocardial infarction (% case)	2.6%

We tested two main hypotheses. First, that whole genome sequencing would detect single low frequency genetic variants that individually explain a similar proportion of phenotypic variance as single common genetic variants. Our second main hypothesis was that some individual common variant – common phenotype associations would be explained by low frequency variants. As a secondary hypothesis that has recently been tested in other studies [7], we also tested whether or not low frequency variant - common phenotype associations would be better captured by low pass sequencing than imputation from the 1000 Genomes Project.

2.3 Materials and Methods

2.3.1 Samples

We selected 680 individuals from the InCHIANTI study [9,17]; a study of aging from the Chianti region in Tuscany, Italy, for low pass whole genome sequencing (**Table 1**). Selection criteria included the availability of microarray genotype data and non-missingness of phenotypic data that included gene expression data and circulating biomarker.

2.3.2 Whole-genome sequencing

Whole-genome sequencing was performed at the Beijing Genomics Institute (BGI), Shenzhen, China using Illumina HiSeq 2000 to obtain a minimum read depth of 6X and median of 7X. An average of 240 million paired-end 90bp reads per sample were aligned to the 1000 Genomes implementation of the build 37 genome reference consortium (GRC) human reference genome [18], using the burrows-wheeler aligner (BWA) version 1.5.9 [19] (**Fig. S1 and Table S1**).

2.3.3 Sequence read processing

Using the sequence reads aligned at BGI through BWA, each genome was scanned for small insertions and deletions (indels) using the Genome Analysis Toolkit (GATK) version 1.6 indel re-aligner [20]. This process detected both de-novo and known indels from dbSNP version 135 [21]. Regions containing indels were then realigned to the reference genome. Duplicated reads across the genome were detected using Picard version 1.59 (available from http://picard.sourceforge.net) and subsequently removed to avoid potential bias when genotyping. In addition, base quality scores in each aligned read were recalibrated using the GATK version 1.6 table recalibrator. Recalibration used read group, reported quality score, sequencing machine cycle and sequence context as covariates.

2.3.4 Sequence variant identification

SNP and indel calling was performed across all 680 genomes using the GATK version 2.2 unified genotyper. False-positive variant calls were filtered using variant quality score recalibration (VQSR). VQSR developed a covarying estimate of the relationship between eight variant call annotations (read depth, mapping quality, quality of read depth, haplotype score, inbreeding coefficient, mapping quality bias, strand bias, and read position bias) and the probability that the call is a true genetic variant. The truth model was determined adaptively based on HapMap 3.3 sites and polymorphic sites from the 1000 genomes Omni 2.5M SNP chip array [22,23].

2.3.5 Quality control of variant capture and sequence-based genotype calls

As a quality control check we first used GATK's variant annotator (version 2.2) to determine the overlap of discovered variants catalogued in HapMap 3.3 [23]**,** 1000 Genomes Omni 2.5, and the 1000 Genomes phase 1 indel dataset [22] (**Fig. S2**).

2.3.6 Imputation of sequence data to recover and refine genotype sites

Haplotype phasing was performed using Beagle version 3.3 [24], and missing data was imputed internally using the filtered and present genotypes only. For SNPs we observed an overall Ti/Tv of 2.19. A summary of the variants captured can be found in **Tables S2 and S3**.

2.3.7 Variant and genotype comparison with 2Mb of high depth sequence (median >30X)

We compared variants captured though our low-pass sequencing experiment with regions known to associated with Parkinson's disease sequenced at high depth (median >30X) in 96 InCHIANTI subjects (total of ~2Mb). SNP and indel calling was performed across the 96 samples using the GATK's unified genotyper (version 2.2). SNPs were filtered using VQSR and indels were hard filtered using GATK version 2.2 variant filtration. Of the 96 subjects, 83 subjects overlapped formed a subset of the 680 whole-genome sequenced subjects.

To assess the quality of the variants captured and the genotypes called in low pass sequencing we created a high quality set of variants and genotypes called in the high depth 2Mb of sequence data. We filtered by 1) masking out polymorphic regions in chromosome 6 and 17 in the 2Mb regions in both datasets; 2) removing sites containing a genotype called at less than 20X coverage in the high-depth sequence dataset from both datasets; 3) removed all non-biallelic sites from the respective dataset.

The degree of overlap was then calculated each way for both of the filtered 2Mb datasets using the GATK version 2.2 variant annotator and the genotype concordance matrices were calculated in overlapping sites using VCFtools version 0.1.9 [25] (**Fig. S3 and Tables S4 – S7**).

2.3.8 Quality control of genotypes derived from sequence-based imputation

As an additional quality control check of internally imputed genotypes we performed genotype concordance checks with the Illumina HumanHap550 GWAS chip. Of the 680 subjects, 7 were selected for exclusion as the fraction of concordant genotypes for each subject was consistent with a sample swap (52% concordance in each instance). For the remaining 673 subjects we observed good concordance with the genotyping array (>98% concordance). For all 673 samples genotyping calls increased after internal imputation performed by Beagle (**Figs. S4 and S5**).

2.3.9 1000 Genomes imputation

To compare whole genome sequencing to imputation from the 1000 genomes reference panel we used haplotypes from the 1000 Genomes Phase I integrated (version 3) release with singletons removed (30,061,896 variants; 28,681,763 SNPs and 1,380,133 indels). Genotype data captured on the Illumina HumanHap550 chip were phased using MACH 1.0.16 [26,27]. Subsequent imputation was performed using Minimac (version 2012.10.9) [28]. We used a multi-ethnic haplotype reference panel that included 1,092 individuals including 379 Europeans (including 98 Tuscans), 181 Americans, 246 Africans, and 289 Asians, in an attempt to capture variants that may be rare in Europeans but more common on haplotypes from different ethnic backgrounds.

2.3.10 Variants included in association analyses

For all association analyses we filtered on biallelic variants with a minor allele count ≥4 and an r^2 imputation quality >0.7. To ensure comparable imputation metrics between the Beagle- and MaCH/Minimac derived dosages we recalibrated the Beagle imputation metric to MaCH's r^2 [27]. As described above, Beagle was used to refine and recover genotypes for the variant sites identified by the low-pass sequencing data.

2.3.11 *cis*-eQTL association analysis

Whole-genome expression profiles of the InCHIANTI subjects were derived from whole blood and captured using the Illumina HT12-v3 BeadChip as previously described [29]. We excluded probes that harboured non-singleton variants within the 50bp probe region captured by our sequencing efforts or the Exome Sequencing Project (ESP) [30]. This resulted in 11,132 probes for association testing. We performed kinship analysis using KING [31] and removed first-degree relatives from the analysis that resulted in 450 remaining individuals. We inverse-normalised the intensity values for the filtered probes and individuals prior to generating residuals that adjusted for age, sex, amplification batch, and hybridization batch to increase the likelihood of the error around the model being normally distributed. Finally, we inverse-normalised the residual values prior to performing the association analyses. We performed association testing in *cis* having defined a *cis* region as ±1Mb the probe transcription start site. Dosages output by BEAGLE were formatted for MACH2QTL [26,27] and variants in *cis* tested against the normalised intensity values of the respective probe.

2.3.12 Circulating biomarker association analyses

A full list of the 93 circulating biomarkers is provided in the **Table S8**. For the 93 circulating biomarkers we similarly performed a double inverse-normalisation for each trait but inversed normalised the raw data values, and adjusted for age and sex only when generating the residuals. We tested the entire genome for associations against each of the circulating biomarkers using all 673 chip-concordant subjects. We used a mixed linear model as implemented in EMMAX [32] to account for relatedness instead of removing subjects from the analysis.

2.3.13 Estimating numbers of independent variants using 2Mb windows

We used LDSelect version 1.0 [33] across 22 2Mb windows (1 per autosomal chromosome) to estimate the average number of independent variants (MAC ≥ 4) we would expect to observe, defining variants as independent if their pair-wise r^2 cut-off < 0.8. We estimated an average total of 2,848 independent variants within a 2Mb window (**Table S9**). In addition, we observed an average of 2,085 and 778 low frequency and common independent variants, respectively, within a 2Mb window. Only those variants with imputation quality > 0.7 were included in this analysis. Using this information, we estimated the number of independent variants tested in the association analyses. As 2Mb represents ~1/1500 of the genome we extrapolated estimates for the number of independent variants for the circulating biomarkers (all and split by minor allele frequency bin) by multiplying by 1500.

2.3.14 Calculating statistical thresholds for association analyses

For *cis*-eQTLs analyses there were a total of 9,187,579 analyzable variants (MAC ≥ 4 and imputation r^2 > 0.7) that fell within 11,132 2Mb windows around each of the gene expression probes. Given the estimated number of independent variants within a 2Mb region was 2,848 we calculated 2,848 x 11,122 gene expression phenotypes = 31,675,456 independent tests. A *P*-value of 1.6x10^{-9} provides a Bonferroni corrected *P*-value of 0.05 and a *P*-value of ~1x10^{-06} provides a false-discovery rate of ~5% given the number of *cis*-eQTLs we identified at that threshold (1,314).

For the 93 circulating biomarkers we first estimated the number of independent variants across the whole genome by multiplying the number of independent (r^2<0.8) variants in a 2Mb window, 2848, by the approximate number of 2 Mb windows, 1500 = 4,272,000. We multiplied this number by number of circulating biomarkers we were testing to give a total of 397,296,000 independent tests. A *P*-value of 8x10^{-10} provides a Bonferroni corrected *P*-value of 0.05.

2.3.15 Conditional analysis using variants in opposing minor allele frequency bins

For associations that reached our statistical thresholds, we conditioned on the dosage of the most significant variant from the opposing minor allele frequency bin (MAF <5% vs. MAF ≥5%). These variants were limited to those either within the 2Mb *cis*-region of the specific expression trait or within 1Mb of the index variant representing a circulating biomarker association. To ensure that a lack of change in significance of the index variant was not driven by the best variant from the opposite minor allele frequency bin belonging to a secondary signal (creating the potential to miss a partially tagging variant that may not have been the most significant in the opposing bin), we performed full conditional analysis on all traits and conditioned the original index variant identified on the best variant from the opposing allele frequency bin from all additional signals that were observed.

To test further whether or not low frequency variants could explain common signals, we conditioned common signals on all independent low frequency variants (r^2<0.2) with P<1x10^{-4} within the region. Association-based variant clumping was performed using PLINK [34] to identify the variants required for this conditional analysis. Of the 1,232 common signals, 661 had ≥1 low frequency variant in the region meeting these criteria. Three hundred and thirty common signals had two or more low frequency variants that we conditioned on.

2.3.16 Replication of *cis*-eQTLs in the San Antonio Family Heart Study

Whole-genome expression profiling was performed using Illumina Sentrix Human Whole Genome (WG-6) Series I as previously described [35] and called genotypes were provided by the T2D-GENES Consortium. Genotypes were derived either directly from high-pass (60X) whole-genome sequencing or through family-based genotype imputation in the remaining individuals not sequenced. In an attempt to harmonize the WG-6 chip and the HT12-v3 chip we limited our replication efforts to a 397 probe subset of the 1,325 whereby the probe sequences matched across the two platforms. Levels of expression were detected for 233/397 probes in 643 SAFHS individuals. Association analyses were performed using mixed-linear models as implemented in RareFAM that adjusts for a kinship matrix when performing association testing (available online from http://genome.sph.umich.edu/wiki/FamRvTest). One variant from the SAFHS replication results was classified as spurious and removed prior to testing for correlation with the initial *cis*-eQTL results as it had an effect size of >9 standard deviations of an inverse normalised distribution of gene expression levels.

2.3.17 Validation of low frequency variants with bespoke genotyping

We selected 10 low frequency SNPs associated with *cis*-eQTLs and 1 low frequency lactic dehydrogenase variant for genotyping at LGC Genomics, United Kingdom. For 9/11 SNPs we obtained >99% concordance overall. There were two that were returned as monomorphic (both *cis*-eQTL variants) (**Table S17**).

2.4 Results

Analysis of our median 7 fold whole genome sequencing data detected 11,657,229 high quality variants (10,144,717 SNPs and 1,512,512 indels) (see Materials and Methods for definition of high quality) and had a minor allele count (MAC) ≥ 4. Of these variants 6,129,221, 5,528,008 and 2,917,071 had a MAF ≥5% (common), MAF <5% (low frequency), and MAF <1% but minor allele count (MAC) ≥4, respectively. We limited tests to those with a MAC ≥4 because we had limited power to detect associations with 3 or less alleles. A full break down of the numbers of variants tested for each of the analyses is shown in **Table 2**, **Figure 2, and Tables S2-S3**.

A number of analyses provided strong evidence that our data were of high quality (**Materials and Methods, Figs. S2-S5 and Tables S4-S7**). We compared genotypes generated by low pass sequencing with those identified from a separate targeted deep-sequencing (128X) experiment of 2Mb of (non-contiguous) sequence from 83 overlapping individuals. This comparison provided an estimate that 99.4% of the variants identified by low pass sequencing were true positives and a false negative (variants missed by low pass sequencing but detected in deep sequence data) rate of 11.9%. Equivalent figures for indels were in keeping with the increased difficulty of scoring these variants from low pass sequencing data at 84% and 20.7% respectively (**Materials and Methods, Fig. S3 and Tables S4-S7**). Finally, a comparison between *non-reference* SNP genotypes generated by low pass sequencing and GWAS chip data provided strong evidence that genotypes of common variants were accurately (mean genotype concordance of 99.7%) genotyped (**Materials and Methods, Figs. S4 and S5**).

Table 2. A breakdown of the number of variants with minor allele count ≥4 tested in the *cis*-eQTL and circulating biomarker analyses. Details of how we estimated the number of independent variants can be found in Materials and Methods.

	All Variants		**MAF < 0.05**		**MAF ≥ 0.05**	
Analysis	**N variants tested**	**N estimated independent**	**N variants tested**	**N estimated independent**	**N variants tested**	**N estimated independent**
cis-eQTLs	9,187,579	3,480,256	3,760,279	2,547,870	5,427,300	950,716
Biomarkers	11,657,229	4,272,000	5,528,008	3,127,500	6,129,221	1,167,000

All gene expression and biomarker phenotypes were inverse-normalised. A list of the 93 circulating biomarkers can be found in **Table S8**. To assess the number of independent variants and tests we were performing, we randomly selected a 2Mb region from each of the 22 chromosomes and used LDselect [33] and an r^2 cut off of 0.8 to define independent signals, a likely conservative cut-off (**Materials and Methods, Table S9**). We conditioned all single common variant - phenotype associations on the most strongly associated single low frequency variants within the locus (1Mb either side), and vice versa conditioned all low frequency – phenotype associations on more strongly associated common variants in the same region (Materials and Methods). We used several approaches to test the robustness of our associations including testing *cis*-eQTL associations in a replication study - gene expression and high pass (60x) whole genome sequence data from 643 individuals from the San Antonio Family Heart Study (SAFHS) (Materials and Methods) **-** and validation of a subset of 11 low frequency variants with bespoke genotyping (Materials and Methods).

We identified 1,314 *cis*-eQTLs at P<1x10^{-06} and 8 biomarker associations at P<8x10^{-10}, and for *cis*-eQTLs observed a continuous distribution between lower frequency variants of larger effect and higher frequency variants of smaller effect (**Fig. 1, Tables S10 and S11**). Of the 6,129,221 common (>5%) and 5,528,008 low frequency variants tested we identified 0.02% and 0.002% respectively as *cis*-eQTLs at P<1x10^{-06}. Low frequency - large effect associations comprised 7% of detectable *cis*-gene expression traits (89 of 1,314) and 1 of 8 biomarker associations. The average effect size of low frequency index variants was 1.36 (range 0.80-2.39) standard deviations and the average effect size of common index variants was 0.61 (range 0.32-1.73) standard deviations (**Tables S10 and S11**). These differences in per-allele effect size were expected given that lower frequency variants need to have larger per-allele effects to be detected (see **Table 3** for power calculations).

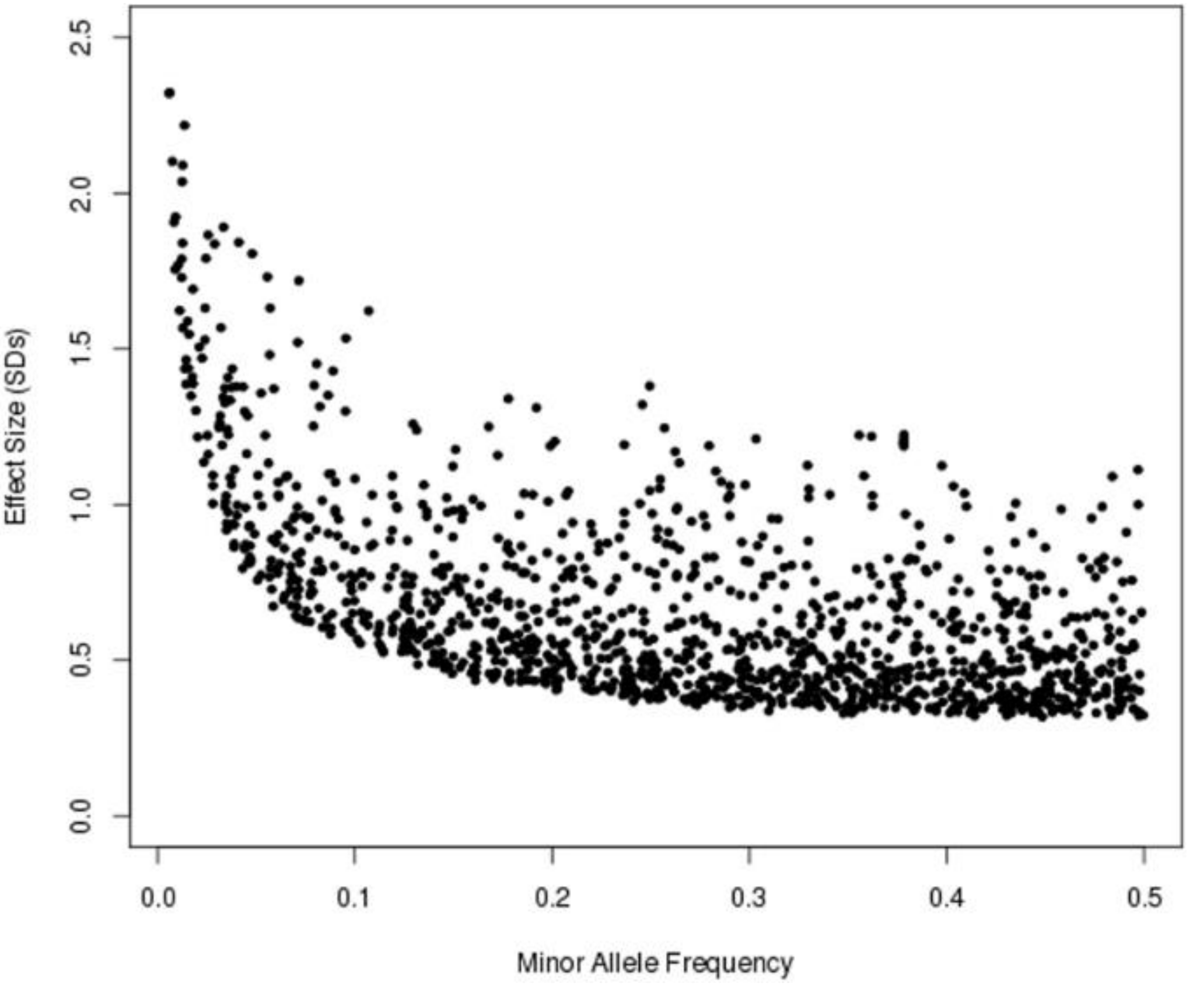


Figure 1. The distributions of effect sizes of index *cis*-eQTL variants by minor allele frequency.

Table 3. Statistical power to detect variants associated with gene expression in 450 individuals at *P*=1x10^{-6} as a function of phenotypic variance explained and standard deviation (SD) effect size.

MAF	Variance	Power	Effect (SD)	Variance	Power	Effect (SD)	Variance	Power	Effect (SD)
0.01	0.05	0.47	1.59	0.06	0.65	1.74	0.07	0.79	1.88
0.02	0.05	0.47	1.13	0.06	0.65	1.24	0.07	0.79	1.34
0.03	0.05	0.47	0.93	0.06	0.65	1.02	0.07	0.79	1.10
0.04	0.05	0.47	0.81	0.06	0.65	0.88	0.07	0.79	0.95
0.05	0.05	0.47	0.73	0.06	0.65	0.79	0.07	0.79	0.86
0.1	0.05	0.47	0.53	0.06	0.65	0.58	0.07	0.79	0.62
0.2	0.05	0.47	0.40	0.06	0.65	0.43	0.07	0.79	0.47
MAF	Variance	Power	Effect (SD)	Variance	Power	Effect (SD)	Variance	Power	Effect (SD)
0.01	0.08	0.89	2.01	0.09	0.95	2.13	0.10	0.98	2.25
0.02	0.08	0.89	1.43	0.09	0.95	1.52	0.10	0.98	1.60
0.03	0.08	0.89	1.17	0.09	0.95	1.24	0.10	0.98	1.31
0.04	0.08	0.89	1.02	0.09	0.95	1.08	0.10	0.98	1.14
0.05	0.08	0.89	0.92	0.09	0.95	0.97	0.10	0.98	1.03
0.1	0.08	0.89	0.67	0.09	0.95	0.71	0.10	0.98	0.75
0.2	0.08	0.89	0.50	0.09	0.95	0.53	0.10	0.98	0.56

Our low pass sequencing approach meant that we were able to accurately capture and analyse a similar number of low frequency variants (5,528,008, including 2,917,071 at allele frequency <1%) as common variants (6,129,221). However, proportional to the number of variants analysed, we detected far fewer low frequency variant associations than common variant associations despite the same statistical power to detect individual variants explaining the same proportion of phenotypic variance. In **Table 3** we show how statistical power remains fixed for variants explaining similar proportions of phenotypic variance, but how standard deviation effect sizes need to be higher for lower frequency variants. In total, we identified 89 low-frequency *cis*-eQTLs and 1 low frequency – circulating biomarker association compared to 1,225 common *cis*-eQTLs and 7 common biomarker associations (**Tables S10 and S11**). Accounting for linkage disequilibrium between variants accentuated this difference – for *cis*-eQTLs the low frequency variant associations represented 0.003% of an estimated 2,547,870 independent low frequency variants (where independence was defined as r^2<0.8), while the common variant associations represented 0.12% of an estimated 950,716 independent common variants (r^2<0.8) (**Table 2 and Fig. 2**). These comparisons were not influenced by differences in quality of genotypes between low frequency and common variants because we only compared variants of high quality. Under an alternative genetic architecture, we could have expected to identify 1105 low frequency variants associated with *cis* gene expression (0.02% of 5,528,008 low frequency variants tested – the same proportion of common variants associated with gene expression), given we had the same statistical power to detect single low frequency - large effect variants that explain a similar proportion of phenotypic variance as single common smaller effect variants. Instead our data are consistent with the argument that only a small proportion of single low frequency variants will have large enough per-allele effects to explain a similar proportion of phenotype variance as single common variants. Our data do not rule out the possibility that many 1000s of low frequency variants of moderate and small effect could collectively account for more phenotype variance than common variants collectively.

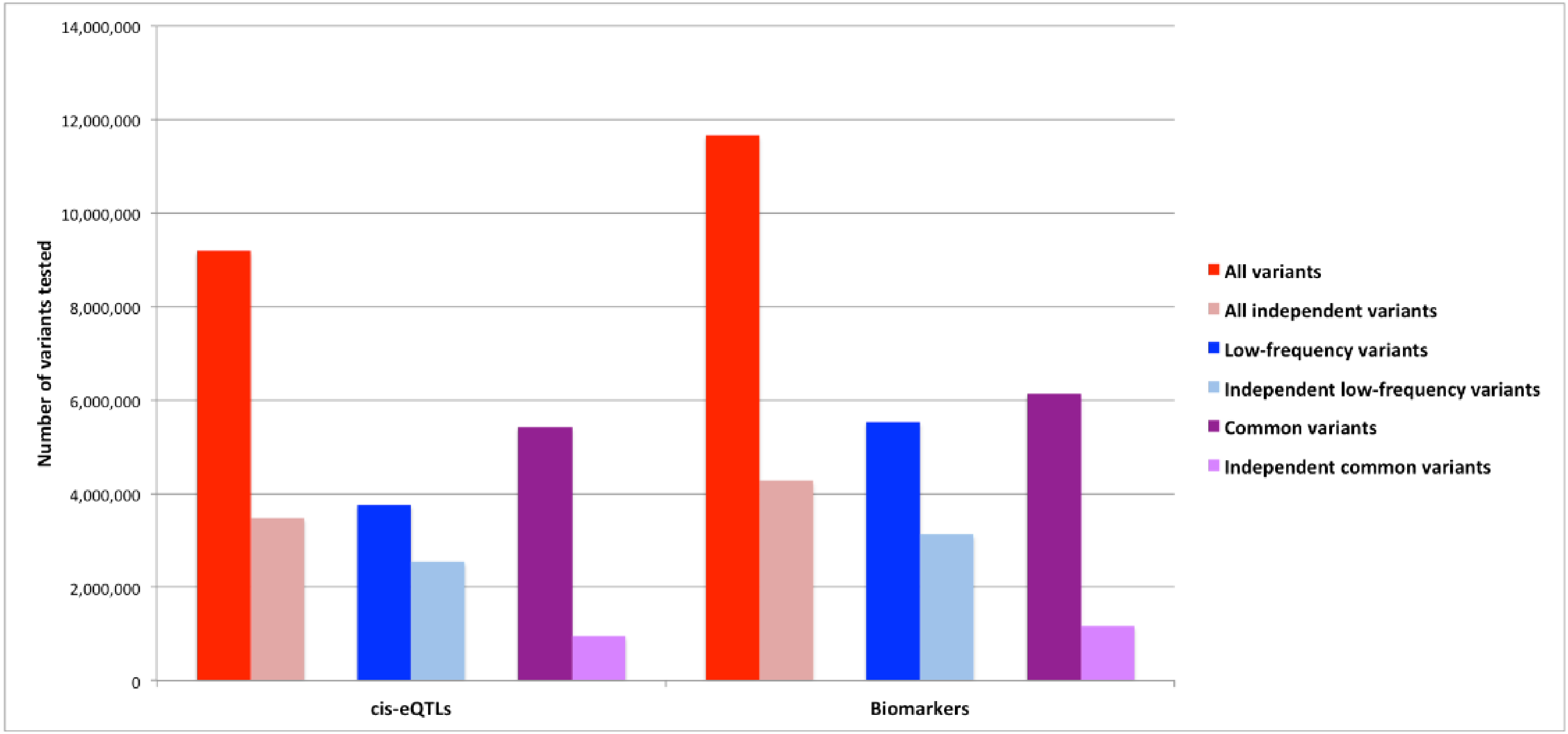


Figure 2. The total number of variants (low frequency and common) tested in the *cis*-eQTL and circulating biomarker analyses.

Analyses of gene expression phenotypes in a second dataset suggested that the associations observed were robust – of 233 *cis*-eQTLs associations where the same gene was probed with the same expression probe sequence in a second study of similar size (N=643 related individuals, Materials and Methods), we detected 166 associated at a Bonferroni corrected *P*-value of <0.0002 and 222 of 236 were directionally consistent. Of these 12 of 17 testable low frequency associations reached *P*<0.0002 and all 17 were directionally consistent. For example, low frequency variants in or near the genes *ACAD9*, *HDHD3*, *SOS1, UTS2*, *RTN1* and *RBPMS2* influenced the expression of those genes with per-allele effects of >1 standard deviation in the replication data, where winner's curse would not have appreciably influenced the effect size. No data were available to replicate the single low frequency variant associated with a biomarker. This biomarker was Lactate dehydrogenase and we could not identify any studies with relevant measures (**Table S12**).

We next assessed the extent to which common variant associations were driven by low frequency associations, and vice versa, by conditioning on the most strongly associated variants in the alternative allele frequency bin. All evidence of association was lost (*P*>0.05) for 13 of 1,232 common variant associations when conditioning on the strongest low frequency variant and 969 of 1,232 remained associated at our statistical thresholds. We next repeated the analyses but conditioned all common variant associations on all independent (r^2<0.2) low frequency variants reaching P<1x10^{-04} in the cis region of the expression probe or within 1Mb of the common index variants for biomarker associations. All evidence of association was lost (*P*>0.05) for 30 of the 661 (5%) common signals that had at least one low frequency association at the same locus. These results strongly suggested that few of the common associations were driven by single or multiple low frequency variants. For low frequency variants, all evidence of association was lost (*P*>0.05) for 11 of 90 associations and 47 of the 90 remained associated at our thresholds when adjusting for the strongest common variant in the region (46 of the 89 low frequency *cis*-eQTL signals and 1 of 1 low frequency biomarker associations) (**Tables S13 and S14**).

The availability of 1000 Genomes reference sequence has improved the ability to accurately capture low frequency variants and may mean that low pass sequencing individual studies is an inefficient use of research funds. We therefore next assessed how well low frequency variant associations would have been detected without any sequence data from the InCHIANTI study but by using data from the 1000 Genomes Project as a reference panel for imputation and the GWAS array (Illumina HumanHap550K) genotypes as a scaffold. We used an imputation reference panel comprised of 2,184 haplotypes from 1,092 individuals sequenced and phased by the 1000 Genomes Project. Of these, 379 individuals were of European descent and included 98 individuals from Tuscany (the same part of Italy as the InCHIANTI study)*.* Of the 90 low frequency signals (89 *cis*-eQTLs, 1 biomarker), detected by sequencing, we did not detect 63% (57) based on the same statistical thresholds using 1000 Genomes imputation alone (all *cis*-eQTLs) (**Tables S15 and S16**). Ignoring statistical thresholds, 85% of all *cis*-eQTL and biomarker associations identified through sequence-based analysis were less strongly associated in the 1000 Genomes imputed dataset or had no proxy within 250Kb of the index variant (**Tables S15 and S16**). However, 62 (69%) of the 90 low frequency variant associations were detected at *P*<0.0001. Twenty of the 28 associations not detected in 1000 Genomes were not well imputed (mean $r^2 = 0.31$) with the remaining 8 being due to the genotypes not being present 1000 Genomes. These results illustrate that imputation from the 1000 Genomes reference panel captures most of the low frequency variants, just not as accurately.

2.5 Discussion

Our study provides an early example of a whole genome sequencing experiment designed to identify low frequency variants associated with common human phenotypes. We could accept our first main hypothesis. We show that, when using the same sample size for detecting both common and low frequency variants, whole genome sequencing has the ability to identify low frequency variants with larger effect sizes (and similar phenotypic variance explained) than those observed for common variants. However, our data suggest that these single-variant low frequency large effect signals may represent a relatively small proportion (here 7%) of detectable associations. Our data are consistent with the majority of single low frequency genetic variants explaining smaller proportions of phenotypic variance than single common genetic variants. This result was perhaps expected, but given our sequencing based approach allowed us to test a similar number of high quality low frequency variants (5,528,008) as common variants (6,129,221) we could reasonably have expected to identify approximately 1105 low frequency *cis*-eQTLs under a different genetic architecture. If extrapolated to other common human phenotypes, our results would indicate that, for example, whole genome sequenced (or perhaps extremely well imputed) sample sizes of 35,000 cases and equivalent controls will be needed to detect approximately 5 single low frequency variants associated with type 2 diabetes (7% of 65 [36]). These are obviously very cautious extrapolations because the genetic architecture of *cis* gene expression and circulating biomarkers may be very different compared to other phenotypes. Nevertheless, there is strong evidence that changes to *cis* gene expression is a common mechanism leading to common disease and quantitative phenotypes [11-

16] and a recent whole genome sequence and imputation based study provides evidence that these estimates may be of the correct order of magnitude – an effective sample size of 13,500 type 2 diabetes cases detected 5 low frequency (MAF <5%) large effect type 2 diabetes associations [8].

Our data enabled us to largely reject our second main hypothesis – that common variant - common phenotype associations are explained by individual or multiple low frequency variants. Only 30 (2%) common variant associations were driven by low frequency (and by necessity larger effect) variants. In contrast, a larger fraction (12%) of low frequency variants was entirely driven by common variant associations at the same locus. We also note that, as with most sequencing studies, low frequency insertion/deletion variants were harder to call and we may have missed true associations caused by these types of variant. Of note, 13% (12) of the 90 low frequency - large effect and 11% (137) of the 1,232 common variant association signals included an indel as the most strongly associated variant.

In keeping with other recent studies we were able to accept our secondary hypothesis. Our data clearly show that whole genome sequencing is more effective than imputation from the current 1000 Genomes Project reference panel. Imputation of missing genotypes missed 63% of low frequency - large effect associations detected by whole genome sequencing at $P<1\text{x}10^{-6}$, and 31% at $P<1\text{x}10^{-4}$. Nevertheless larger reference panels will improve the ability of imputation to capture low frequency variants and it is notable that 1000 Genomes imputation captured at least half the genotype information ($r^2 > 0.52$) for 75% of the low frequency signals. As noted by other studies [7], these findings emphasize the need for larger reference panels from which to impute missing genotypes into extremely large GWAS datasets.

There were a number of limitations to our study. First, our conclusions are based largely on testing 1000s of *cis* gene expression phenotypes rather than the whole genome for a small number of phenotypes. However, the costs of whole genome sequencing have so far limited single study sample sizes to less than 3000 samples, and our approach had the advantage that we were well powered to detect many common variant – common phenotype associations. This advantage meant we were able to make a fair comparison between common and low frequency variants. Our data may also be relevant to disease phenotypes because numerous studies have shown that many disease associations are enriched for *cis* gene expression effects [11-16]. A second limitation is that we did not assess more detailed phenotypes or genotype combinations. For example, a recent study has shown that low frequency variants may be a frequent cause of allele and exon specific changes to gene expression [37]. Furthermore we could not test the role of most rare variants (<0.5%) in our study because our sample size limited analyses to those occurring at 0.3% frequency or more (minor allele count ≥ 4). However, our approach meant we were able to analyse 2.9 million accurately called variants with allele frequencies of less than 1%.

In conclusion, our approach provided an unbiased assessment of the relative contribution of low frequency and common genetic variation to common quantitative phenotypes of relevance to human disease. Our study shows that low pass whole genome sequencing can identify low frequency - large effect variants in common human phenotypes using sample sizes sufficiently large to provide statistical power to detect large numbers of common variant associations.

2.6 References

1. Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, et al. (2013) Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature 498: 232-235.

2. Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, et al. (2012) A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature 488: 96-99.

3. Kong A, Thorleifsson G, Frigge ML, Masson G, Gudbjartsson DF, et al. (2014) Common and low-frequency variants associated with genome-wide recombination rate. Nat Genet 46: 11-16.

4. Morrison AC, Voorman A, Johnson AD, Liu X, Yu J, et al. (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet 45: 899-901.

5. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, et al. (2014) A polygenic burden of rare disruptive mutations in schizophrenia. Nature 506: 185-190.

6. Service KS, Teslovich TM, Fuchsberger C, Ramensky V, Yajnik P, et al. (2014) Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. PLoS Genet 10: e1004147.

7. Genome of the Netherlands C (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46: 818-825.

8. Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, et al. (2014) Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet 46: 294-298.

9. Melzer D, Perry JR, Hernandez D, Corsi AM, Stevens K, et al. (2008) A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet 4: e1000072.

10. Wood AR, Perry JR, Tanaka T, Hernandez DG, Zheng HF, et al. (2013) Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation. PLoS One 8: e64343.

11. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, et al. (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 43: 246-252.

12. de Jong S, van Eijk KR, Zeegers DW, Strengman E, Janson E, et al. (2012) Expression QTL analysis of top loci from GWAS meta-analysis highlights additional schizophrenia candidate genes. Eur J Hum Genet 20: 1004-1008.

13. Dubois PC, Trynka G, Franke L, Hunt KA, Romanos J, et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 42: 295-302.

14. Montgomery SB, Lappalainen T, Gutierrez-Arcelus M, Dermitzakis ET (2011) Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet 7: e1002144.

15. Small KS, Hedman AK, Grundberg E, Nica AC, Thorleifsson G, et al. (2011) Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat Genet 43: 561-564.

16. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, et al. (2013) Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 45: 1238-1243.

17. Ferrucci L, Bandinelli S, Benvenuti E, Di Iorio A, Macchi C, et al. (2000) Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. J Am Geriatr Soc 48: 1618-1625.

18. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, et al. (2011) Modernizing reference genome assemblies. PLoS Biol 9: e1001091.

19. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

20. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.

21. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311.

22. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

23. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.

24. Browning BL, Yu Z (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet 85: 847-861.

25. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, et al. (2011) The variant call format and VCFtools. Bioinformatics 27: 2156-2158.

26. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10: 387-406.

27. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34: 816-834.

28. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44: 955-959.

29. Wood AR, Hernandez DG, Nalls MA, Yaghootkar H, Gibbs JR, et al. (2011) Allelic heterogeneity and more detailed analyses of known loci explain additional phenotypic variation and reveal complex patterns of association. Hum Mol Genet 20: 4082-4092.

30. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337: 64-69.

31. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, et al. (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26: 2867-2873.

32. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42: 348-354.

33. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74: 106-120.

34. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

35. Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet 39: 1208-1216.

36. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44: 981-990.

37. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506-511.

2.7 Supplementary Information

Tables S1-S17. Accessible from: https://goo.gl/LA7XbC

Figure S1. The average percentage of the genome covered at different read depths across the 680 InCHIANTI subjects

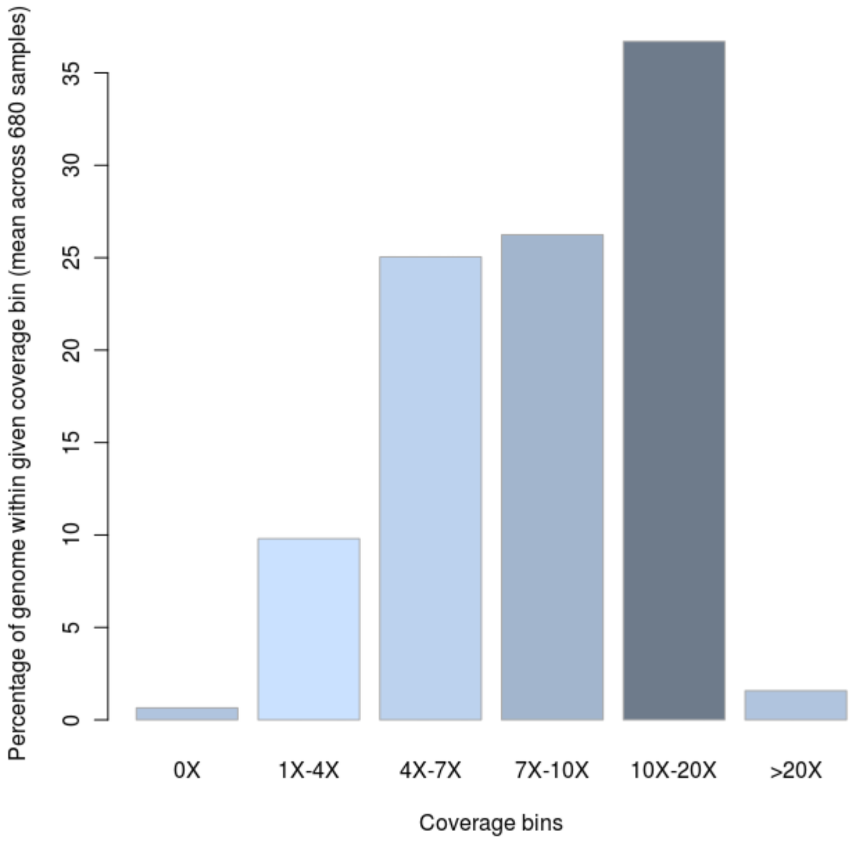
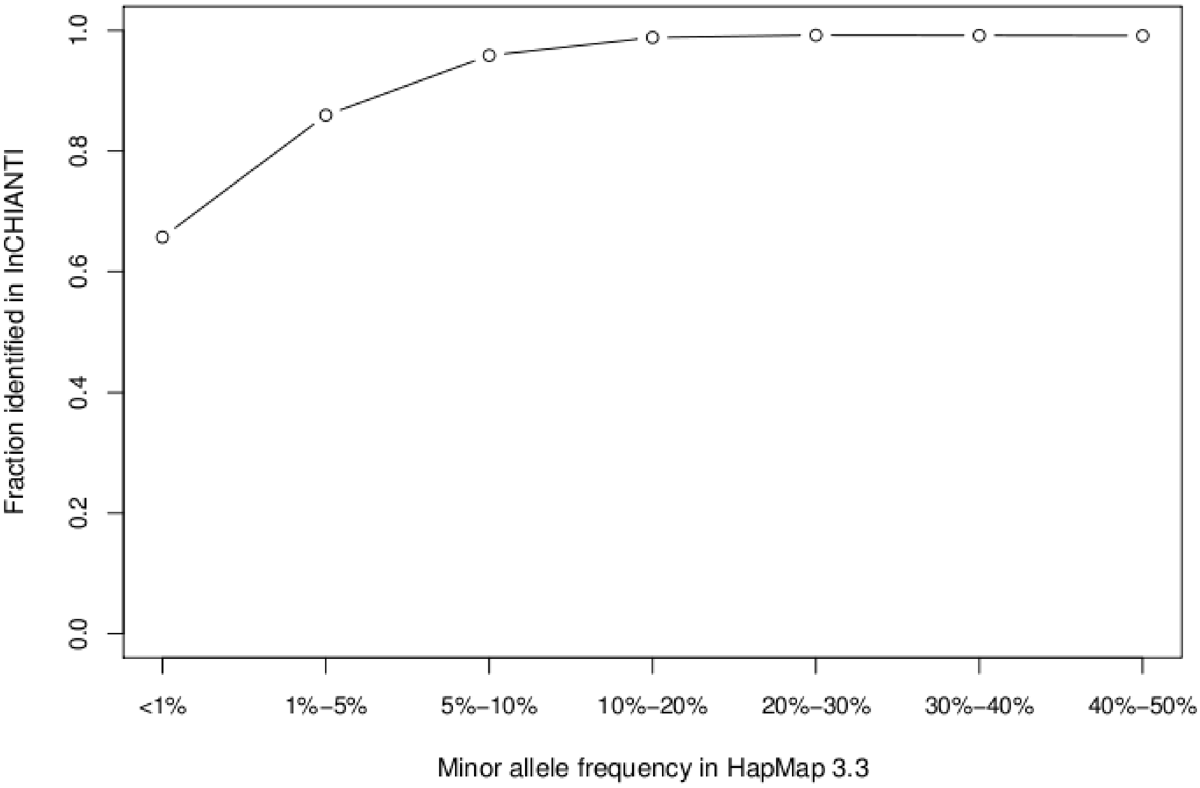
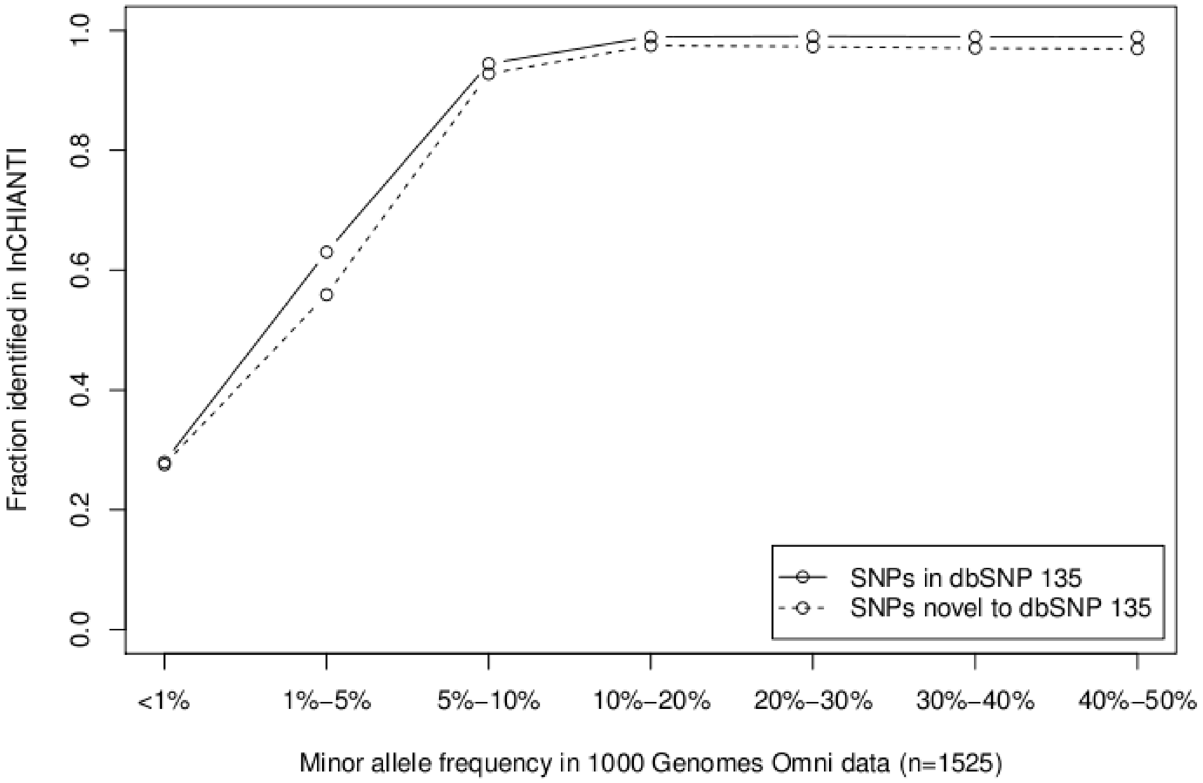


Figure S2. The fraction of SNPs found in HapMap, 1000 Genomes Omni 2.5 genotyping array and indels present in the 1000 Genomes project phase 1 indel datasets, binned by catalogued minor allele frequency. **a)** 1,294,681 (97.4%) SNPs present in HapMap Phase 3 release 3: 1,329,031 SNPs from CEU, GBR, TSI and Intersecting populations; **b)** 1,677,600 (76.9%) SNPs with 1,038,907 novel to dbSNP 135 present in 2,181,344 SNPs with 1,504,747 novel to dbSNP 135) on the 1000 Genomes Omni2.5 high-density genotyping array; **c)** 831,414 (57.6%) indels with 273,595 novel to dbSNP 135 present in 1,443,514 indels with 700,122 novel to dbSNP 135 from the 1000 Genomes phase 1 dataset.

a)



b)


Fraction identified in InCHIANTI
0.0
0.2
0.4
0.6
0.8
1.0
<1%
1%–5%
5%–10%
10%–20%
20%–30%
30%–40%
40%–50%
Minor allele frequency in 1000 Genomes Omni data (n=1525)
SNPs in dbSNP 135
SNPs novel to dbSNP 135

c)

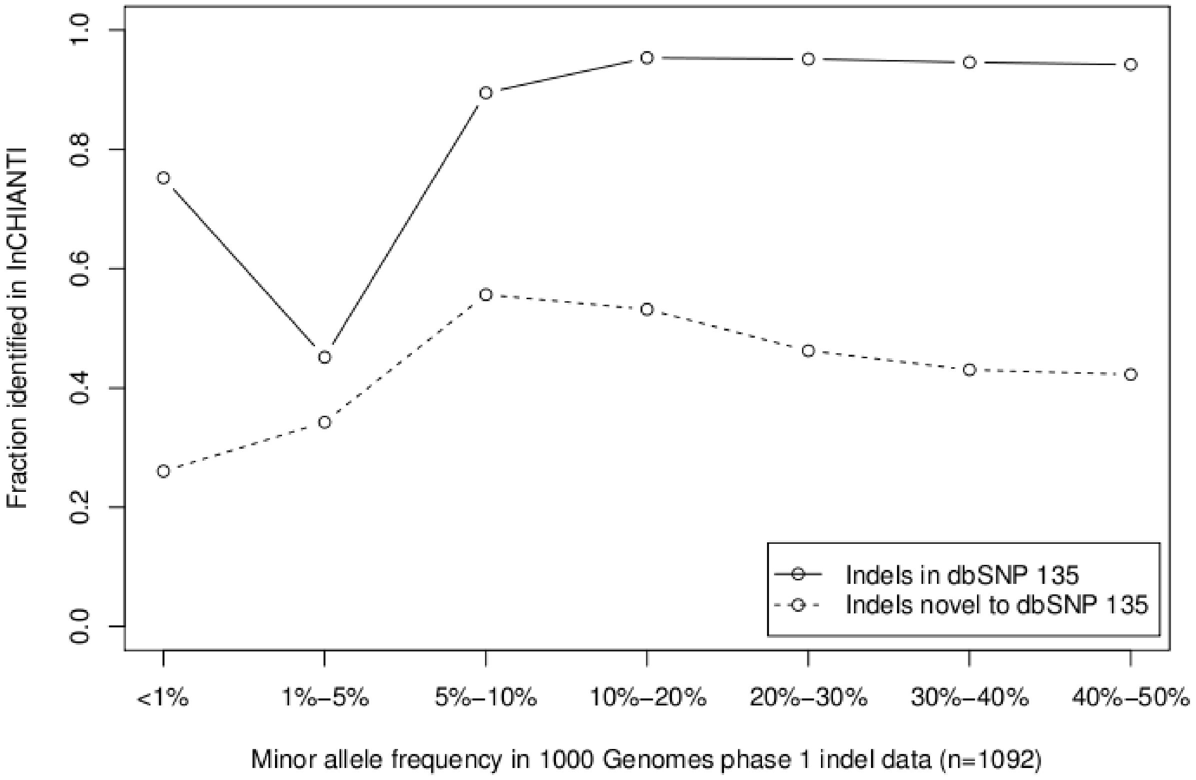
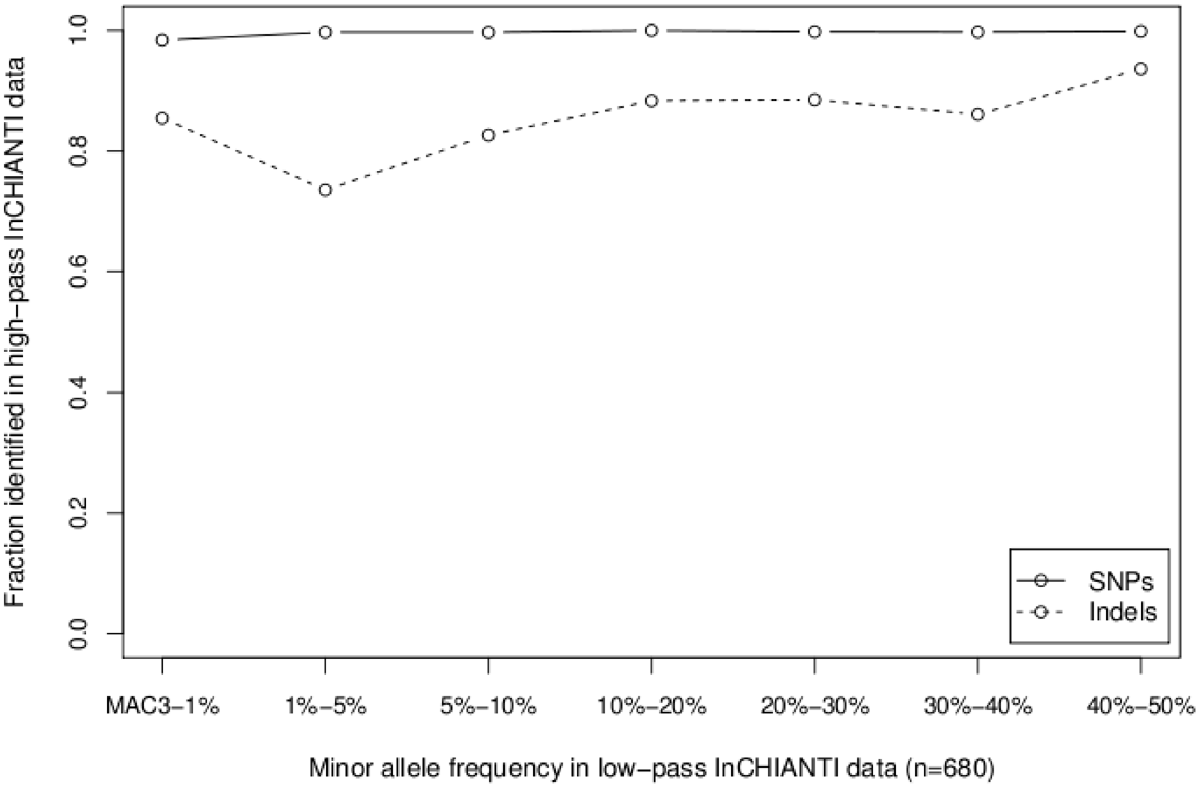

Fraction identified in InCHIANTI
0.0
0.2
0.4
0.6
0.8
1.0
<1%
1%–5%
5%–10%
10%–20%
20%–30%
30%–40%
40%–50%
Minor allele frequency in 1000 Genomes phase 1 indel data (n=1092)
Indels in dbSNP 135
Indels novel to dbSNP 135

Figure S3. a) The fraction of 10,389 variants discovered through low-pass sequencing (9,402 SNPs and 987 indels) that overlap with 10,167 variants identified through high-pass sequencing (9,342 SNPs: 99.4% and 825 indels: 84%); **b)** the fraction of 11,649 variants discovered through high-pass sequencing (10,609 SNPs and 1,040 indels) that overlap with 10,167 variants identified through low-pass sequencing (9,342 SNPs: 88.1% and 825 indels: 79.3%)

a)



b)

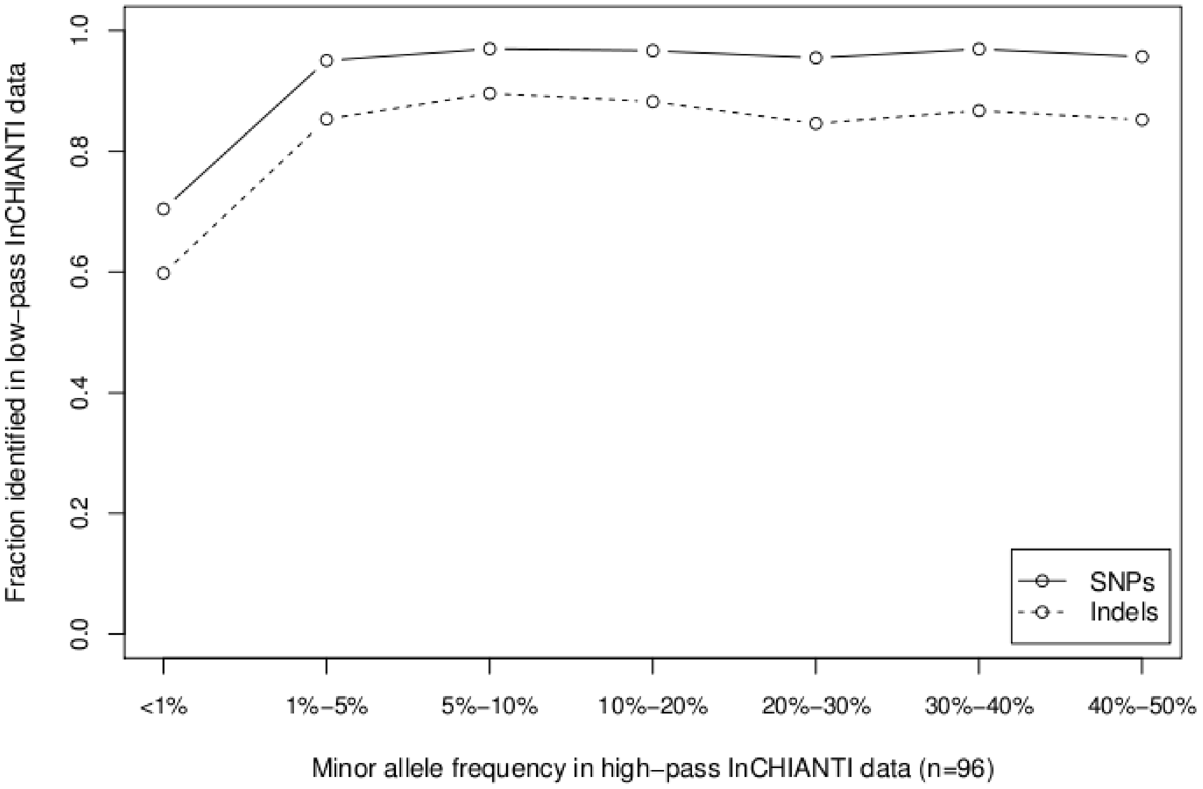

1.0
0.8
0.6
0.4
0.2
0.0
Fraction identified in low−pass InCHIANTI data
SNPs
Indels
<1%
1%−5%
5%−10%
10%−20%
20%−30%
30%−40%
40%−50%
Minor allele frequency in high−pass InCHIANTI data (n=96)

Figure S4. Plot of overall sequence-to-chip genotype concordance post Beagle refinement of the 673 “concordant” subjects. Subjects are ranked (indexed) in order of least to most concordance with the microarray *post* Beagle Imputation.

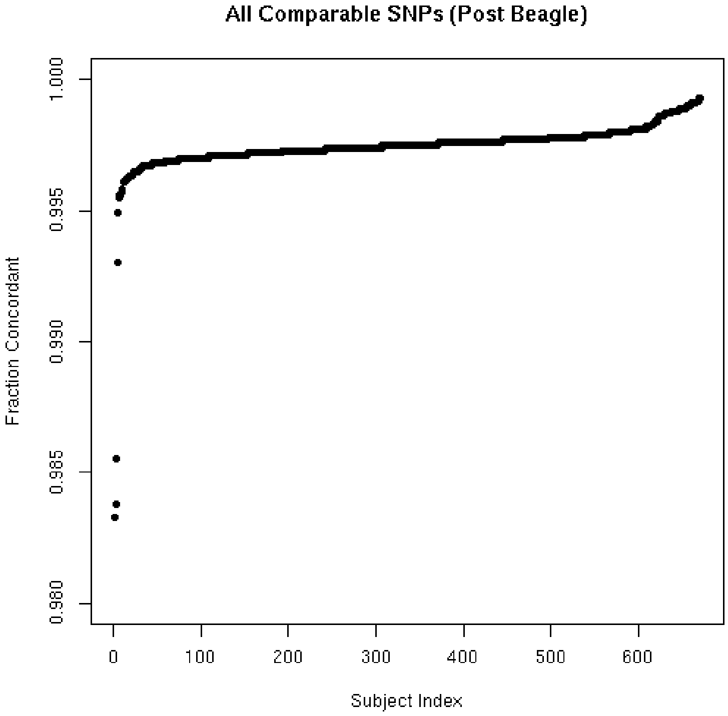
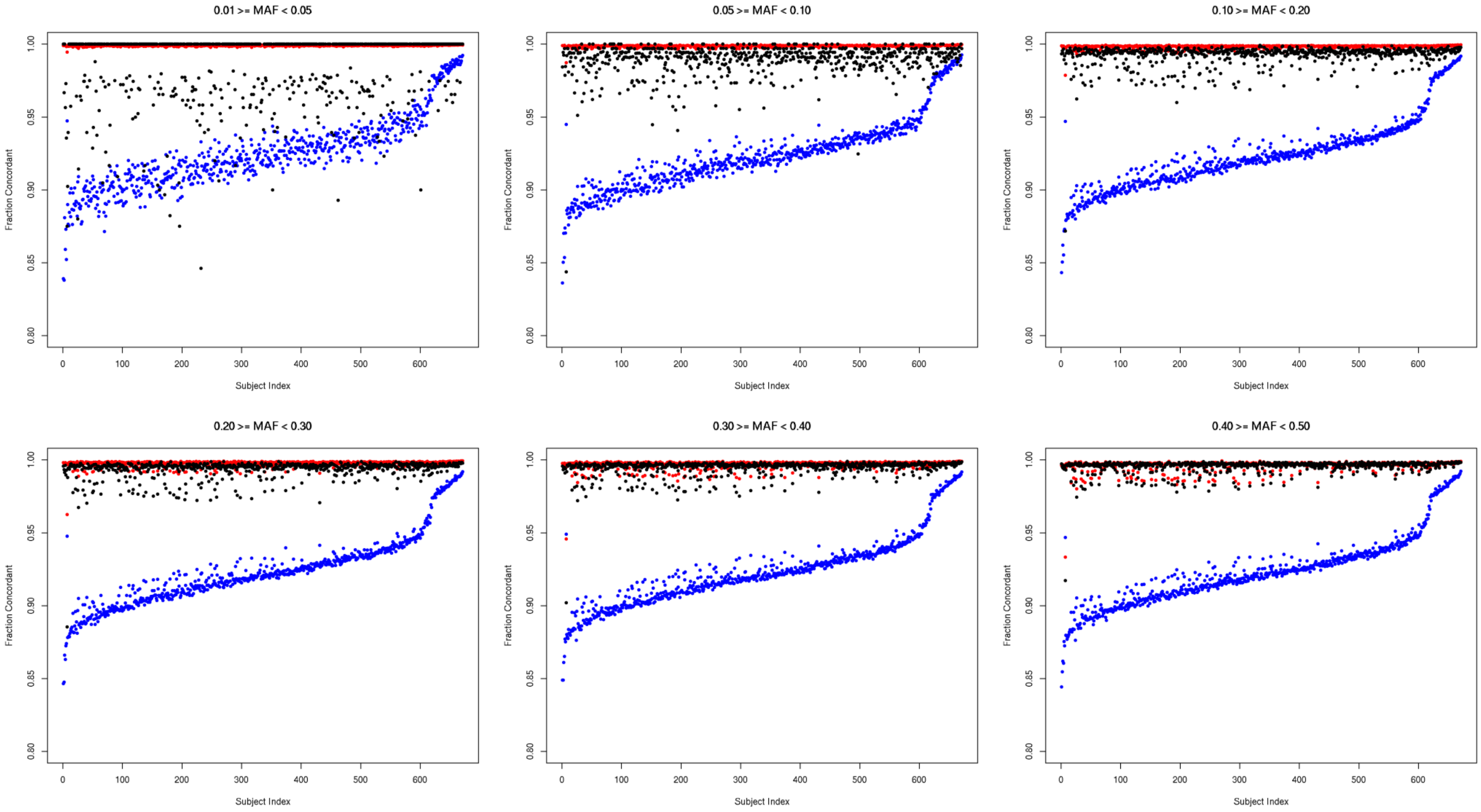
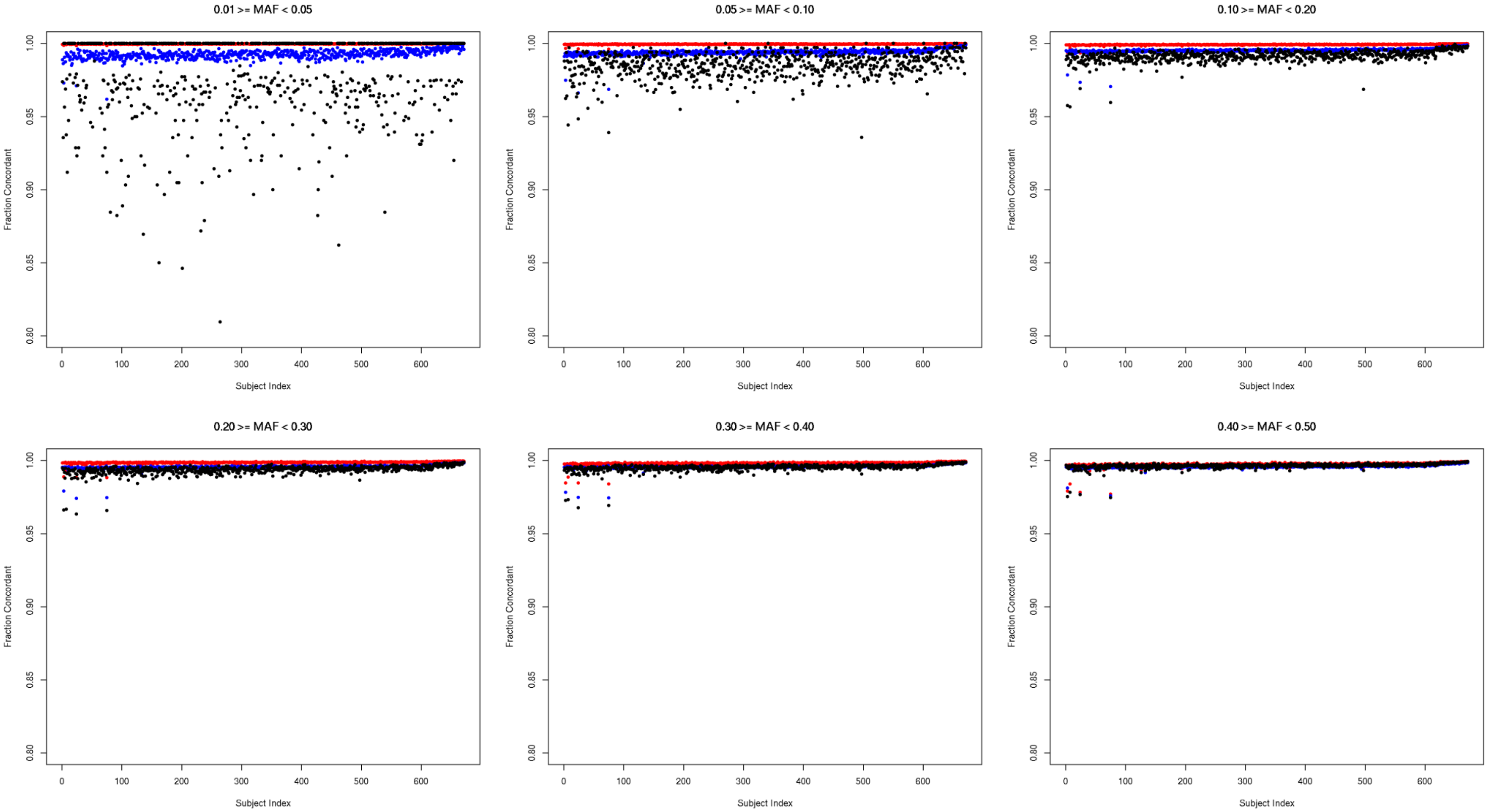


Figure S5. Sequence to Illumina HumanHap550 genotyping concordance for 673 individuals stratified by genotype as defined on chip. Red = common homozygous genotypes on chip; Blue = heterozygous genotypes on chip; Black = rare homozygous genotypes on chip. **a)** before beagle phasing and imputation; **b)** after beagle phasing and imputation. Subjects are ranked (index) in order of least to most concordant *prior* to beagle imputation overall.

a)



b)


0.01 >= MAF < 0.05
0.05 >= MAF < 0.10
0.10 >= MAF < 0.20
0.20 >= MAF < 0.30
0.30 >= MAF < 0.40
0.40 >= MAF < 0.50
Fraction Concordant
Subject Index
0.80
0.85
0.90
0.95
1.00
0
100
200
300
400
500
600

CHAPTER 3

AMY1 gene copy number detected in 657 whole genome sequenced individuals is not associated with body mass index

Adapted from: Usher, C.L., Handsaker, R.E., Esko, T., Tuke, M.A., Weedon, M.N., Hastie, A.R., Cao, H., Moon, J.E., Kashin, S., Fuchsberger, C. *et al.* (2015) Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nature genetics*, 47, 921-925.

3.1 Abstract

A recent publication by Falchi et al [1] identified an association between a multi-allelic CNV encompassing the salivary amylase gene (*AMY1*) and obesity. *AMY1* copy number was reported to have a larger effect on obesity than SNPs at *FTO*, the largest effect detected in GWAS. We aimed to replicate these findings with whole-genome sequence data in 657 individuals from the InCHIANTI cohort, a longitudinal study of ageing in Tuscany, Italy. We called *AMY1* copy number using two different computational methods: MrsFAST/mrCaNaVaR, a read-depth based aligner and CNV calling suite, and GenomeSTRiP, a read-depth based CNV caller that classifies copy number based on population-level information. We then tested for association between the resulting CNV calls and BMI. We observed a distribution of *AMY1* CNVs consistent with most individuals carrying an odd number of *AMY1* copies, which sum to an even number in the diploid human genome. We did not observe any association between *AMY1* copy number and BMI with either mrCaNaVaR (P = 0.64) or GenomeSTRiP called genotypes (P = 0.53). The *FTO* (P = 0.074) SNP and a polygenic score (P = 0.001) of 11 BMI SNPs were associated with BMI in the same dataset. *AMY1* copy number did not associate with BMI in the InCHIANTI cohort despite good statistical power to replicate the reported effect sizes. These results were also replicated in three additional European cohorts. These results contrast with previous results that *AMY1* copy number exerts a far-stronger effect on BMI and obesity than SNPs at FTO and other loci.

3.2 Introduction

The three amylase genes reside in a structurally complex part of the genome with multiple complex configurations of deletions, and duplications [2,3]. Each of the three amylase genes vary widely in copy number, with *AMY1* varying from 2-17 copies [4] and *AMY2A* from 0-8 [5]. Amylase has a role in starch metabolism whereby *AMY1* is secreted by the salivary gland and *AMY2* is secreted by the pancreas. A recent publication [1] identified significant association of the multi-allelic CNV encompassing the salivary amylase gene (*AMY1*) with body mass index (BMI) and obesity. The study found associations between *AMY* locus copy number and obesity/BMI using an array-based family study in 342 Swedish individuals. *AMY1* associations were replicated using qPCR copy number measures in ~6,200 individuals from cohorts of northern European and Asian descent. A reduced copy number of the *AMY1* gene was associated with an odds ratio of 1.19 per copy for obesity compared to normal weight individuals. The effect of *AMY1* copy number reported to explain 11% of the genetic contribution to obesity, far greater than the effect of SNPs at *FTO*, the largest effect previously detected in GWAS [6]. A previous GWAS of 339,224 individuals [7] was unable to detect this association. These findings have raised questions about the completeness of GWAS [1].

Complex CNVs such as that of *AMY1* are extremely difficult to characterise [8], and CNV association studies often involve rough copy number estimates that can be confounded by influences such as methods of detection or DNA quality [9-11]. Here, we aimed to replicate published findings and identify and validate copy number of the *AMY1* locus using detection methods applied to 657 whole human genomes from the InCHIANTI cohort [12]. We then aimed to test the significance of the relationship between BMI and *AMY1* copy number.

3.3 Materials and Methods

3.3.1 Samples

We selected 680 individuals from the InCHIANTI study [12], a study of aging from the Chianti region in Tuscany, Italy, for low pass whole genome sequencing [13]. Individuals were selected for sequencing based on the availability of gene expression and circulating biomarker data.

3.3.2 Whole genome sequencing

Whole-genome sequencing was performed at the Beijing Genomics Institute (BGI), Shenzhen, China using the Illumina HiSeq 2000 to obtain a minimum mean read depth of 6X (median 7X). An average of 240 million paired-end 90bp reads per sample were aligned to the 1000 Genomes implementation of the Genome Reference Consortium's build 37 of the human reference genome [14], using two different alignment and copy number calling methods.

3.3.3 Sequence read mapping

Burrows-Wheeler Aligner (BWA) version 1.5.910 [15] and MrsFAST-Ultra version 3.3.1 [16] were used to align the sequencing reads. BWA and MrsFAST are optimised for different purposes, but can be applied to different copy number calling methods:

3.3.3.1 BWA read mapping

BWA is a general purpose aligner that maps each read to its most likely single position in the reference genome. We mapped reads to the reference genome as paired-end. Each read had a mapping quality assigned to it giving the likelihood of the read mapping to its position over other positions in the genome. A mapping quality of zero was assigned to reads that had an equal likelihood of being mapped to many regions in the genome. If the mapping quality of a read was zero, its actual position was randomly assigned amongst all of its possible mappings.

3.3.3.2 MrsFAST read mapping

MrsFAST aligns each read to the reference genome one or more times wherever there is a match that fits within predefined thresholds. This method has been shown in some studies to be a requirement of absolute copy number detection [17]. The reads were treated as single-end to allow reads to map to multiple positions. A hamming distance threshold of 4 was set to determine how close a match should be to the reference genome. The hamming threshold is defined in this case as the number of base changes, and/or gaps that need to be added for there to be a perfect alignment. Repeat masking was applied to the reference for MrsFAST alignment. We masked repeat sequences in the reference genome using both RepeatMasker Open-3.0 and Tandem Repeats Finder 4.07b [18]. RepeatMasker masks nucleotides that are in regions of reduced DNA complexity and interspersed regions of DNA such as retrotransposons. Tandem Repeats Finder searches for contiguous copies of patterns of 12 or less nucleotides. These two methods were combined to mask a total of 46.2% (1,457,927,398 bases) of the GRC build 37 reference genome. In addition to the 7.5% (237,019,517 bases) of un-mappable regions in the reference genome, gave a total of 53.7% of the GRC build 37 reference genome as masked and not callable for absolute copy number in MrsFAST aligned short read sequence data.

3.3.4 Defining the *AMY1* region

We extracted 'triplicated' regions to maximise likelihood that each read mapping to the *AMY1* locus is at its genuine position. Although we masked out short repeats and low-complexity DNA, we also needed to account for the three *AMY1* paralogs occurring in the reference genome. The reference paralogs are known as *AMY1A*, *AMY1B* and *AMY1C*. Reads mapping to the genome greater than three times were discarded. To detect these regions we first split the reference genome into 90bp k-mers corresponding to the length of our sequencing reads. Each unique 90bp k-mer was then mapped back to the reference genome allowing each read to be mapped multiple times using MrsFAST with a hamming distance threshold of 4% of each read (4bp). We then removed reads that mapped to the genome greater than or equal to three times. We calculated read depth using the regions that are covered by the remaining reads.

3.3.5 Calling absolute copy number

We used two separate methods for detecting absolute copy number in the *AMY1* regions (**Fig. S1**). To detect copy number from BWA aligned reads we used Genome STRiP multi-allelic CNV (mCNV) caller [19,20]. To detect copy number from MrsFAST alignments, we used mrCaNaVar version 0.51 [17]. Both of these approaches estimate absolute copy number using next generation sequence data based solely on GC-bias corrected read depth. Both approaches take GC-bias as the ratio between the number of G and C bases in a given region and the number of reads mapped to it. A positive or negative deviation from the GC curve indicates a mapping confounder that can distort read depth and subsequently copy number.

3.3.5.1 Genome STRiP copy number calling

Genome STRiP calculates absolute copy number for each individual based on 'triplicated' loci defining *AMY1* as previously described. GC-bias is calculated by masking regions unlikely to be copy number polymorphic and then creating GC profiles of multiplicative correction [21]. Total average read depth across all diploid masked regions of the genome were regressed against the average read depth across regions of a specific GC content using LOESS linear regression to give a GC-bias correction factor [21]. Copy number was then calculated by taking a mean GC corrected depth of all reads with a mapping quality threshold of 30 and calculating the number of reads per copy. The reads were then classified into copy number categories on a population level for all 657 samples using constrained Gaussian mixture models [19]. The mixture model is based on copy number classes of integers ranging from zero to the maximum detected copy number in that population for the specified region.

3.3.5.2 mrCaNaVaR copy number calling

mrCaNaVaR calculated absolute copy number values for the whole genome in 1kb non-overlapping depth windows. GC base percentage was calculated for each of these windows. Depth was GC-bias normalised using LOESS regression based on GC percentage for all windows in the genome [21]. Copy number for each window was then calculated from its deviation from overall GC corrected bin depth distribution for each sample. We then intersected the 1kb copy number bins against our defined triplicated amylase region using Bedtools version 2.20.1 [22]. We were then able to produce a copy number estimate for each sample based on the mean of all intersected regions.

3.3.6 Quality control

Of the full 680 subjects, 7 were selected for exclusion based on SNP chip discordance in previously called SNP data [13]. A further 16 genotypes were removed for having missing phenotype data in either BMI, age or sex resulting in 657 genomes being used for analysis. We checked the distribution of called copy number in the 657 subjects against that of published data and peaks around each copy number integer.

3.3.7 Association analysis

We used linear regression in STATA version 13 to analyse evidence of association between copy number and BMI in the resulting 657 subjects. We analysed the mrCaNaVaR and Genome STRiP based copy number datasets separately. In both datasets, we used age and sex as covariates in the linear regression model.

We used Quanto version 1.2.4 to estimate power to detect association between *AMY1* copy number and BMI based on a published 0.66 to 4.40% genetic variance in European subjects.

We also analysed known associations with BMI and obesity from previous GWAS [7]. Falchi et al. reported that *AMY1* copy number had a larger effect on obesity than SNPs at *FTO*, the largest effect detected in GWAS. We used linear regression with STATA version 13 to detect an association between *FTO* SNP rs1558902 and BMI with age and sex as covariates. We also used the same linear regression to detect evidence of an association between a polygenic score composed of 11 SNPs known to associate with BMI (**Table 1**).

Table 1. Published GWAS SNPs used to calculate polygenic score for BMI

Spelotes et al	SNP	Effect Allele	Beta (SE)	EAF	VarExp	NPC
FTO	rs1558902	A	0.39 (0.02)	0.42	0.34%	14.03
MC4R	rs6567160	C	0.23 (0.03)	0.24	0.10%	4.149
SEC16B	rs543874	G	0.22 (0.03)	0.19	0.07%	2.905
FLJ35779	rs2112347	T	0.10 (0.02)	0.63	0.02%	0.821
BDNF	rs11030104	A	0.19 (0.03)	0.78	0.07%	2.855
MTCH2	rs3817334	T	0.06 (0.02)	0.41	0.01%	0.412
FAIM2	rs7138803	A	0.12 (0.02)	0.38	0.04%	1.65
MAP2K5	rs16951275	T	0.13 (0.02)	0.78	0.03%	1.228
GPRC5B	rs12446632	G	0.17 (0.03)	0.87	0.04%	1.628
SH2B1	rs3888190	A	0.15 (0.02)	0.4	0.05%	2.062
NEGR1	rs3101336	C	0.13 (0.02)	0.61	0.04%	1.643
				∑	0.81%	34.02

3.3.8 Copy number validation

We validated our copy number calling in 54 samples using droplet digital PCR (ddPCR). The ddPCR protocol provided by BioRad was followed, with the addition of samples being thermocycled for 10 more cycles (50 cycles total). Probe assays were selected for having a broad working range of annealing temperatures, producing good clustering of fluorescence types, covering regions not known to contain SNPs, and targeting areas not homologous to the other amylase genes. Copy number calls for all 54 InCHIANTI samples were obtained from a single ddPCR run. Given that *AMY1* and *AMY2A* share parity, we checked each individual's two replicate *AMY1* calls for concordance with their *AMY2A* call. If both *AMY1* calls were concordant, they were averaged. If only one was concordant, only the concordant *AMY1* genotype was used. If both calls were not concordant, they were averaged. This resulted in better clustering at integers (average deviation from integer 0.152, compared to straight averaging 0.179).

3.4 Results

3.4.1 *AMY1* copy number distribution in the InCHIANTI population

We noted that the Genome STRiP method produced a visually less noisy output than the MrsFAST/mrCaNaVaR method (**Fig. 1**). Observing the Genome STRiP data, we observed a distribution of *AMY1* CNVs consistent with most individuals carrying an odd number of *AMY1* copies that these sum to an even copy number in diploid genomes.

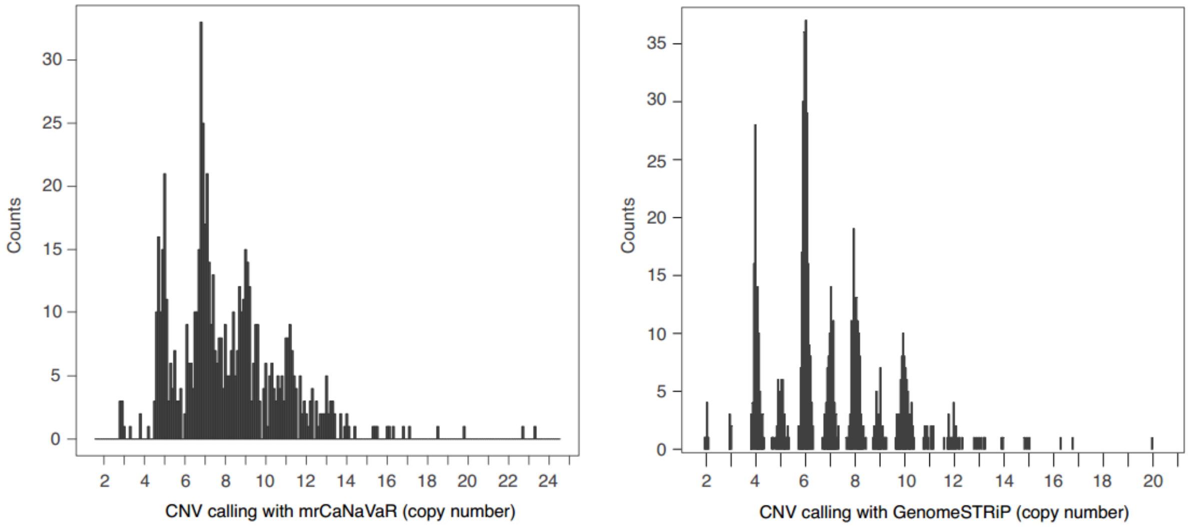


Figure 1. Comparison of *AMY1* copy number distributions using 657 InCHIANTI subjects. Both programs were run on the same, low pass whole genome sequencing dataset (InCHIANTI). Genome STRiP was run with BWA aligned data, and mrCaNaVaR was run with mrsFAST aligned data

3.4.2 Quality control of InCHIANTI copy number calls

Our *AMY1* CNVs called by the Genome STRiP method showed strong concordance with ddPCR based results in 54 InCHIANTI individuals (**Fig. 2**).

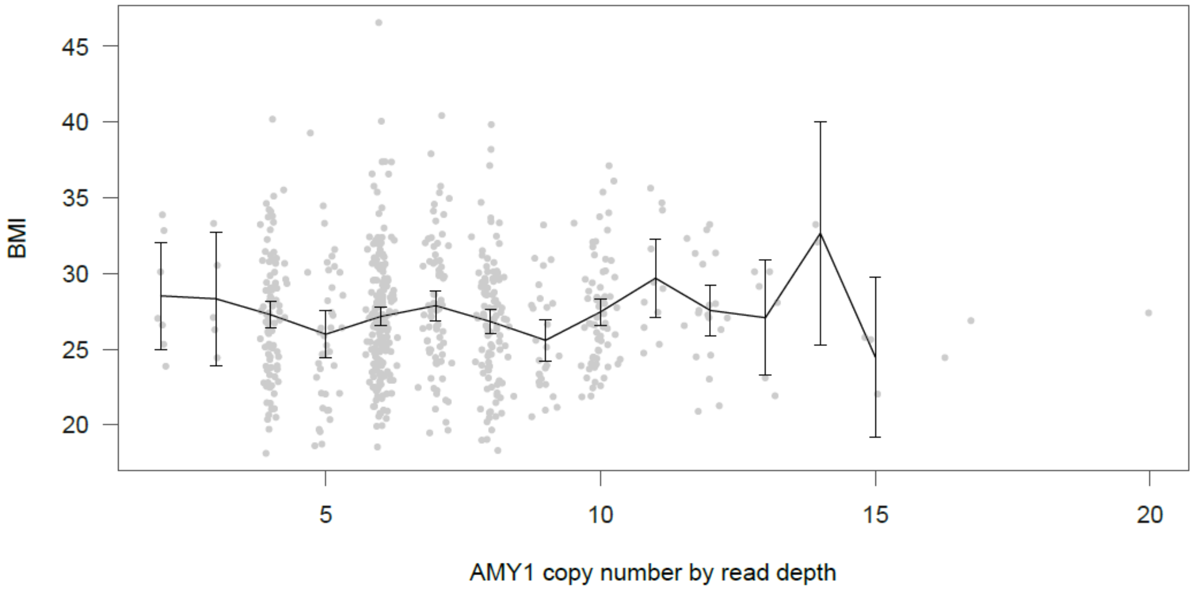


Figure 2. The association analysis of Genome STRiP derived *AMY1* copy number with BMI in the InCHIANTI cohort.

3.4.3 Calculating statistical power to detect association between *AMY1* copy number and BMI

At a significance level of *P* = 0.05, we estimated our statistical power thresholds for *AMY1* copy number to be associated with BMI to be: 77%, (95% CIs = 42% and 92%). Our power calculation is based on a point estimate on total BMI variance of 1.111% (95% CIs = 0.471% and 1.76%).

3.4.4 Association of *AMY1* copy number with BMI

We did not observe any association between *AMY1* copy number and BMI with either mrCaNaVaR (*P* = 0.64) or GenomeSTRiP called genotypes (*P* = 0.53) (Fig. 3, Table 2). Both the *FTO* (*P* = 0.074) variant and a polygenic score (*P* = 0.001) of 11 BMI SNPs were more strongly associated with BMI than *AMY1* copy number (Table 2).

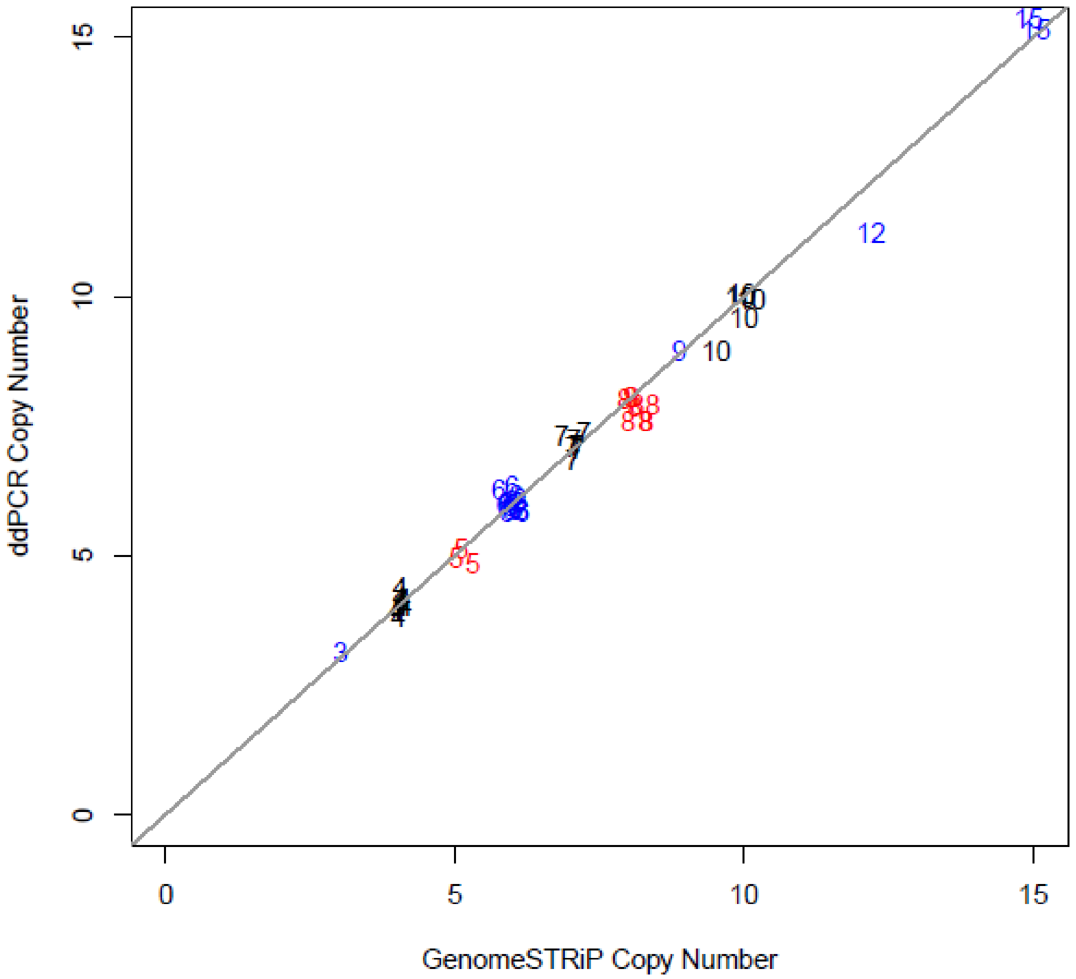


Figure 3. Concordance in copy number calling between Genome STRiP and ddPCR. The line plotted indicates full concordance.

Table 2. Association with BMI for *FTO* SNP, previously BMI-associated SNPs in multiple BMI loci and *AMY1* copy number

Gene	Variant	Power	Beta	*P* value
FTO	rs1558902	0.32	0.41 (-0.04-0.86)	0.074
Polygenic score	11 SNPs	0.64	0.50 (0.20-0.81)	0.001
AMY1	Copy Number: GenomeSTRiP	0.77	0.04 (-0.08-0.15)	0.526
	Copy Number: mrCaNaVaR		0.04 (-0.11-0.18)	0.635

3.5 Discussion

These results contrast with the recent findings that *AMY1* copy number exerts a far-stronger effect on BMI and obesity than SNPs at *FTO* and other loci [1]. In our study, *AMY1* copy number did not associate with BMI or obesity (**Fig. 2**, **Table 2**). Our findings have been replicated in 2 additional cohorts (1,000 Estonians and 2,863 individuals from the GoT2D cohort) and published in Nature Genetics [23]. We believe that the difference from the reported observation likely comes from our use of higher-resolution approaches for both molecular analysis and analysis of sequence data (**Fig. 3**).

We considered the possibility that our study could have failed to see a real effect. Our study is based in individuals from the Chianti region of Tuscany, a population that commonly consumes starches such as breads and pasta [24]. Our study was replicated in an Estonian study cohort, in addition to two European cohorts with elevated body weight. The Estonian diet is slightly different than that of other European countries [25], though it appears to be similarly rich in starch [26]. Fully understanding human genetic variation and its relationship to phenotypes will require characterising hundreds of complex loci like the amylase locus that mutate at high frequencies in ways that cause large-scale changes in the dosage and expression of genes. Some of these loci could be capable of rapid evolution [27-29]. The amylase locus offers insights to guide future studies of structurally complex loci. The high apparent complexity observed in measurements from diploid genomes may arise from a modest number of common structural forms that appear in different combinations in different diploid genomes. Structurally complex loci appear to share features with both common and rare variation reflecting both ancient and recent mutations. These loci may be best understood through combinations of tagging, imputation, and direct molecular measurement. Whole genome sequencing of large cohorts will ultimately reveal the extent to which this and many other structurally complex loci contribute to human phenotypic variation.

A recent publication found that a high number of copies of *AMY1* were associated with lower obesity in Mexican children [30]. They evaluated the number of *AMY1* copies in 597 Mexican children (293 obese children and 304 normal weight controls) using microfluidic digital PCR. The effect of *AMY1* copy number on obesity status was assessed using a logistic regression model adjusted for age and sex. They found that a high copy number of *AMY1* was associated with a reduced risk of obesity (OR per estimated copy 0.84, number of copies ranging from one to 16 in the population, $P = 4.25 \times 10^{-6}$). Although higher resolution methods than qPCR were used in this study, they did not correct for population stratification. Association studies in admixed populations such as Latinos are known to be confounded by population stratification [31]. Genetic stratification among indigenous populations within Mexico has been characterised and has been found to be highly complex [32]. It is critical to correct for population stratification before reporting an association in a sample from this population.

3.6 References

1. Falchi M (2014) Low copy number of the salivary amylase gene predisposes to obesity. Nat Genet 46: 492-497.

2. Groot PC (1989) The human [alpha]-amylase multigene family consists of haplotypes with variable numbers of genes. Genomics 5: 29-42.

3. Groot PC (1990) Evolution of the human [alpha]-amylase multigene family through unequal, homologous, and inter- and intrachromosomal crossovers. Genomics 8: 97-105.

4. Perry GH (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39: 1256-1260.

5. Groot PC, Mager WH, Frants RR (1991) Interpretation of polymorphic DNA patterns in the human [alpha]-amylase multigene family. Genomics 10: 779-785.

6. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889-894.

7. Locke AE (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518: 197-206.

8. Cantsilieris S, White SJ (2013) Correlating multiallelic copy number polymorphisms with disease susceptibility. Hum Mutat 34: 1-13.

9. Barnes C (2008) A robust statistical method for case-control association testing with copy number variation. Nat Genet 40: 1245-1252.

10. Clayton DG (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37: 1243-1246.

11. Zanda M (2014) A genome-wide assessment of the role of untagged copy number variants in type 1 diabetes. PLoS Genet 10: e1004367.

12. Melzer D (2008) A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet 4: e1000072.

13. Wood AR, Tuke MA, Nalls M, Hernandez D, Gibbs JR, et al. (2015) Whole-genome sequencing to understand the genetic architecture of common gene expression and biomarker phenotypes. Hum Mol Genet 24: 1504-1512.

14. Church DM (2011) Modernizing reference genome assemblies. PLoS Biol 9: e1001091.

15. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

16. Hach F (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Methods 7: 576-577.

17. Alkan C (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41: 1061-1067.

18. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.

19. Handsaker RE (2015) Large multiallelic copy number variations in humans. Nat Genet 47: 296-303.

20. Handsaker RE, Korn JM, Nemesh J, McCarroll SA (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43: 269-276.

21. Sudmant PH (2010) Diversity of human copy number variation and multicopy genes. Science 330: 641-646.

22. Quinlan AR (2014) BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics 47: 11 12 11-34.

23. Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, et al. (2015) Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. Nat Genet 47: 921-925.

24. Ferrucci L (2000) Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. J Am Geriatr Soc 48: 1618-1625.

25. Leitsalu L (2014) Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol.

26. Mottus R (2012) Personality traits and eating habits in a large sample of Estonians. Health Psychol 31: 806-814.

27. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.

28. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464: 713-720.

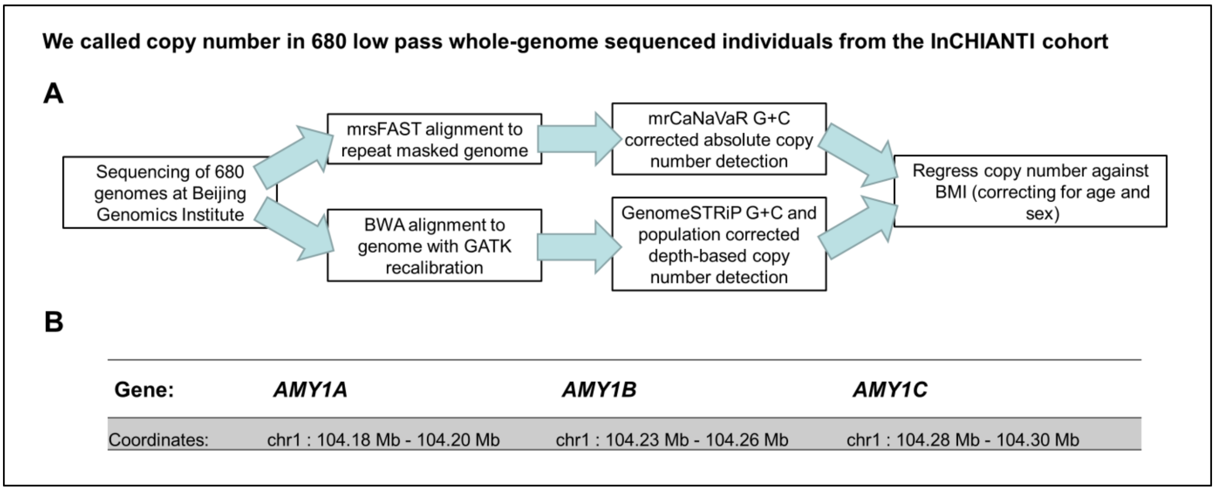
29. Tognon G (2014) Mediterranean diet, overweight and body composition in children from eight European countries: cross-sectional and prospective results from the IDEFICS study. Nutr Metab Cardiovasc Dis 24: 205-213.

30. Mejia-Benitez MA (2015) Beneficial effect of a high number of copies of salivary amylase AMY1 gene on obesity risk in Mexican children. Diabetologia 58: 290-294.

31. Choudhry S, Coyle NE, Tang H, Salari K, Lind D, et al. (2006) Population stratification confounds genetic association studies among Latinos. Hum Genet 118: 652-664.

32. Moreno-Estrada A, Gignoux CR, Fernandez-Lopez JC, Zakharia F, Sikora M, et al. (2014) Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science 344: 1280-1285.

3.7 Supplementary Information



We called copy number in 680 low pass whole-genome sequenced individuals from the InCHIANTI cohort

A
Sequencing of 680 genomes at Beijing Genomics Institute
mrsFAST alignment to repeat masked genome
BWA alignment to genome with GATK recalibration
mrCaNaVaR G+C corrected absolute copy number detection
GenomeSTRiP G+C and population corrected depth-based copy number detection
Regress copy number against BMI (correcting for age and sex)
B

Gene:	*AMY1A*	*AMY1B*	*AMY1C*
Coordinates:	chr1 : 104.18 Mb - 104.20 Mb	chr1 : 104.23 Mb - 104.26 Mb	chr1 : 104.28 Mb - 104.30 Mb

Figure S1. (A) Processing and analysis steps for low pass whole genome sequencing and copy number variant calling, **(B)** Triplicate read-mapped regions were taken from these three loci, and used to detect read counts and call copy number using both Genome STRiP and mrCaNaVaR

CHAPTER 4

Refining the CNV comorbidity map using 120,286 UK Biobank individuals

4.1 Abstract

Copy Number Variants (CNVs) in the form of large deletions and duplications have been associated with many neurodevelopmental disorders. These large copy number events are extremely rare and little is known about their prevalence in the general population. Furthermore, studies that have detected these associations have generally been subject to ascertainment bias. We scanned for 69 CNV sites that were previously implicated as being associated with developmental delay in SNP microarray data from 120,286 UK Biobank individuals. We aimed to further refine their prevalence in the general population and assess their relationship with developmental delay and common traits such as obesity. We found that the UK Biobank further strengthens the associations of these CNVs with developmental delay with the exception of four instances: one deletion and three duplications. Copy Number Variants found to be associated with developmental delay also have an impact on surrogate measures of developmental delay: Height, BMI, Fluid Intelligence, Income, and other complex traits. These findings underscore the fact that large, rare, pathogenic deletions and duplications are prevalent in control populations, but further work is needed to understand their full extent and effect on individuals that do not present with clinically detectable symptoms.

4.2 Introduction

The clinical interpretation of large deletions and duplications is difficult because they are very rare in the general population. A previous study by Coe et al of 30k children and 19k adult controls (a follow up to a study of 15k cases and 8k controls) identified and assessed the clinical significance and frequency of 69 regions of the genome harbouring deletions, duplications or both [1,2]. Cases were comprised of primarily paediatric cases with intellectual disability, developmental delay and/or Autism and controls were composed of 12 adult population cohorts of European origin. The study found 44 deletions and 28 duplications to be associated with developmental delay when comparing cases to controls. These data have been used as the basis for a “CNV morbidity map”. These large CNVs are greatly enriched in children with neurodevelopmental delay but these individuals usually present with a wide range of problems, some of which may be secondary to the disease process [3]. Many cases for example, present with obesity or growth defects, and a large proportion of CNVs are associated with obesity in the OMIM Morbid Map and DECIPHER clinical databases [4,5], but this may be a consequence of neurodevelopmental problems rather than a phenotype specific to the CNV and genes it has deleted or duplicated.

Here we aimed to refine the published CNV morbidity map by calling copy number variation in the previously described 69 regions associated with autosomal genomic disorders in 120,286 adult individuals from the UK Biobank – 6 times the size of the previously used largest control or population based study [6]. These individuals represent a control population, similar to that used by the previous study in that they are greatly depleted of individuals with neurodevelopmental delay because they had to actively sign up to the study. We hypothesised that the larger control sample size would reveal an additional prevalence of CNVs in healthy individuals, reducing or eliminating developmental delay significance in a proportion of CNVs. We also predicted that this reduction in significance would be more apparent in duplications as they exhibit a less obvious deleterious genetic impact compared to deletions. We detected deletions or duplications in 34 of the 69 published CNV regions, refined their frequency in the UK Biobank and used four common phenotypes available to assess their impact in the general population – height, BMI, a fluid intelligence test and income as surrogate measures of cognitive ability along with 22 additional phenotypic measures.

4.3 Materials and Methods

4.3.1 Selection of CNVs

We selected 69 regions of the genome that were previously identified as harbouring large rare pathogenic CNVs [1,2]. These CNVs were either deleted or duplicated loci at higher frequencies in cases of children with developmental delay.

4.3.2 CNV Detection

We initially identified CNVs present in 150,729 UK Biobank subjects from the first UK Biobank release (see http://biobank.ctsu.ox.ac.uk). We then refined this to 120,286 individuals of British descent. British-descent was defined as individuals who both self-identified as white British and were confirmed as ancestrally Caucasian using principal components analyses (http://biobank.ctsu.ox.ac.uk). Of these individuals, 120,286 were classified as unrelated, with a further 7,980 first- to third-degree relatives. We did not correct for relationships between individuals and therefore did not include related individuals in this study.

CNVs were called in the UK Biobank using PennCNV-Affy [7]. The PennCNV-Affy ‘validate’ function was used to detect subjects with a CNV present in the region. The CNV was then validated by confirming the presence of an intersecting CNV call using the PennCNV-Affy ‘test’ HMM-based algorithm. PennCNV-Affy was used to allow for a large number of samples to be processed in manageable amount of time and because of its low false discovery rate. A previous study in 12 monozygotic twins found that PennCNV had a lower mean CNV call discordance within twin pairs than Genotyping Console (GTC) used in the published study [8]. We developed a pipeline that was first used to transpose and quantile normalise probe intensity data suitable for the CNV detection pipeline, and then execute PennCNV processing. Details of the pipeline are given in section 4.3.3 below.

4.3.3 Probe intensity normalisation

Raw UK Biobank probeset intensities were quantile normalised to allow intensity data in each subject to be compared to one another [9]. During this process the data were also converted to the Birdseed format [10] required for further PennCNV-Affy processing.

A C++ program was developed to transpose and quantile normalise each sample according to chromosome and genotyping batch (**Appendix I**). Intensities were sorted numerically across each chromosome with missing values being allocated an overall median value to facilitate normalisation. Tied data items - including missing values - were broken to necessitate ranking by incrementing each respective tied data value by 1×10^{-5} (each raw data point had a precision of 6 with a maximum range of 5). Probe intensities across each chromosome within each sample were then added to an order statistic tree data structure to allow probe intensity rankings to be retrieved with $O(\log n)$ computational complexity. The mean intensity across each genotyping batch was then calculated for each sorted position. Each value in the original dataset was then queried against the order statistic tree and its resulting ranking was used as an index to extract from the list of sorted mean intensities. Mean intensities were then substituted in place of the equivalently ranked raw intensities whilst ignoring missing values. Each transformed intensity value was then $\log_2$ transformed for processing in PennCNV-Affy. Mean values of all intensities across each chromosome were checked to ensure that they were the same within each genotyping batch.

4.3.4 CNV Calling

PennCNV-Affy was used to infer genotype clusters and then generate combined probeset intensity values normalised to a reference dataset $\log_2(R_{obs}/R_{exp})$ where R is the combined probeset intensity known as Log R Ratio (LRR). PennCNV-Affy also generated B Allele Frequencies (BAF) representing the proportion of the combined probeset intensity accounted for by the B allele probe.

Two separate methods in PennCNV were then used to detect CNVs: ‘test’ and ‘validate’. PennCNV ‘test’ detects deviations from a normal LRR/BAF state across all probes in the genome using a Hidden Markov Model (HMM). The transition matrix and LRR/BAF distribution parameters for the HMM were based on previous studies using Affymetrix SNP6 array data. PennCNV ‘validate’ uses the 69 published CNV breakpoints and deletion/duplication frequencies as priors to assess the probability of a given copy number for that region in each subject without using the HMM.

4.3.5 CNV Scoring

LRR and BAF plots were generated for each CNV detected in each individual (**Appendix II**). Visual inspection was carried out by three independent reviewers. A CNV was not included if at least one reviewer concluded that the LRR and BAF distributions did now show sufficient evidence for have its called number of copies (**Table S7**). A CNV passed scoring overall if at least 80% of the called subjects passed visual inspection, and was also validated using PennCNV-Affy ‘test’.

4.3.6 Case – control significance testing

Fishers exact tests were used to calculate the statistical significance of the prevalence of each CNV in controls compared to cases with developmental delay. The R version 3.0.2 ‘fisher.test’ package was used to apply the Fisher’s exact test to 2 x 2 contingency matrices. We first used this method to replicate the published *P* values as a sanity check. Tests were then applied using UK Biobank CNV counts as control data.

4.3.7 Phenotype selection

Twenty six continuous phenotypes were selected as reciprocal indicators of morbidity and used in a phenotype wide scan “PheWAS” (**Table S4**). We identified 2 of the 26 common phenotypes as surrogate indicators of developmental delay to assess the impact in the general population –a fluid intelligence test and income. We also focused on the CNV associations with height and BMI, given that children with developmental delay often have stunted growth or obesity but it is not certain to what extent these features are secondary or primary to the genetic defect.

4.3.7.1 Height

Many disorders of developmental delay are associated with abnormal growth but this means that cohorts of cases with developmental delay could be biased towards individuals of abnormal growth because clinical geneticists may be more likely to refer an Individual for testing if accompanied by growth defects. Furthermore, recent studies have shown that common genetic variation from across a large fraction of the genome is involved in variation in height [11], suggesting that CNVs involving many genes could alter skeletal growth and final adult height.

4.3.7.2 BMI

Large rare CNVs have been associated with obesity, for example 16p11.2 [12,13]. BMI is also highly polygenic so many genes are likely to be involved [14]. BMI also has many of the same ascertainment bias issues as height.

4.3.7.3 Fluid Intelligence

Fluid Intelligence is a measure of an individual's ability to be adaptable and solve problems in a variety of situations. The fluid intelligence score is based on the number of correct logic and reasoning ability related questions answered in 2 minutes. A lower fluid intelligence score may then be an indicator of an individual with developmental delay.

4.3.7.4 Income

Income is a likely indicator of socioeconomic status and educational attainment throughout life. It can therefore be a secondary indicator of developmental delay.

4.3.8 Phenotype normalisation

We residualised and inverse normalised the 26 phenotypes to enable linear regression with minimal outliers and confounding. We then used Age, Sex, recruitment centre and principal components 1-5 to correct for population confounding during residualisation.

4.3.9 Phenotype association testing

Linear regression was carried out using STATA 13. We tested each of the 26 inverse normalised phenotypes against both deletions and duplications that were present in 10 or more UK Biobank individuals. We corrected for chip type (UKB Axiom or UK BiLEVE) at runtime.

4.4 Results

4.4.1 Frequency of deletions previously associated with known autosomal genomic disorders

We detected 937 deletion events (**Appendix II**) present in 25 of the 69 published loci (**Table S1**). We found that the published associations with developmental delay were generally strengthened (**Fig. 1A**), with exceptions detailed below.

4.4.1.1 Deletions present in cases, but not controls (Category A)

Of 34 deletions present in previous studies of cases that were not present in controls, we also found them to be absent from our 120,286 UK Biobank individuals. The number of probes present in each of these regions was comparable to that in the controls of Coe et al (**Table S2**) and zero of these regions intersected with previously reported problematic signature array regions prone to artefacts due to genomic waviness (as defined by Coe et al. 2014). The statistical confidence of these deletions being more frequent in cases than controls is now strengthened due to the larger sample size of the UK Biobank.

4.4.1.2 Deletions detected in published cases and not controls but present in UK Biobank (Category B)

Twelve deletions present in at least one case from the previous study but absent from their controls were detected in at least one individual in the UK Biobank (**Table 1**). Nevertheless, for 11 of these deletions, the statistical confidence of these being more frequent in cases than controls strengthened when using the UK Biobank as a control population because of its larger size. These 12 deletions included several known syndromes including 5 individuals with 22q DiGeorge deletion and one individual with Williams syndrome. The twelfth CNV that in the 3p11.2 deletion was present in 17 of 120,286 individuals, a frequency of 0.00014% which is not statistically different from the frequency of 0.00031% in the series of cases from Coe et al.

4.4.1.3 A deletion not present in published cases or controls but present in UK Biobank (Category C)

We found one deletion of 3Mb at 10q11 in 20 UK Biobank individuals that was not seen in Coe et al cases or controls (**Table 3**). This region was targeted because it was published as being duplicated in 10 cases, but no controls.

4.4.1.4 Deletions present in published cases and controls, and also UK Biobank (Category D)

For 12 additional regions, deletions were detected in the previous study in cases and controls, but at a higher frequency in cases, with *P*values ranging from 3.19 x 10^{-21} to 0.515. For 6 of these 12 deletions, using the UK Biobank as controls strengthened the statistical confidence that they were pathogenic. The exceptions were the chr17:14.07Mb-15.41Mb deletion including *HNPP* that was not previously different in frequency in published cases and controls, and remained so using UK Biobank; deletion chr13:20.81Mb-21.01Mb that we observed 106 times and a 290kb deletion of chr15:22.80Mb-23.09Mb, that was present in the UK Biobank at higher frequency

(0.38%) than in the previously reported control (0.14%) set but remained more frequent in cases (0.69%) with strong statistical confidence but weaker effect size (OR = 1.8).

4.4.1.5 Regions where there were either no deletions detected, or deletions did not pass quality control (Category E)

Two further regions failed our QC steps, and 8 regions were not detected as deleted in either cases or controls from the previous study or in UK Biobank. These regions were annotated because at least one published individual had a duplication.

Table 1. Ten deletions with 10 or more individuals present in UK Biobank, their significance and effect size on 4 surrogate developmental delay phenotypes, and their prevalence in 26 Phenotypes *: P=0.05 to 0.01, **: P=0.01-0.001 ***: P<0.001

Deletion	**Cat**	**Size**	**Coe case freq (n)**	**Coe control freq (n)**	**UKBB freq (n)**	**P Coe vs Coe**	**P coe vs UKBB**	**Intelligence, beta (95%CIs)**	**Income beta (95%CIs)**	**BMI beta (95%CIs) (kgm2)**	**Height beta (95%CIs) (cm)**	**PheWAS N<0.05**
chr15:31132708-32482708	B	1.35 Mb	65	0	10	2.85e-15	6.06e-36	-2.34** (-4.17, -0.51)	-0.82* (-1.63, 0.00)	2.91* (-0.06, 5.88)	-4.44* (-8.35, -0.52)	11/26
chr3:87237310-87557310	B	320kb	9	0	17	0.00972	0.0511	-	-	-	-5.19*** (-8.19, -2.19)	1/26
chr10:49389994-52389994	C	3 Mb	0	0	20	1	1	-	-	-	-	1/26
chr2:111383531-113093529	D	1.7mb	20	3	14	0.00483	4.88e-07	-1.63* (-3.18, -0.08)	-0.95** (-1.57, -0.33)	-	-	4/26
chr16:29652499-30202499	D	550kb	101	6	27	2.07e-16	1.74e-47	- (n=6)	-1.32*** (-1.80, -0.84)	7.39*** (5.58, 9.20)	-7.71*** (-10.09, -5.33)	15/26
chr16:15502499-16292499	D	790kb	36	7	41	0.000666	5.73e-08	- (n=11)	0.72*** (-1.10, -0.34)	-	-	4/26
chr17:14069275-15499275	D	1.43mb	13	8	65	0.515	0.775	-	-	-	-	1/26
chr16:21942499-22462499	D	520kb	50	11	74	0.000177	7.3e-08	-1.13* (-2.12, -0.13)	-0.50** (-0.79, -0.22)	-	-1.88* (-3.32, -0.44)	11/26
chr13:20812000-21012000	D	200kb	34	17	106	0.195	0.0937	-	-	-	-	1/26
chr15:22798636-23088559	D	290kb	200	27	462	3.19e-21	2.48e-11	1.10*** (-1.45, -0.75)	0.26*** (-0.37, -0.15)	-	-1.58*** (-2.16, -1.01)	12/26

4.4.2 Frequency of duplications previously associated with known autosomal genomic disorders

We detected a total of 1,231 duplication events (**Appendix II**) present in 24 of the 69 published loci (**Table S2**). We found that the published associations with developmental delay were generally strengthened (**Fig. 1B**), with exceptions detailed below.

4.4.2.1 Duplications present in cases, but not controls (Category A)

Of 25 duplications present in previous studies of cases that were not present in controls, we also found them to be absent from our 120,286 UK Biobank individuals. The greater sample size of UK Biobank resulted and the fact that we did not observe any of these duplications strengthened the statistical confidence that these duplications are pathogenic. Six duplications did not reach nominal significance, with only one copy of each occurring in cases. The number of probes present in each of these regions was comparable to that in the controls of the previously published controls and as with deletions, zero of these regions intersected with previously reported problematic regions [1].

4.4.2.2 Duplications detected in published cases and not controls but present in UK Biobank (Category B)

Ten duplications present in at least one case from the previously published study were absent from their controls but we detected at least one individual carrying the duplication in UK Biobank. Of these 10 duplications, the evidence for pathogenicity was strengthened for 6 whilst 4 remained unassociated or had very borderline evidence of pathogenicity. Two of these duplications were present in 10 or more individuals (**Table S2**). Three duplications were not present in previously published cases or controls but had one present in UK Biobank (**Table 3**). These regions were targeted because they were deleted in cases, but not controls in previous publication.

4.4.2.3 Duplications detected in published cases, only 1 control but not present in UK Biobank (Category C)

There were also 6 duplications present in many previously published cases, one of their controls, but none of the UK Biobank. For five of these, the greater sample size of UK Biobank and the fact that we did not observe any of these duplications strengthened the statistical confidence that these duplications are pathogenic (**Table S2**). The final duplication was only present in one case and one previously published control.

4.4.2.4 Duplications detected in published cases and controls and also present in UK Biobank (Category D)

For 10 additional regions, duplications were detected in the previous publication in cases and controls, but at a higher frequency in cases, with p-values ranging from 1.35×10^{-11} to 0.31. For seven of these, we observed stronger statistical confidence of pathogenicity. Two duplications at 16p13.11 and chr16:15.5Mb-16.3Mb were no longer of nominal significance with UK Biobank controls. One duplication was not different between cases and controls whether using previous controls or UK Biobank.

Table 2. Ten duplications with 10 or more individuals present in UK Biobank, their significance and effect size on the 4 phenotypes, and prevalence in 26 Phenotypes *: P=0.05 to 0.01, **: P=0.01-0.001 ***: P<0.001

Duplication	Cat	Size	Coe case freq (n)	Coe control freq (n)	UKBB freq (n)	P Coe vs Coe	P coe vs UKBB	Intelligence, beta (95%CIs)	Income beta (95%CIs)	BMI beta (95%CIs) (kgm2)	Height beta (95%CIs) (cm)	PheWAS N<0.05
chr2:96726273-97676273	B	950kb	4	0	12	0.128	0.381	-	-	-2.44* (-5.15, 0.27)	-	3/26
chr2:111383531-113093529	B	1.71 Mb	7	0	26	0.0272	0.469	-	-	-1.91* (-3.75, -0.07)	-2.85* (-5.28, -0.42)	6/26
chr17:34815887-36205887	D	1.39 Mb	23	3	32	0.00147	0.00012	-1.87* (-3.42, -0.33)	-	-	-	3/26
chr16:21942499-22462499	D	520kb	11	4	61	0.212	0.854	-	-	-1.10* (-2.31, 0.12)	-	3/26
chr17:14069275-15499275	D	1.43 Mb	17	5	35	0.0691	0.0168	-	-0.44* (-0.89, 0.00)	-	-	7/26
chr16:29652499-30202499	D	550kb	62	9	37	3.5e-07	7.01e-21	-2.09** (-3.39, -0.80)	-0.49* (-0.90, -0.09)	-1.62* (-3.16, -0.07)	2.64** (0.61, 4.68)	13/26
chr15:31132708-32482708	D	1.35 Mb	28	11	81	0.0834	0.0679	-0.68* (-0.95, -0.41)	-0.94*** (-1.74, -0.14)	-	-	9/26
chr22:19020000-20290000	D	1.27 Mb	97	12	75	1.35e-11	1.05e-26	-1.32** (-2.29, -.36)	-0.55*** (-0.82, -0.28)	1.97** (0.88, 3.06)	-1.47 (-2.89, -0.04)	12/26
chr16:15502499-16292499	D	790Kb	68	27	248	0.0112	0.197	-0.71** (-1.17, -0.25)	-0.22** (-0.37, -0.07)	0.76* (0.16, 1.35)	-	13/26
chr15:22798636-23088559	D	290Kb	128	60	586	0.0112	0.863	-	-	-	-	3/26

4.4.2.5 Duplications not present in published cases or controls but present in UK Biobank (Category E)

Four separate deletions, each present in one individual and each greater than 1Mb were not seen in Coe et al cases or controls. These regions were targeted because they were published as being deleted in cases and controls, but not duplicated (**Table 3**).

4.4.2.6 Regions where there were either no duplications detected, or duplications did not pass quality control (Category F)

Fourteen regions were not detected as duplicated in either cases or controls from the previous study or in UK Biobank. These regions were annotated because at least one individual from the previous published study had a deletion.

Table 3. Summary of deletions and duplications detected at N ≥ 1 in UK Biobank that are not present in previously published cases and controls

Chromosome	Size (Mb)	Syndrome	UKBB N	Type
chr10:49389994-52389994	3	10q11 duplication syndrome (Coe et al)	20	Deletion
chr1:168733376-173733377	5	1q24 deletion syndrome	1	Duplication
chr10:81690020-88940020	7.25	10q23.1 deletion syndrome	1	Duplication
chr3:115237310-115647310	4.1	3q13 deletion syndrome	1	Duplication
chr2:50146496-51256496	1.1	-	1	Duplication

4.4.3 Summary of variants tested

Twenty seven deletion and 24 duplication loci were detected in the 69 published morbidity map regions [1]. Ten deletions (**Table 1**) and 10 duplications (**Table 2**) that were each present in greater than or equal to ten individuals were selected for phenotype association testing.

4.4.4 Height, BMI, Fluid Intelligence and Income association

Seven deletions and 9 duplications present in 10 or more individuals were associated at $p<0.05$ in at least one of 4 of our main phenotypes (fluid intelligence, income, height and BMI). Deletions had a generally stronger association, with 4 deletions at $P \leq 0.0001$ in at least one trait compared to 2 duplications.

4.4.5 Association of deletions and duplications with all 26 phenotypes

After testing each of the ten deletions against the 26 phenotypes, there were 61 associations at $p < 0.05$ (**Fig. 2A, Table 1, Table S1**). After testing each of the ten duplications against the 26 phenotypes, there were 61 associations at $p<0.05$ (**Fig. 2B, Table 2, Table S2**).

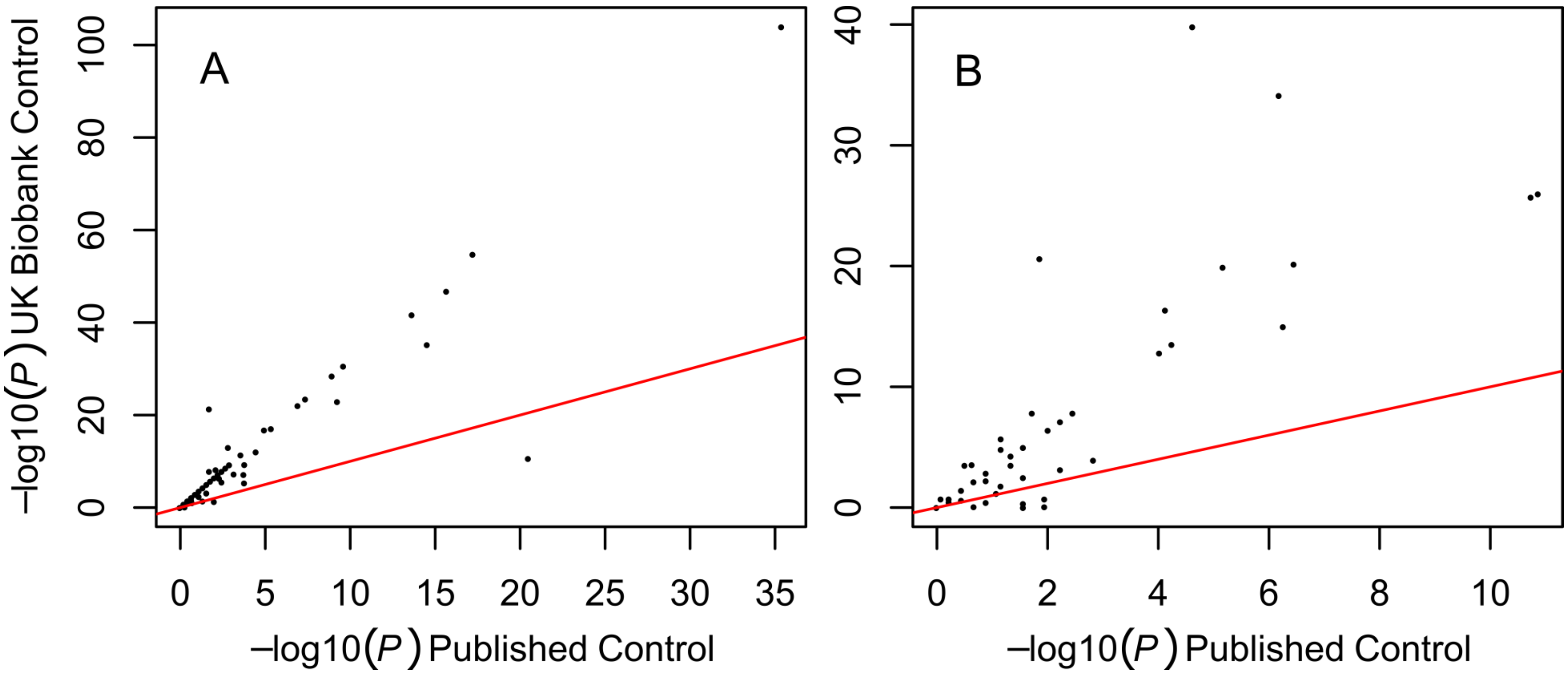


Figure 1. Quantile-quantile plots comparing the evidence of pathogenicity for CNVs when using case frequencies from Coe et al, but using frequencies from Coe et al controls or UK Biobank controls. **(A)** and **(B)** show deletions and duplications respectively.

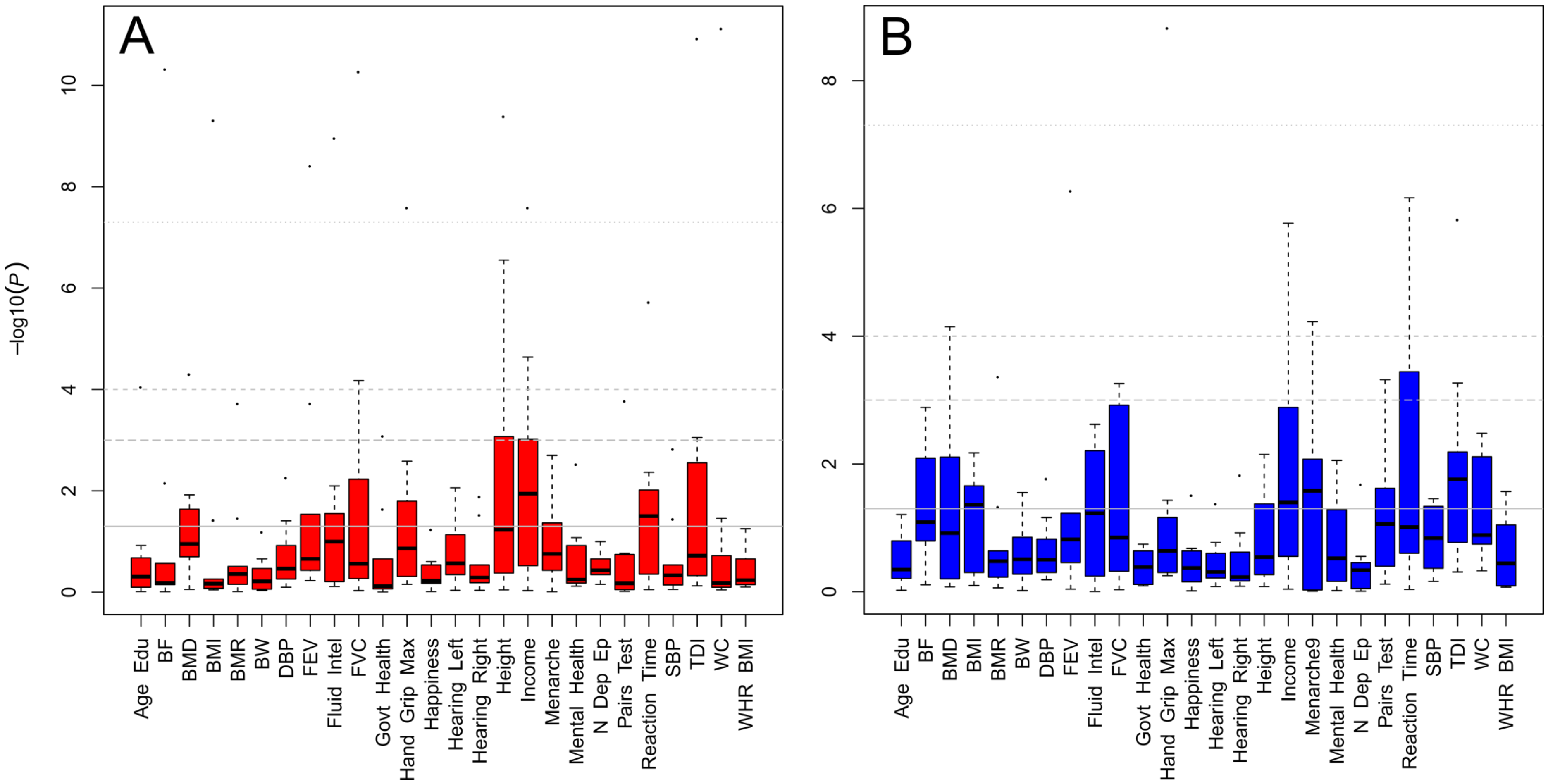


Figure 2. Plots showing 26 phenotypes and the distribution of their phenotype associations with **(A)** ten deletions and **(B)** ten duplications where there are 10 or more carriers in UK BIobank. Full phenotype names are provided in **Table S4**

4.5 Discussion

4.5.1 UK Biobank controls increase the strength of evidence for pathogenicity of the developmental delay CNVs with four exceptions

The UK Biobank data strengthen the evidence that most CNVs previously reported by Coe et al. are pathogenic, causing developmental delay. We hypothesised that the significance of CNVs could have been overestimated due to the SNP arrays used in controls varying in technology and cohort size, thus lowering the prevalence, however for most CNVs this is not the case. We also hypothesised that significance of duplications would be lower as an increased gene dosage would have less deleterious biological mechanisms to cause developmental delay which also appears not to be apparent for most duplications. The 4 exceptions to these findings are given in the next four subsections. Notably, 3 of the 4 exceptions described below are duplications, and the single deletion was novel in the previously published study indicating more potentially false associations with developmental delay among duplications.

4.5.1.1 The 320Kb deletion at 3p11.2 is not associated with developmental delay

This deletion at chr3:87,237,310-87,557,310 was published as a ‘newly detected CNV’ associated with developmental delay in that the CNV was not called in previous studies. The deletion was detected in nine individuals with developmental delay. The significance of these newly detected CNVs are based on a median of the gene-level copy number empirical *P* values [1] in an attempt to account for a somewhat complex breakpoint structure in these CNVs. Their simulated *P* value was 7.5 x 10^{-5} and a *q* value of 0.049. We found using the standard fisher's exact test a *P* value of 0.0097. Using UK Biobank as a control altered this *P* value to 0.051 eliminating its nominal association. The CNV was detected in UK Biobank data with 79 supporting probes, greater than the number of probes present in 5 of the 8 array types used in the published data. This variant is associated with decreased height, but does not show any evidence of being associated with decreased cognitive function however due to the sub-optimal probe count, further studies of this region need to be carried out to fully understand the impact of this region.

4.5.1.2 The 1.7Mb duplication at 2q13 is no longer associated with developmental delay

The duplication at chr2:111,383,531-113,093,529 is no longer nominally associated with developmental delay. The variant was previously published with a p-value of 0.027 (q-value 0.0928 whereas the published q-value cut-off is between 0.1 and 0.25). The variant is well captured with 410 supporting probes present in 26 individuals. The duplication is associated with both a lower BMI and height and present at $p<0.05$ in 6 of the 26 phenotypes at tested which include decreased lung function and basal metabolic rate. The duplication is also associated with an increased reaction time, which may also be indicative of cognitive function. These association statistics indicate that this duplication has a potential phenotypic impact, but it may not necessarily be one of decreased cognitive function and/or developmental delay.

4.5.1.3 The 290Kb duplication at 15q11.2 is not associated with developmental delay

The duplication at chr15:22,798,636-23,088,559 present in 586 individuals was published as having a p-value of 0.0112 (q-value 0.0513). A p-value of 0.863 was reported when using UK Biobank as controls. This variant in UK Biobank is supported by 137 probes, greater than the number available in six of the eight arrays used in the published study. The variant is not significant in any of the surrogate measures of developmental delay, but is associated at $p<0.05$ with three of the 26 tested phenotypes including increased social deprivation a higher score in the pairs matching test and an earlier age of menarche in females.

4.5.1.4 The 790Kb duplication at 16p13.11 is not associated with developmental delay

This variant at chr16:15,502,499-16,292,499 is detected in 248 individuals. The published study found this duplication to be associated with developmental delay with a *P* value of 0.0112 and a q value of 0.0513. Using UK Biobank controls the *P* value became 0.197 losing its nominal significance. This variant is supported by 339 probes which is higher than the number on 7 of the 8 arrays used in the previous publication. Paradoxically, this variant is associated with a decreased fluid intelligence score, a decreased income and a 0.76kgm^2 increase in BMI. The variant is associated with 13 of the 26 tested traits overall including decreased lung function, earlier menopause and a nominal association with an increased pairs matching score.

4.5.2 CNVs found to be associated with developmental delay have a stronger impact on Height, BMI, Fluid Intelligence, Income and other common traits

We found that CNVs that associated with developmental delay were more likely to be associated with many common traits in our "PheWAS" scan of 26 common continuous traits (we did not test disease or other binary traits because there would be too few numbers). In the 10 deletions occurring in ≥ 10 individuals, those that associated with developmental delay also associate with ≥ 1 surrogate developmental delay phenotype and are nominally significant in between 4 and 15 of the 26 phenotypes tested. The 4 deletions that are not significantly associated with developmental delay were each only nominally significant in exactly 1 of the 26 phenotypes tested which included increased reaction time, decreased hand grip strength, decreased height and increased bone mineral density. Duplications occurring in ≥ 10 individuals were all associated with between 3 and 13 of the 26 phenotypes including ≥ 1 surrogate developmental delay phenotype with the exception of one duplication (chr15:22798636-23088559) present in 586 individuals that was not associated with developmental delay but was associated with early menarche, higher pairs matching test score and high social deprivation. These results implicate a higher level of pleiotropy in deletions associated with developmental delay whereas duplications have a high level of pleiotropy regardless of association with developmental delay.

4.5.3 Potential biases and limitations of this study

There are several limitations in this study, with perhaps the most notable being the detection technology used. Even though the UK Biobank array has been custom built with these pathogenic CNVs in mind, some of the loci have less than 100 probes. When observing probe intensity plots (**Appendix II**) it is apparent that the variance introduced by signal noise means that many probes are needed to rely on a CNV call. There is also a level of uncertainty that we are picking up exactly the same CNV calls that we would if we used the aCGH signature arrays used in the published cases. This is however a problem in the previous study that is compounded by their use of multiple technologies and different technologies in the cases compared to controls.

The UK Biobank is biased against individuals with developmental delay as the participants are aged between 40 and 69 and a recruitment questionnaire was required to sign up to the study. Socioeconomic status within the UK Biobank provides a minimum bound for developmental delay given that the UK Biobank is likely to be biased towards higher functioning and better supported Individuals with intellectual disability. The bias against developmental delay in the UK Biobank therefore underscores its suitability as a control group in this study.

4.6 References

1. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, et al. (2014) Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet 46: 1063-1071.

2. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, et al. (2011) A copy number variation morbidity map of developmental delay. Nat Genet 43: 838-846.

3. O'Donovan MC, Kirov G, Owen MJ (2008) Phenotypic variations on the theme of CNVs. Nat Genet 40: 1392-1393.

4. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res 43: D789-798.

5. Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, et al. (2014) DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation. Nucleic Acids Res 42: D993-D1000.

6. Collins R (2012) What makes UK Biobank special? Lancet 379: 1173-1174.

7. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665-1674.

8. Castellani CA, Melka MG, Wishart AE, Locke ME, Awamleh Z, et al. (2014) Biological relevance of CNV calling methods using familial relatedness including monozygotic twins. BMC Bioinformatics 15: 114.

9. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.

10. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40: 1253-1260.

11. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46: 1173-1186.

12. Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, et al. (2010) A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature 463: 671-675.

13. Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, et al. (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature 478: 97-102.

14. Locke AE (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518: 197-206.

4.7 Supplementary Information

Table S1. Summary of all 69 CNV regions and their deletions prevalence and significance. HS = Hotspot CNVs flanked by segmental duplications, MB = multi-breakpoint CNVs. HSMB = situations where a combination of both hotspot and multi-breakpoint CNVs have been detected.

Chr	Start (HG19)	End (HG19)	Size (bp)	Type	Category	Deletion Syndrome	Del Case N	Del Control Coe N	Del UKBB N	Del Coe *P*	Del UKBB *P*	Control vs Control *P*
chr1	10,001	10,077,413	10,067,412	MB	A	1p36 deletion syndrome	77	0	0	5.84E-18	1.77E-55	1
chr1	168,733,376	173,733,377	5,000,001	MB	A	1q24 (FMO deletions andDNM3)	12	0	0	0.00207	2.97E-09	1
chr10	81,690,020	88,940,020	7,250,000	HSMB	A	10q23.1 (SFTPD to GLUD1NRG3 inclusive)	11	0	0	0.00347	1.52E-08	1
chr10	81,960,020	88,800,020	6,840,000	HS	A	10q23 deletion	11	0	0	0.00347	1.52E-08	1
chr11	43,983,424	46,063,424	2,080,000	MB	A	Potocki-Shaffer syndrome	6	0	0	0.0455	5.45E-05	1
chr11	67,753,424	71,282,352	3,528,928	HS	A	SHANK2 FGFs deletion	1	0	0	0.598	0.195	1
chr15	24,818,907	28,426,405	3,607,498	HS	A	Prader-Willi/Angelman	40	0	0	1.13E-09	3.69E-29	1
chr15	72,962,947	75,532,947	2,570,000	HS	A	15q24 A to C deletion	7	0	0	0.0272	1.06E-05	1
chr15	72,962,947	76,012,945	3,049,998	HS	A	15q24 A to D deletion	2	0	0	0.357	3.79E-02	1
chr15	74,012,947	75,532,947	1,520,000	HS	A	15q24 Covers B to C deletion	13	0	0	0.00124	5.77E-10	1
chr15	74,012,947	75,532,947	1,520,000	HS	A	15q24 Covers B to C deletion	13	0	0	0.00124	5.77E-10	1
chr15	74,012,947	76,012,945	1,999,998	HS	A	15q24 B to D deletion	2	0	0	0.357	0.0379	1
chr15	74,012,947	78,132,945	4,119,998	HS	A	15q24 B to E deletion	2	0	0	0.357	0.0379	1
chr15	83,182,945	84,738,996	1,556,051	HS	A	15q25.2 deletion	1	0	0	0.598	0.195	1
chr16	3,779,999	3,859,999	80,000	MB	A	Rubinstein-Taybi syndrome	4	0	0	0.128	1.44E-03	1
chr16	21,352,499	29,442,499	8,090,000	HS	A	Shaffer locus deletion	3	0	0	0.213	0.00738	1
chr16	21,612,499	29,042,499	7,430,000	HS	A	16p11.2-p12.2	3	0	0	0.213	0.00738	1

						microdeletion syndrome						
chr17	50,000	2,593,250	2,543,250	MB	A	17p13.3 deletion (both YWHAE and PAFAH1B1)	16	0	0	0.000264	4.26E-12	1
chr17	2,363,250	2,923,250	560,000	MB	A	17p13.3 deletion (including PAFAH1B1)	11	0	0	3.47E-03	1.52E-08	1
chr17	16,709,275	20,479,408	3,770,133	HSMB	A	Smith-Magenis syndrome deletion	24	0	0	4.29E-06	8.76E-18	1
chr17	43,704,217	44,184,217	480,000	HS	A	17q21.31 deletion	31	0	0	1.16E-07	9.24E-23	1
chr17	58,065,218	60,305,218	2,240,000	HS	A	17q23.1q23.2 deletion	1	0	0	5.98E-01	1.95E-01	1
chr2	200,161,755	200,511,755	350,000	MB	A	2q33.1 (SATB2)	13	0	0	0.00124	5.77E-10	1
chr2	239,705,243	242,471,327	2,766,084	MB	A	2q37 deletion	33	0	0	4.16E-08	3.50E-24	1
chr22	51,113,134	51,173,134	60,000	MB	A	Phelan-McDermid syndrome deletion	43	0	0	2.40E-10	2.71E-31	1
chr3	115,237,310	115,647,310	410,000	MB	A	3q13 (GAP43)	9	0	0	0.00972	4.02E-07	1
chr4	1,870,202	2,010,202	140,000	MB	A	Wolf-Hirschhorn deletion	24	0	0	4.29E-06	8.76E-18	1
chr4	80,780,976	83,280,976	2,500,000	MB	A	4q21 (BMP3)	11	0	0	0.00347	1.52E-08	1
chr5	10,000	11,727,000	11,717,000	MB	A	Cri du Chat syndrome	4	0	0	0.128	0.00144	1
chr5	87,964,244	88,224,244	260,000	MB	A	5q14 (MEF2C)	10	0	0	0.00581	7.83E-08	1
chr5	175,717,394	177,057,394	1,340,000	HS	A	Sotos syndrome deletion	10	0	0	5.81E-03	7.83E-08	1
chr6	100,813,279	100,943,279	130,000	MB	A	6q16 deletion	1	0	0	0.598	1.95E-01	1
chr8	8,092,590	11,892,591	3,800,001	HS	A	8p23.1 deletion	8	0	0	0.0163	2.06E-06	1
chr9	137,860,179	141,080,179	3,220,000	MB	A	9q34 deletion	5	0	0	7.62E-02	2.80E-04	1
chr12	65,073,733	68,643,733	3,570,000	MB	B	12q14 microdeletion syndrome	3	0	1	0.213	0.0252	1
chr15	31,132,708	32,482,708	1,350,000	HS	B	15q13.3 deletion	65	0	10	2.85E-15	6.06E-36	1
chr15	85,138,996	85,698,996	560,000	HS	B	Cooper 15q25.2	7	0	3	0.0272	0.000728	1
chr17	553,250	1,353,250	800,000	MB	B	17p13.3 deletion (including YWHAE)	17	0	3	1.58E-04	5.14E-10	1
chr17	29,165,874	30,215,887	1,050,013	HS	B	NF1 microdeletion syndrome	7	0	3	0.0272	7.28E-04	1
chr2	96,726,273	97,676,273	950,000	HS	B	2q11.2 deletion	6	0	8	0.0455	3.89E-02	1
chr22	19,020,000	20,290,000	1,270,000	HS	B	DiGeorge/VCFS deletion	158	0	5	3.97E-36	1.17E-104	1

chr22	21,910,000	23,650,000	1,740,000	HS	B	22q11.2 distal deletion	20	0	2	3.37E-05	9.36E-13	1
chr3	87,237,310	87,557,310	320,000	MB	B	3p11.2 (CHMP2B to POU1F1)	9	0	17	9.72E-03	5.11E-02	1
chr3	195,715,603	197,355,603	1,640,000	HS	B	3q29 deletion	11	0	3	0.00347	3.08E-06	1
chr7	72,742,064	74,142,064	1,400,000	HS	B	Williams syndrome deletion	61	0	1	2.24E-14	2.15E-42	1
chr7	74,962,064	76,662,064	1,700,000	HS	B	Wms-distal deletion	5	0	2	0.0762	0.00413	1
chr10	49,389,994	52,389,994	3,000,000	HS MB	C		0	0	20	1.00E+00	1.00E+00	1
chr1	145,288,643	145,628,643	340,000	HS	D	TAR deletion	25	2	28	0.000163	4.82E-06	0.936279175
chr1	146,573,376	147,393,376	820,000	HS	D	1q21.1 deletion	68	6	38	5.50E-10	1.24E-23	0.593602025
chr13	20,812,000	21,012,000	200,000	MB	D	13q12 deletion	34	17	106	0.195	0.0937	0.562410047
chr15	22,798,636	23,088,559	289,923	HS	D	15q11.2 deletion	200	27	462	3.19E-21	2.48E-11	1
chr16	15,502,499	16,292,499	790,000	HS	D	16p13.11 deletion	36	7	41	0.000666	5.73E-08	0.516340936
chr16	21,942,499	22,462,499	520,000	HS	D	16p12.1 deletion	50	11	74	0.000177	7.30E-08	0.657882728
chr16	28,772,499	29,112,499	340,000	HS	D	16p11.2 distal deletion	27	1	2	1.09E-05	1.73E-17	0.36398112
chr16	29,652,499	30,202,499	550,000	HS	D	16p11.2 deletion	101	6	27	2.07E-16	1.74E-47	0.312278953
chr17	14,069,275	15,499,275	1,430,000	HS	D	HNPP	13	8	65	0.515	0.775	0.819327942
chr17	34,815,887	36,205,887	1,390,000	HS	D	17q12 deletion (ACACA)	20	2	1	0.00145	1.04E-13	0.053321504
chr2	111,383,531	113,093,529	1,709,998	HS	D	2q13 deletion	20	3	14	0.00483	4.88E-07	0.432493057
chr3	191,517,306	193,017,306	1,500,000	MB	D	3q28-29 (FGF12)	13	1	1	0.00772	6.62E-09	0.260427914
chr12	6,469,739	6,809,739	340,000	MB	E		0	0	0	1.00E+00	1.00E+00	1
chr15	99,362,477	102,521,392	3,158,915	MB	E	15q26 deletion	11	1	0	0.0188	1.52E-08	0.140015729
chr17	57,655,218	58,075,218	420,000	HS	E	17q23 deletion	0	0	0	1	1.00E+00	1
chr2	50,146,496	51,256,496	1,110,000	MB	E	2p16.1 (NRXN1)	30	9	0	0.019	4.75E-22	2.06E-08
chr2	57,746,496	61,736,496	3,990,000	HS	E	2p15-16.1 microdeletion syndrome	0	0	0	1	1.00E+00	1
chr2	59,646,496	63,146,496	3,500,000	MB	E		0	0	0	1	1	1
chr2	100,693,568	108,443,568	7,750,000	HS	E	2q11.2q13 deletion	0	0	0	1.00E+00	1.00E+00	1
chr3	9,525,000	11,025,000	1,500,000	MB	E		0	0	0	1	1	1
chr7	66,482,565	72,272,064	5,789,499	HS	E	Wms-prox deletion	0	0	0	1	1	1
chr9	32,010,000	39,010,000	7,000,000	MB	E		0	0	0	1.00E+00	1.00E+00	1

Table S2. Summary of all 69 CNV regions and their duplication prevalence and significance. HS = Hotspot CNVs flanked by segmental duplications, MB = multi-breakpoint CNVs. HSMB = situations where a combination of both hotspot and multi-breakpoint CNVs have been detected.

Chr	**Start (HG19)**	**End (HG19)**	**Size (bp)**	**Type**	**Category**	**Duplication Syndrome**	**Dup Case N**	**Dup Control Coe**	**Dup UKBB N**	**Dup Coe *P***	**Dup UKBB *P***	**Control vs Control *P***
chr11	43,983,424	46,063,424	2,080,000	MB	A	None	2	0	0	0.357	0.0379	1
chr15	72,962,947	75,532,947	2,570,000	HS	A	None	3	0	0	0.213	0.00738	1
chr15	72,962,947	76,012,945	3,049,998	HS	A	None	3	0	0	0.213	7.38E-03	1
chr15	74,012,947	75,532,947	1,520,000	HS	A	None	7	0	0	2.72E-02	1.00E+00	1
chr15	74,012,947	75,532,947	1,520,000	HS	A	None	7	0	0	0.0272	1.00E+00	1
chr15	74,012,947	76,012,945	1,999,998	HS	A	None	1	0	0	0.598	0.195	1
chr15	99,362,477	102,521,392	3,158,915	MB	A	15q26 overgrowth syndrome	4	0	0	1.28E-01	1.44E-03	1
chr16	21,352,499	29,442,499	8,090,000	HS	A	None	2	0	0	0.357	0.0379	1
chr16	21,612,499	29,042,499	7,430,000	HS	A	None	2	0	0	0.357	3.79E-02	1
chr17	50,000	2,593,250	2,543,250	MB	A	17p13.3 duplication (both YWHAE and PAFAH1B1)	6	0	0	0.0455	5.45E-05	1
chr17	16,709,275	20,479,408	3,770,133	HSMB	A	Potocki-Lupski syndrome duplication	19	0	0	5.63E-05	3.14E-14	1
chr17	29,165,874	30,215,887	1,050,013	HS	A	None	7	0	0	0.0272	1.06E-05	1
chr17	43,704,217	44,184,217	480,000	HS	A	17q21.31 duplication	3	0	0	0.213	0.00738	1
chr2	59,646,496	63,146,496	3,500,000	MB	A	2p15-16.1 proximal (PEX13 to AHSA2)	9	0	0	0.00972	4.02E-07	1

chr2	100,693,568	108,443,568	7,750,000	HS	A	None	1	0	0	0.598	1.95E-01	1
chr2	239,705,243	242,471,327	2,766,084	MB	A	None	1	0	0	0.598	0.195	1
chr22	51,113,134	51,173,134	60,000	MB	A	None	11	0	0	0.00347	1.52E-08	1
chr3	9,525,000	11,025,000	1,500,000	MB	A	3p25.3 (JAGN1 to TATDN2)	10	0	0	0.00581	7.83E-08	1
chr4	1,870,202	2,010,202	140,000	MB	A	None	11	0	0	0.00347	1.52E-08	1
chr5	10,000	11,727,000	11,717,000	MB	A	None	1	0	0	0.598	0.195	1
chr5	175,717,394	177,057,394	1,340,000	HS	A	None	3	0	0	2.13E-01	7.38E-03	1
chr7	66,482,565	72,272,064	5,789,499	HS	A	Wms-prox duplication	1	0	0	0.598	1.95E-01	1
chr7	74,962,064	76,662,064	1,700,000	HS	A	Wms-distal duplication	1	0	0	5.98E-01	1.95E-01	1
chr9	32,010,000	39,010,000	7,000,000	MB	A	9p13	18	0	0	9.43E-05	1.61E-13	1
chr9	137,860,179	141,080,179	3,220,000	MB	A	9q34 duplication	6	0	0	0.0455	5.45E-05	1
chr10	49,389,994	52,389,994	3,000,000	HS MB	B	10q11	10	0	8	0.00581	0.000729	1
chr10	81,960,020	88,800,020	6,840,000	HS	B	None	4	0	1	0.128	6.07E-03	1
chr15	24,818,907	28,426,405	3,607,498	HS	B	PWS duplication	48	0	8	1.82E-11	1.98E-26	1
chr15	83,182,945	84,738,996	1,556,051	HS	B	None	1	0	1	0.598	0.352	1
chr15	85,138,996	85,698,996	560,000	HS	B	None	2	0	3	0.357	2.52E-01	1
chr2	96,726,273	97,676,273	950,000	HS	B	2q11.2 duplication	4	0	12	0.128	0.381	1
chr2	111,383,531	113,093,529	1,709,998	HS	B	2q13 duplication	7	0	26	0.0272	4.69E-01	1
chr22	21,910,000	23,650,000	1,740,000	HS	B	22q11.2 distal duplication	7	0	5	0.0272	0.00332	1
chr7	72,742,064	74,142,064	1,400,000	HS	B	WBS duplication	28	0	5	5.46E-07	1.05E-15	1
chr8	8,092,590	11,892,591	3,800,001	HS	B	8p23.1 duplication	6	0	1	0.0455	3.18E-04	1
chr1	10,001	10,077,413	10,067,412	MB	C	None	28	1	0	6.70E-06	1.26E-20	0.14001573
chr12	6,469,739	6,809,739	340,000	MB	C	12p13 (SCNN1A to PIANP)	23	1	0	7.37E-05	4.50E-17	0.14001573

chr13	20,812,000	21,012,000	200,000	MB	C	None	5	1	0	0.23	0.00028	0.14001573
chr16	3,779,999	3,859,999	80,000	MB	C	None	8	1	0	0.0686	2.06E-06	0.14001573
chr17	553,250	1,353,250	800,000	MB	C	17p13.3 duplication (including YWHAE	11	1	0	0.0188	1.52E-08	0.14001573
chr6	100,813,279	100,943,279	130,000	MB	C	None	1	1	0	0.838	0.195	0.14001573
chr15	22,798,636	23,088,559	289,923	HS	D	None	128	60	586	0.0112	0.863	0.99989281
chr15	31,132,708	32,482,708	1,350,000	HS	D	15q13.3 duplication	28	11	81	0.0834	0.0679	0.75731026
chr16	15,502,499	16,292,499	790,000	HS	D	16p13.11 duplication	68	27	248	0.0112	0.197	0.98505229
chr16	21,942,499	22,462,499	520,000	HS	D	None	11	4	61	0.212	0.854	0.98592899
chr16	29,652,499	30,202,499	550,000	HS	D	16p11.2 duplication	62	9	37	3.50E-07	7.01E-21	0.18698307
chr17	2,363,250	2,923,250	560,000	MB	D	17p13.3 duplication (including PAFAH1B1)	8	1	1	0.0686	1.54E-05	0.26042791
chr17	14,069,275	15,499,275	1,430,000	HS	D	CMT1A	17	5	35	0.0691	0.0168	0.67636756
chr17	34,815,887	36,205,887	1,390,000	HS	D	17q12 duplication (ACACA)	23	3	32	0.00147	0.00012	0.8855369
chr22	19,020,000	20,290,000	1,270,000	HS	D	22q11.2 duplication	97	12	75	1.35E-11	1.05E-26	0.56918152
chr3	195,715,603	197,355,603	1,640,000	HS	D	3q29 duplication	6	2	1	3.10E-01	3.18E-04	0.0533215
chr1	168,733,376	173,733,377	5,000,001	MB	E		0	0	1	1	1	1
chr10	81,690,020	88,940,020	7,250,000	HSMB	E		0	0	1	1	1	1
chr2	50,146,496	51,256,496	1,110,000	MB	E		0	0	1	1	1	1
chr3	115,237,310	115,647,310	410,000	MB	E		0	0	1	1	1	1
chr1	145,288,643	145,628,643	340,000	HS	F	None	56	11	0	2.37E-05	1.54E-40	4.04E-10
chr1	146,573,376	147,393,376	820,000	HS	F	1q21.1 duplication	48	5	0	6.50E-07	7.55E-35	5.38E-05
chr11	67,753,424	71,282,352	3,528,928	HS	F	None	0	0	0	1	1	1
chr12	65,073,733	68,643,733	3,570,000	MB	F	None	0	0	0	1	1	1

chr15	74,012,947	78,132,945	4,119,998	HS	F	None	0	0	0	1	1.00E+00	1
chr16	28,772,499	29,112,499	340,000	HS	F	None	29	8	0	0.0137	2.44E-21	1.48E-07
chr17	57,655,218	58,075,218	420,000	HS	F	None	0	0	0	1	1	1
chr17	58,065,218	60,305,218	2,240,000	HS	F	None	0	0	0	1.00E+00	1.00E+00	1
chr2	57,746,496	61,736,496	3,990,000	HS	F	None	0	0	0	1.00E+00	1.00E+00	1
chr2	200,161,755	200,511,755	350,000	MB	F		0	0	0	1	1	1
chr3	87,237,310	87,557,310	320,000	MB	F		0	0	0	1	1	1
chr3	191,517,306	193,017,306	1,500,000	MB	F		0	0	0	1.00E+00	1.00E+00	1
chr4	80,780,976	83,280,976	2,500,000	MB	F		0	0	0	1.00E+00	1.00E+00	1
chr5	87,964,244	88,224,244	260,000	MB	F		0	0	0	1.00E+00	1.00E+00	1

Table S3. The 69 previously published regions and the number of genotyping probes present within each genotyping array. Affymetrix Axiom is the array used by UK Biobank with the remaining 8 arrays used in the previous study

CNV (HG19)	Affy Axiom UK Biobank	Affy 6 SNP	Affy 6 CN	Illumina hh240k	Illumina hh317k	Illumina hh550k	Illumina hh610k Quad	Illumina hh650k	Human 1.2m Duo Custom
chr1:10001-10077413	3,578	2,734	2,494	849	1,093	1,685	2,224	2,154	4,511
chr1:145288643-145628643	160	45	79	12	16	23	50	36	161
chr1:146573376-147393376	263	300	275	116	84	167	212	229	425
chr1:168733376-173733377	1,279	1,820	1,706	419	535	820	1,049	1,092	2,550
chr2:50146496-51256496	222	460	419	138	176	287	315	356	469
chr2:57746496-61736496	840	1,257	1,321	287	462	664	858	889	1,467
chr2:59646496-63146496	745	956	1,118	196	387	528	663	665	1,303
chr2:96726273-97676273	280	107	281	17	66	74	96	90	406
chr2:100693568-108443568	1,800	2,723	2,589	673	860	1,353	1,756	1,808	3,268
chr2:111383531-113093529	410	390	486	111	151	229	321	291	633
chr2:200161755-200511755	39	89	102	17	18	24	34	41	94
chr2:239705243-242471327	1,320	896	783	293	396	585	791	753	1,349
chr3:9525000-11025000	661	579	444	177	222	367	420	439	835
chr3:87237310-87557310	79	94	105	26	32	54	62	66	99
chr3:115237310-115647310	156	148	131	35	63	90	97	116	175
chr3:191517306-193017306	431	438	641	74	177	222	278	279	606
chr3:195715603-197355603	591	354	503	58	209	240	299	288	741
chr4:1870202-2010202	60	10	41	5	5	7	12	10	61
chr4:80780976-83280976	544	737	777	170	213	321	421	476	740
chr5:10000-11727000	4,396	5,848	4,050	1,549	1,769	2,905	3,463	3,863	6,086
chr5:87964244-88224244	44	49	86	18	24	40	45	48	78
chr5:175717394-177057394	538	255	393	45	124	150	196	197	738
chr6:100813279-100943279	66	46	47	16	16	29	32	37	54
chr7:66482565-72272064	1,305	1,829	1,897	434	585	902	1,108	1,190	1,799
chr7:72742064-74142064	336	143	366	38	76	98	156	125	596
chr7:74962064-76662064	380	236	486	51	93	115	232	158	563
chr8:8092590-11892591	1,563	2,118	1,224	592	652	1,098	1,229	1,444	2,062
chr9:32010000-39010000	2,045	2,262	2,257	539	848	1,202	1,512	1,598	2,942
chr9:137860179-141080179	1,670	683	751	262	326	497	682	637	1,813
chr10:49389994-52389994	752	1,043	809	284	317	519	648	703	1,147

chr10:81690020-88940020	1,695	2,785	2,498	741	842	1,396	1,672	1,881	3,109
chr10:81960020-88800020	1,568	2,675	2,390	716	801	1,339	1,581	1,802	2,849
chr11:43983424-46063424	712	924	627	272	302	520	589	687	976
chr11:67753424-71282352	1,136	1,070	993	299	398	589	739	799	1,530
chr12:6469739-6809739	159	73	89	21	38	51	69	58	218
chr12:65073733-68643733	848	1,365	1,140	366	406	663	788	939	1,563
chr13:20812000-21012000	94	117	62	48	42	78	87	98	121
chr15:22798636-23088559	137	84	144	56	50	94	112	107	155
chr15:24818907-28426405	1,223	1,410	1,325	497	469	854	999	1,123	1,731
chr15:31132708-32482708	424	471	538	123	155	240	298	310	490
chr15:72962947-76012945	870	587	966	161	252	367	510	470	1,289
chr15:72962947-75532947	729	560	855	153	236	345	430	443	1,090
chr15:74012947-75532947	510	317	509	104	147	224	277	285	799
chr15:74012947-76012945	651	344	620	112	163	246	357	312	998
chr15:74012947-78132945	1,045	778	1,283	218	316	465	665	610	1,632
chr15:74012947-75532947	510	317	509	104	147	224	277	285	799
chr15:83182945-84738996	343	332	588	95	160	230	268	280	579
chr15:85138996-85698996	231	151	183	70	71	129	173	146	368
chr15:99362477-102521392	1,335	1,304	1,053	496	402	786	1,010	1,013	1,636
chr16:3779999-3859999	43	29	27	7	9	12	16	22	33
chr16:15502499-16292499	339	250	275	85	103	163	261	211	698
chr16:21352499-29442499	2,418	2,277	2,593	609	804	1,247	1,671	1,600	3,231
chr16:21612499-29042499	2,352	2,216	2,463	589	791	1,228	1,583	1,567	3,141
chr16:21942499-22462499	168	60	198	17	31	38	51	55	175
chr16:28772499-29112499	100	22	60	7	14	17	61	21	134
chr16:29652499-30202499	174	56	136	16	32	39	54	50	303
chr17:50000-2593250	1,202	543	784	207	340	488	653	571	1,333
chr17:553250-1353250	461	184	213	85	133	193	265	228	408
chr17:2363250-2923250	191	137	151	59	76	115	144	144	272
chr17:14069275-15499275	528	679	501	238	247	424	485	545	767
chr17:16709275-20479408	978	653	1,190	168	237	331	621	463	1,375
chr17:29165874-30215887	365	204	317	50	92	124	169	166	449
chr17:34815887-36205887	548	417	566	135	132	236	269	295	683
chr17:43704217-44184217	401	181	174	31	43	61	74	81	178
chr17:57655218-58075218	130	47	128	14	18	26	38	34	186
chr17:58065218-60305218	446	362	781	109	131	190	323	292	831

chr22:19020000-20290000	508	390	453	90	225	271	328	337	847
chr22:21910000-23650000	488	542	824	408	198	415	536	552	1,058
chr22:51113134-51173134	51	8	32	3	7	8	11	9	26
Total	52,664	53,570	54,899	15,080	19,055	29,761	37,775	38,969	75,831

Table S4: 26 selected phenotypes. The four italicised phenotypes are the surrogate measures of developmental delay

Trait
Height
Income
Body Mass Index
Fluid Intelligence Test Results
Age Full-time education completed
Body fat mass
Bone Mineral Density
Basal Metabolic Rate
Birthweight
Systolic Blood Pressure
Diastolic Blood Pressure
Forced Expiratory Volume
Forced vital capacity
Waist-hip-ratio – BMI prediction
Government health advice not followed
Max hand grip
Happiness
Hearing Left
Hearing Right
Menarche Age
Mental Health
N Depression Episodes
Pairs matching test results
Reaction Time Test Results
Townsend Deprivation Index
Waist Circumference

Table S5: All nominal association signals between deletions with ≥ 10 subjects and 26 phenotypes. Full phenotype names can be found in **Table S4**

Deletion	**Phenotype**	**Normalised Beta**	***P***	**95% Conf. Interval**	**95% Conf. Interval**
chr16:29652499-30202499	WC	1.314842	7.60E-12	0.9384411	1.6912
chr16:29652499-30202499	TDI	1.302269	1.20E-11	0.9259142	1.6786
chr16:29652499-30202499	BF	1.340512	4.80E-11	0.9410151	1.74
chr16:29652499-30202499	FVC	-1.492716	5.40E-11	-1.938696	-1.0467
chr16:29652499-30202499	Height	-1.202022	4.10E-10	-1.578819	-0.82522
chr16:29652499-30202499	BMI	1.195622	4.90E-10	0.8189919	1.5722
chr15:22798636-23088559	Fluid Intelligence	-0.5247701	1.10E-09	-0.6936044	-0.35593
chr16:29652499-30202499	FEV1	-1.332914	3.90E-09	-1.776406	-0.88942
chr16:29652499-30202499	Hand Grip	-1.091212	2.60E-08	-1.47556	-0.70686
chr16:29652499-30202499	Income	-1.241913	2.60E-08	-1.679243	-0.80458
chr15:22798636-23088559	Height	-0.2391234	2.80E-07	-0.330382	-0.14786
chr15:22798636-23088559	Reaction Time	0.2226718	1.90E-06	0.1310275	0.31431
chr15:22798636-23088559	Income	-0.2117539	2.30E-05	-0.3098496	-0.11365
chr15:22798636-23088559	BMD	0.1910984	5.00E-05	0.0987592	0.28343
chr16:21942499-22462499	FVC	-0.4870333	6.70E-05	-0.7263944	-0.24767
chr15:22798636-23088559	Age Edu	-0.2032467	9.00E-05	-0.3049999	-0.10149
chr15:22798636-23088559	Pairs Test	0.1749629	1.70E-04	0.0836138	0.26631
chr15:22798636-23088559	BMR	-0.1764364	1.90E-04	-0.269095	-0.08377
chr16:21942499-22462499	FEV1	-0.4527552	1.90E-04	-0.6907745	-0.21473
chr16:29652499-30202499	Gov Health	0.8609771	8.30E-04	0.3563209	1.3656
chr3:87237310-87557310	Height	-0.8079058	8.50E-04	-1.28279	-0.33302
chr15:31132708-32482708	TDI	1.048135	8.90E-04	0.4296853	1.6665
chr16:15502499-16292499	Income	-0.5819057	9.70E-04	-0.9276954	-0.23611
chr16:21942499-22462499	Income	-0.4219683	1.40E-03	-0.6810861	-0.16285
chr16:21942499-22462499	SBP	0.3683188	1.50E-03	0.1405262	0.59611
chr16:29652499-30202499	Menarche	-0.9780877	2.00E-03	-1.597822	-0.35835
chr16:21942499-22462499	Hand Grip	-0.3521346	2.60E-03	-0.5815745	-0.12269
chr15:22798636-23088559	TDI	0.1392559	2.80E-03	0.047898	0.23061
chr15:22798636-23088559	Mental Health	0.1463013	3.00E-03	0.0497224	0.24288
chr2:111383531-113093529	Income	-0.8223152	4.30E-03	-1.386939	-0.25769
chr16:21942499-22462499	Reaction Time	0.3410362	4.30E-03	0.1067333	0.57533
chr16:21942499-22462499	DBP	0.3251406	5.50E-03	0.0957543	0.55452
chr15:31132708-32482708	FVC	-0.9098203	5.90E-03	-1.557865	-0.26177
chr15:31132708-32482708	Menarche	1.11643	6.20E-03	0.3164063	1.9164
chr16:21942499-22462499	BF	0.3175139	7.00E-03	0.0867841	0.54824

chr15:31132708-32482708	Fluid Intelligence	-1.186379	8.00E-03	-2.062518	-0.31023
chr16:29652499-30202499	Hearing Left	1.07057	8.70E-03	0.2705644	1.8705
chr10:49389994-52389994	Reaction Time	0.5792585	9.60E-03	0.1410057	1.0175
chr2:111383531-113093529	BMD	-0.6985362	1.20E-02	-1.242029	-0.15504
chr16:29652499-30202499	Hearing Right	1.018385	1.30E-02	0.2184294	1.818
chr17:14069275-15499275	Hand Grip	-0.2998442	1.60E-02	-0.5448808	-0.0548
chr16:21942499-22462499	Height	-0.2754811	1.80E-02	-0.5031535	-0.0478
chr16:29652499-30202499	Reaction Time	0.4737537	1.80E-02	0.0817516	0.86575
chr15:31132708-32482708	Reaction Time	0.747439	1.80E-02	0.1276787	1.3671
chr2:111383531-113093529	TDI	0.6112118	2.20E-02	0.0885046	1.1339
chr13:20812000-21012000	BMD	0.2235422	2.30E-02	0.0303839	0.4167
chr15:22798636-23088559	Gov Health	-0.1259984	2.30E-02	-0.2346025	-0.01739
chr16:21942499-22462499	Fluid Intelligence	-0.5316567	2.80E-02	-1.006903	-0.05641
chr15:22798636-23088559	Hand Grip	-0.1027744	2.80E-02	-0.1943266	-0.01122
chr15:31132708-32482708	FEV1	-0.7200331	2.90E-02	-1.364448	-0.07561
chr15:31132708-32482708	Hearing Right	0.9676511	3.00E-02	0.0913444	1.8439
chr15:31132708-32482708	Income	-0.8197365	3.00E-02	-1.558992	-0.08048
chr15:31132708-32482708	Height	-0.6817934	3.10E-02	-1.300968	-0.06261
chr16:21942499-22462499	BMR	-0.2490802	3.50E-02	-0.480107	-0.01805
chr2:111383531-113093529	Fluid Intelligence	-0.7961965	3.50E-02	-1.536728	-0.05566
chr15:31132708-32482708	WC	0.6659698	3.50E-02	0.0474249	1.2845
chr16:15502499-16292499	SBP	-0.3266635	3.60E-02	-0.6326625	-0.02066
chr15:31132708-32482708	BMI	0.6539486	3.80E-02	0.0350476	1.272
chr16:15502499-16292499	DBP	-0.3225518	3.90E-02	-0.6286011	-0.0165
chr16:15502499-16292499	Menarche	0.4218677	4.30E-02	0.0131739	0.83056

Table S6: All nominal association signals between duplications with ≥ 10 subjects and 26 phenotypes. Full phenotype names can be found in **Table S4**

Duplication	**Phenotype**	**Normalised Beta**	***P***	**95% Conf. Interval**	**95% Conf. Interval**
chr17:14069275-15499275	Hand Grip	-1.035867	1.50E-09	-1.371969	-0.69976
chr16:15502499-16292499	FEV1	-0.3314618	5.30E-07	-0.4610408	-0.20188
chr22:19020000-20290000	Reaction Time	0.581565	6.80E-07	0.3521403	0.81098
chr15:31132708-32482708	TDI	0.5338413	1.50E-06	0.316488	0.75119
chr15:31132708-32482708	Income	-0.5977845	1.70E-06	-0.8423144	-0.35325
chr22:19020000-20290000	Income	-0.5090041	5.90E-05	-0.7574527	-0.26055
chr16:15502499-16292499	Menarche	-0.3552882	5.90E-05	-0.5286583	-0.18191
chr16:29652499-30202499	BMD	-0.6533272	7.10E-05	-0.9755044	-0.331
chr16:29652499-30202499	Reaction Time	0.6157008	1.80E-04	0.2934739	0.93792
chr15:31132708-32482708	BMD	-0.4022634	3.20E-04	-0.6214065	-0.18312
chr2:111383531-113093529	Reaction Time	0.6998756	3.60E-04	0.3155041	1.0842
chr15:22798636-23088559	Menarche	-0.2089679	3.80E-04	-0.3243045	-0.09363
chr2:111383531-113093529	BMR	-0.7045439	4.30E-04	-1.096518	-0.31257
chr22:19020000-20290000	Pairs Test	0.4030418	4.80E-04	0.176685	0.62939
chr22:19020000-20290000	TDI	0.3986387	5.40E-04	0.1727537	0.62452
chr2:111383531-113093529	FVC	-0.7148715	5.50E-04	-1.120269	-0.30947
chr16:15502499-16292499	FVC	-0.2207261	9.00E-04	-0.3510454	-0.0904
chr22:19020000-20290000	FVC	-0.3928109	1.20E-03	-0.6303855	-0.15523
chr16:15502499-16292499	BF	0.2060179	1.30E-03	0.0803321	0.3317
chr16:15502499-16292499	Income	-0.2235328	1.30E-03	-0.3596054	-0.08746
chr16:29652499-30202499	Fluid Intelligence	-0.9611624	2.40E-03	-1.580716	-0.3416
chr16:15502499-16292499	Fluid Intelligence	-0.3312719	3.10E-03	-0.5505222	-0.11202
chr22:19020000-20290000	WC	0.3385275	3.30E-03	0.1126106	0.56444
chr22:19020000-20290000	Fluid Intelligence	-0.6444623	6.20E-03	-1.106305	-0.18261
chr16:15502499-16292499	TDI	0.1724727	6.50E-03	0.0481577	0.29678
chr16:15502499-16292499	WC	0.1727811	6.60E-03	0.0481992	0.29736
chr22:19020000-20290000	BMI	0.3149141	6.70E-03	0.0873446	0.54248
chr16:29652499-30202499	Height	0.4420509	7.10E-03	0.120121	0.76398
chr16:21942499-22462499	BF	-0.3494294	7.70E-03	-0.6064873	-0.09237
chr2:111383531-	WC	-0.5212194	7.70E-03	-0.9048443	-0.13759

113093529					
chr17:14069275-15499275	BMD	-0.4705459	7.80E-03	-0.8169804	-0.12411
chr22:19020000-20290000	BF	0.3094642	8.10E-03	0.0803194	0.5386
chr17:14069275-15499275	TDI	0.445126	8.30E-03	0.1145118	0.77574
chr16:29652499-30202499	Menarche	0.6042885	8.40E-03	0.1546583	1.0539
chr16:29652499-30202499	Mental Health	0.4557851	8.80E-03	0.1147116	0.79685
chr15:31132708-32482708	Menarche	-0.4672409	9.30E-03	-0.8192764	-0.1152
chr15:22798636-23088559	Pairs Test	0.1075248	9.40E-03	0.0263705	0.18867
chr16:15502499-16292499	BMD	0.1611572	1.10E-02	0.0363456	0.28596
chr2:96726273-97676273	Menarche	1.144153	1.10E-02	0.2677756	2.02
chr17:34815887-36205887	Fluid Intelligence	-0.924841	1.40E-02	-1.665358	-0.18432
chr17:34815887-36205887	Hearing Right	0.8635691	1.50E-02	0.170747	1.5563
chr15:22798636-23088559	TDI	0.10064	1.50E-02	0.0195857	0.18169
chr2:111383531-113093529	BMI	-0.4733959	1.60E-02	-0.8572435	-0.08954
chr2:111383531-113093529	Height	-0.4740238	1.60E-02	-0.8580413	-0.09
chr17:14069275-15499275	DBP	-0.4040618	1.70E-02	-0.7352916	-0.0728
chr17:34815887-36205887	Income	-0.4976708	1.70E-02	-0.9055277	-0.08981
chr2:96726273-97676273	TDI	-0.6706185	2.00E-02	-1.235199	-0.10603
chr16:29652499-30202499	N Dep Ep	1.032854	2.10E-02	0.1566896	1.909
chr2:96726273-97676273	BMI	-0.6624781	2.20E-02	-1.227456	-0.0975
chr15:31132708-32482708	Pairs Test	0.2508288	2.40E-02	0.033005	0.46865
chr15:31132708-32482708	Fluid Intelligence	-0.4386381	2.50E-02	-0.8229652	-0.05431
chr16:15502499-16292499	BMI	0.1401646	2.70E-02	0.0157605	0.26456
chr15:31132708-32482708	WHR-BMI	-0.2455209	2.70E-02	-0.4633482	-0.02769
chr16:15502499-16292499	BW	-0.1857166	2.80E-02	-0.3509261	-0.0205
chr16:29652499-30202499	TDI	0.3579006	2.90E-02	0.0363358	0.67946
chr16:29652499-30202499	Happiness	0.6836079	3.10E-02	0.0639338	1.3032
chr16:29652499-30202499	Income	-0.3981735	3.50E-02	-0.7678401	-0.0285
chr16:29652499-30202499	SBP	-0.3470937	3.50E-02	-0.6692032	-0.02498
chr22:19020000-20290000	Hand Grip Max	-0.2411037	3.70E-02	-0.4674696	-0.01473
chr16:21942499-22462499	Reaction Time	0.2634418	4.00E-02	0.0124536	0.51442
chr16:21942499-22462499	BMI	-0.2633772	4.10E-02	-0.5160948	-0.01065
chr16:15502499-16292499	Pairs Test	0.1298384	4.10E-02	0.005262	0.25441

chr16:29652499-30202499	BF	-0.3381259	4.20E-02	-0.6643861	-0.01186
chr15:31132708-32482708	Hearing Left	0.4069673	4.20E-02	0.0149358	0.79899
chr22:19020000-20290000	Height	-0.2351322	4.20E-02	-0.4612833	-0.0089
chr16:15502499-16292499	SBP	0.1284589	4.40E-02	0.0036805	0.25323
chr16:29652499-30202499	BMI	-0.3271512	4.60E-02	-0.6489432	-0.00535
chr17:14069275-15499275	Income	-0.4160416	4.60E-02	-0.8239055	-0.00817
chr15:31132708-32482708	SBP	0.2216817	4.60E-02	0.0039431	0.43942
chr17:14069275-15499275	BMR	-0.3567136	4.70E-02	-0.7087388	-0.00468
chr17:14069275-15499275	SBP	-0.3344085	4.80E-02	-0.6655865	-0.00323
chr22:19020000-20290000	Mental Health	0.2340828	4.90E-02	0.0015136	0.4666

Table S7. Results of CNV visual inspection

Locus	**Deletion score**	**Dup score**
chr1:145288643-145628643	Second round – bad	Second round - bad
chr1:146573376-147393376	Good	Second round - bad
chr2:96726273-97676273	Second round – good	good
chr2:111383531-113093529	Good	good
chr3:87237310-87557310	Good	bad
chr3:191517306-193017306	Good	Second round - bad
chr3:195715603-197355603	Good	good
chr4:1870202-2010202	Bad	bad
chr5:175717394-177057394	Bad	bad
chr6:100813279-100943279	Bad	bad
chr7:72742064-74142064	Good	good
chr7:74962064-76662064	Good	good
chr8:8092590-11892591	Bad	good
chr9:137860179-141080179	Bad	bad
chr10:49389994-52389994	Good	good
chr10:81690020-88940020	Good	good
chr10:81960020-88800020	Good	good
chr12:6469739-6809739	Bad	bad
chr13:20812000-21012000	Good	bad
chr15:22798636-23088559	Good	good
chr15:31132708-32482708	Good	good
chr15:85138996-85698996	Second round – good	good
chr16:3779999-3859999	Good	bad
chr16:15502499-16292499	Good	good
chr16:21942499-22462499	Good	good
chr16:28772499-29112499	Second round – good	Second round - bad
chr16:29652499-30202499	Second round – good	good
chr17:50000-2593250	Bad	bad
chr17:553250-1353250	Second round – good	Second round - bad
chr17:2363250-2923250	Bad	good
chr17:14069275-15499275	Good	good
chr17:16709275-20479408	Bad	bad
chr17:29165874-30215887	Second round – good	Second round - good
chr17:34815887-36205887	Good	good
chr17:43704217-44184217	Bad	bad
chr17:57655218-58075218	Bad	bad
chr22:19020000-20290000	Good	good
chr22:21910000-23650000	Good	good
chr22:51113134-51173134	Bad	bad

CHAPTER 5

A Genome-wide association study of large genetic deletions and duplications in 120,286 individuals

5.1 Abstract

Since the first Genome-wide Association Studies (GWAS), Copy Number Variation (CNV) has given researchers an early glimpse into the polygenicity of rare genetic variation. In the pre next-generation sequencing era, genome-wide association studies were limited to being able to detect common SNPs (>5% frequency in the population) but CNVs could be detected at singleton resolution. The sample size of GWAS has been limited however, and many of the CNVs that have been implicated in complex traits and disease risk have been too rare to study We detected CNVs genome-wide in a sample of 120,286 individuals from the UK Biobank where we had power to detect these rare events, and used novel approaches to test their association with 204 complex traits. We found two deletions and two duplications out of a total of 2,258 detected CNVs to be both genome-wide and nominally associated with many traits including novel signals associated with sedentary lifestyle phenotypes such as television watching. These results suggest that there are copy number variants associated with many different complex traits in the general population. More work is now needed to classify these variants more accurately in order to detect more novel associations, and even larger sample sizes will allow the penetrance of these CNVs to be more fully understood with regards to complex traits and common disease risk.

5.2 Introduction

Large duplications and deletions have been implicated as causal to severe developmental delay, mental disorders, obesity, and other congenital disorders [1-4]. These so called pathogenic duplications and deletions are highly pleiotropic with variable penetrance [2] and have also been detected in subjects without severe phenotypes. It is also now known that there are many putative benign CNVs in the population [5] and it has become a challenge within the field to pinpoint which CNVs are pathogenic. Large CNVs can delete or duplicate many genes in one single event, but it is not clear how they act upon the gene to alter the phenotype. There are several potential mechanisms that are thought to mediate the effect large deletions and duplications on phenotypes, for example: copy number correlating with gene dosage effects and the combination of an encompassing deletion and loss-of-function variant deleting both copies of a gene [6,7]. It has also been hypothesised that as deletions and duplications decrease and increase gene dosage respectively, for example, deletions and duplications at 16p11.2 are implicated as exhibiting a 'mirrored' effect where a deletion at this locus is associated with increased BMI and a duplication is associated with decreased BMI [3]. There are well-established methods to detect copy number variation (CNVs) in the population, but these CNVs tend to be extremely rare and can have a high false discovery rate. Studies such as the UK Biobank allow us to exploit samples sizes large enough to adequately characterise and associate these large CNVs with continuous phenotypes in the general population.

We investigated the genome-wide effect of large (≥50Kb) deletions and duplications in the UK Biobank by testing their genome-wide association with 204 continuous phenotypes. We called CNVs genome-wide across 120,286 UK Biobank individuals using an established method and then utilised a CNV scoring model to weight CNVs according to their likelihood of being a true positive and tested their association with a range of complex traits.

5.3 Materials and Methods

5.3.1 Probe intensity normalisation and CNV calling

We normalised probe intensity data using an in-house pipeline (refer to Chapter 4, sections 4.3.2 and 4.3.3). CNVs were then cluster-normalised using PennCNV-Affy and PennCNV HMM caller (refer to Chapter 4, sections 4.3.4).

5.3.2 CNV scoring

CNV calls were then scored using PennCNV-pipeline [8]. Each CNV was assigned a real number score of between -1 and 1, with a value of 0 indicating a 100% probability of the CNV being a false positive (**Box S1**). A negative score represents a deletion call, and a positive score indicates duplication (**Fig. S1A**).

5.3.3 CNV selection for association analyses

Calls were grouped by 90% reciprocal intersection to combine multi-breakpoint CNVs and different probesets in different batches. For each CNV call, if at least 90% of both of the CNV calls intersect with each other, they were classified as being the same CNV.

CNVs were then stratified into deletions and duplications. CNVs with a negative score were labelled as deletions and positive scores were labelled as duplications. The negative deletion values were then converted to positive values to allow for consistent directionality between deletions and duplications when associating with traits.

After CNV classification and stratification into duplication and deletions, variants were then only used for inclusion in association analyses if they were present in ≥ 10 individuals with a score of ≥ 0.7 to ensure a 95% statistical power to detect phenotype associations with an effect size (Beta) of ≥ 0.8 (**Table S8**).

5.3.4 Phenotype selection and normalisation

We selected 204 continuous phenotypes in the UK Biobank (**Table S1**). The phenotypes were selected as indicators of overall health, including diet, physical activity socioeconomic status and anthropometric traits. These phenotypes were all residualised correcting for age, sex, recruitment centre. Population stratification was accounted for by correcting for the first five principal components of the population. The phenotypes were then inverse normal quantile transformed to account for associations driven by outliers.

5.3.5 Association analyses

We tested all 204 phenotypes against 798 scored deletions and 1,520 duplications. Linear regression in STATA version 13 was used to test for association between each these deletions and against inverse quantile normalised continuous traits.

5.3.6 *P* value thresholds

We calculated a Bonferroni significance threshold for deletions and duplications separately to account for multiple testing. A total of 161,976 tests were carried out for deletions giving a significance threshold of 3.09 x 10^{-7}, and a total of 329,840 tests were carried out for duplications giving a significance threshold of 1.67 x 10^{-7}.

5.4 Results

5.4.1 Summary of variants tested

There were 3,029,650 deletion and duplication events detected overall with an event being defined as a single CNV detected in a single individual. The majority of quality scores in these events were less than 0.2 (**Fig. S1A**). After merging CNVs with a 90% reciprocal overlap and selecting only CNVs present in ≥ 10 individuals, the number of events was filtered to 467,748 (197,006 deletions and 270,742 duplications). After this filtering procedure, many of the events that had a quality score of less than 0.2 were removed (**Fig. S1B**). A total of 794 deletions and 1,463 duplication events were carried forward for testing (**Table S2A and Table S2B**).

5.4.2 Significant associations

We found 2 deletions and 2 duplications that were associated with ≥ 1 phenotype at genome-wide significance.

5.4.2.1 The 16p11.2 deletion is genome-wide significantly associated with 18 traits

This deletion is published as being associated with multiple traits [4,9] and has a mean quality score of 0.87 across 40 individuals (**Table S2A**). Here we confirm this pleiotropy, or influence on multiple potentially unrelated traits by observing its significance in 18 traits with *P* values ranging from 8.5×10^{-8} to 9.7×10^{-14} (**Table 1**) and its nominal association with a further 65 traits (**Table S4**). The deletion is significantly associated with known published traits including increased fat mass (leg, trunk, body and arm), anthropometric traits (decreased height, waist-hip ratio and higher BMI), socioeconomic status (increased Townsend Deprivation Index, decreased Income and Number of Vehicles) and lung function (decreased Forced Expiratory Volume, Peak Expiratory Flow and Forced Vital Capacity).

5.4.2.2 The 16p12.1 deletion is genome-wide significant against increased hours of television watching

The 16p12.1 deletion is present in 78 individuals with a mean quality score of 0.92 is significantly associated with increased television watching time ($P = 4.3 \times 10^{-9}$) (**Table 1, Fig. 1**) and also nominally associated with lung function phenotypes, sedentary lifestyle, decreased fluid intelligence and longevity (**Table S3**).

5.4.2.3 The 17p12 *CMT1A* duplication is significantly associated with hand grip strength

Duplications at this locus are known to be associated with Charcot-Marie disease. The duplication encompasses the *PMP22* gene which encodes a myelin protein responsible for insulating axons in the peripheral nervous system [10]. The duplication is present in 33 UK Biobank individuals with a mean quality score of 0.99 (**Table 2, Fig. 2**). The duplication is associated with maximum hand grip strength ($P = 1.4 \times 10^{-9}$), and hand grip strength corrected for height ($P = 3.6 \times 10^{-8}$).

5.4.2.4 The 15q13.2-13.3 duplication is significantly associated with household ownership of fewer vehicles

The 15q13.2-13.3 duplication is associated with a decreased vehicle ownership ($P = 2 \times 10^{-8}$) as surrogate measure of socioeconomic status (**Table 2, Fig. 3**). The duplication is present in 60 individuals and is nominally associated with a further 35 traits including a higher Townsend Deprivation Index and lower income (**Table S5**). The duplication is also associated with an increased red blood cell distribution width ($P = 4.7 \times 10^{-7}$) and decreased bone mineral density ($P = 3.2 \times 10^{-5}$).

Table 1. Deletions found to be genome-wide significant when tested against continuous traits. Full phenotype names can be found in **Table S1**

Deletion	Trait	SIN Beta	P	low 95% CI	high 95% CI
chr16:29596230-30190891	LFP	1.612803	9.70E-14	1.188223	2.0373
chr16:29596230-30190891	WC	1.448171	3.10E-13	1.058776	1.8375
chr16:29596230-30190891	BF	1.513584	9.00E-13	1.098471	1.9286
chr16:29596230-30190891	BMI	1.336544	1.80E-11	0.9469122	1.7261
chr16:29596230-30190891	AFP	1.416062	2.30E-11	1.000919	1.8312
chr16:29596230-30190891	FVC	-1.553319	2.30E-11	-2.008664	-1.0979
chr16:29596230-30190891	BMI Rural	1.323336	2.80E-11	0.9337429	1.7129
chr16:29596230-30190891	TDI	1.316318	3.50E-11	0.9269578	1.7056
chr16:21946524-22440319	TV Watching	0.8585542	4.30E-11	0.6033926	1.1137
chr16:29596230-30190891	WHR	1.265148	1.90E-10	0.8760565	1.6542
chr16:29596230-30190891	N Vehicles	-1.339631	2.10E-10	-1.75262	-0.92664
chr16:29596230-30190891	Height	-1.255512	2.80E-10	-1.645325	-0.86569
chr16:29596230-30190891	Height Norm	-1.685404	5.80E-10	-2.218496	-1.1523
chr16:29596230-30190891	TFP	1.306288	7.00E-10	0.8910309	1.7215
chr16:29596230-30190891	FEV1	-1.399517	1.40E-09	-1.852321	-0.94671
chr16:29596230-30190891	BF Mass	1.233629	5.80E-09	0.8183725	1.6488
chr16:29596230-30190891	Hand Grip Max	-1.135978	2.20E-08	-1.534138	-0.73781
chr16:29596230-30190891	PEF	-1.296347	3.00E-08	-1.75487	-0.83782
chr16:29596230-30190891	Income	-1.260254	8.50E-08	-1.721299	-0.7992

Table 2. Duplications found to be genome-wide significant when tested against continuous traits. Full phenotype names can be found in **Table S1**

Duplication	Trait	SIN Beta	P	low 95% CI	high 95% CI
chr17:14047617-15483611	Hand Grip Max	-1.09673	1.40E-09	-1.451734	-0.74172
chr15:30912719-32516949	N Vehicles	-0.7934671	2.00E-08	-1.070443	-0.51649
chr17:14047617-15483611	Ht Hand Grip	-1.032203	3.60E-08	-1.399472	-0.66493

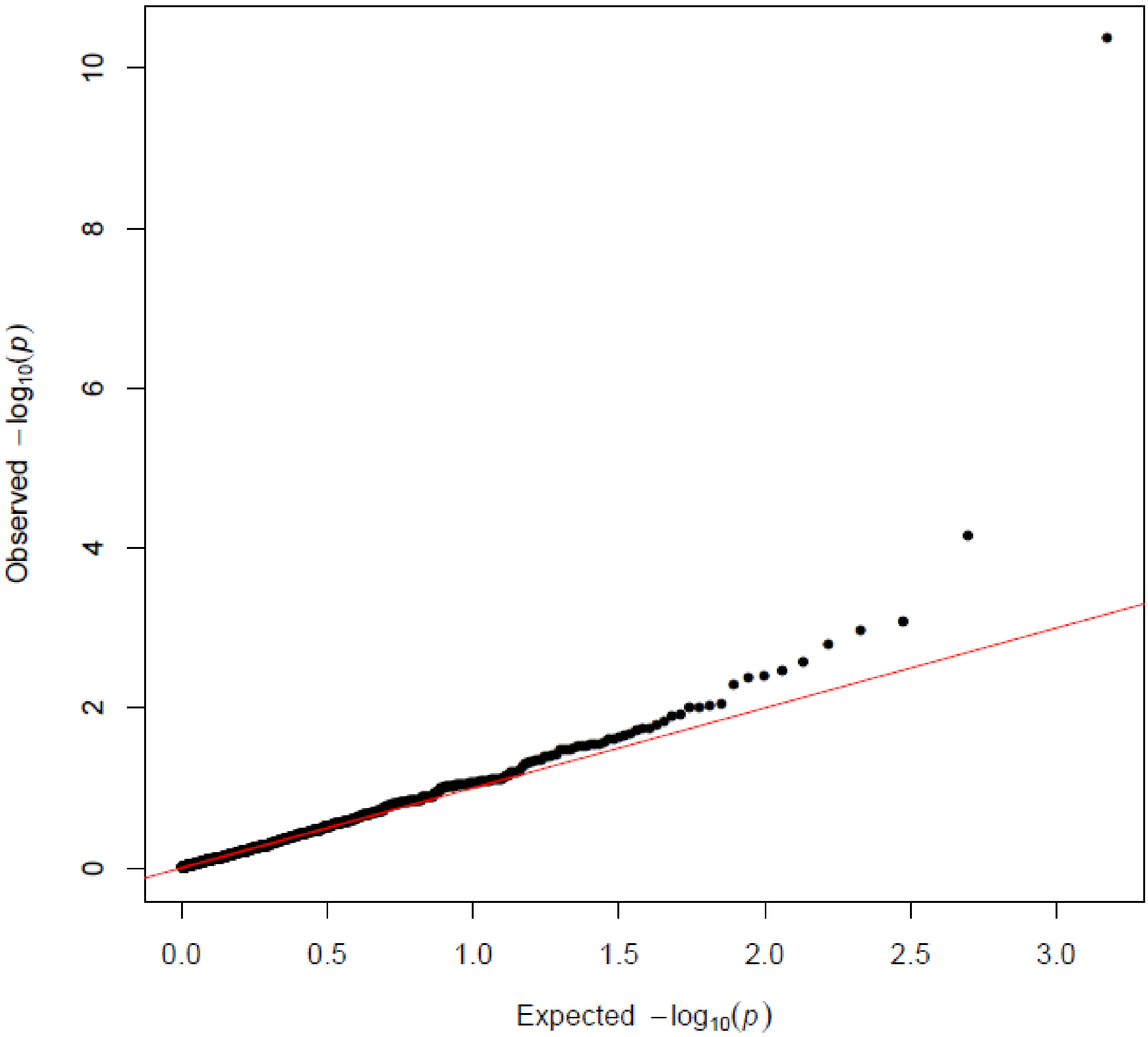


Figure 1. Quantile-quantile plot of genome wide deletions associated with TV watching with strongest signal at 16p12.1

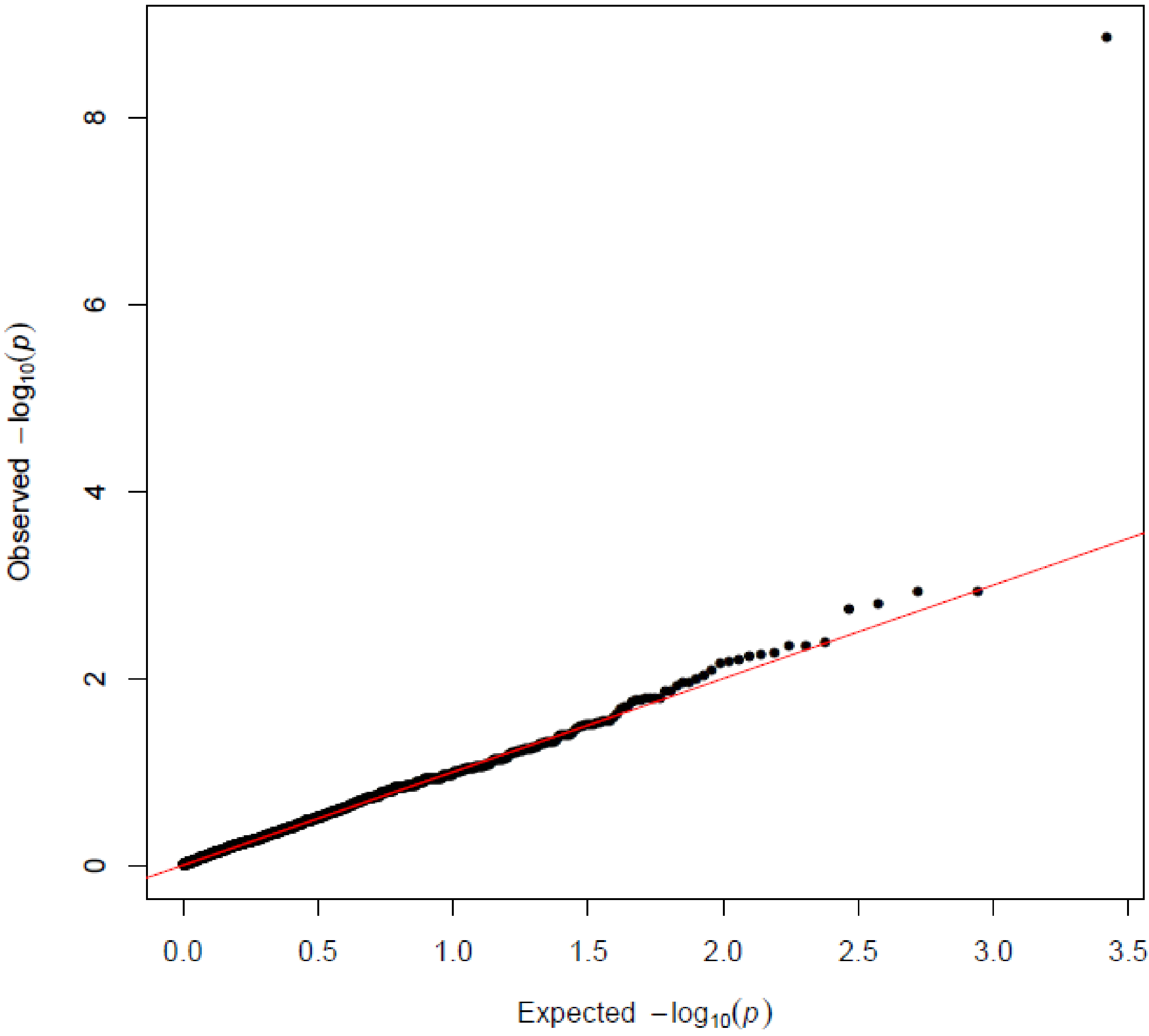


Figure 2. Quantile-quantile plot of P values for genome-wide duplication association with maximum hand grip strength, the 17p12 *CMT1A* duplication is the strongest signal

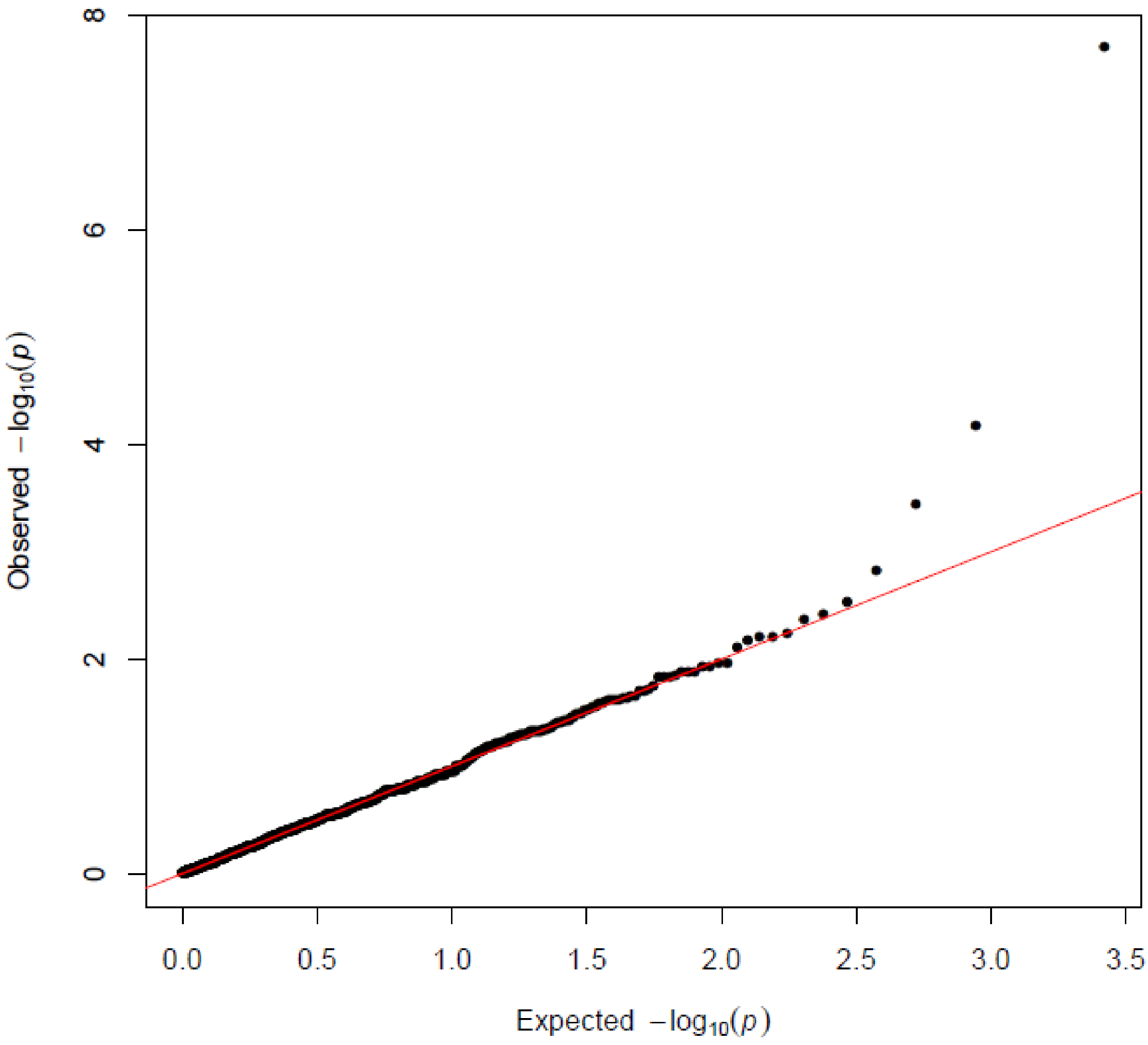


Figure 3. Quantile-quantile plot of P values for genome-wide duplication association with number of vehicles owned, a surrogate measure of social deprivation with 15q13.2-13.3 duplication as the strongest signal

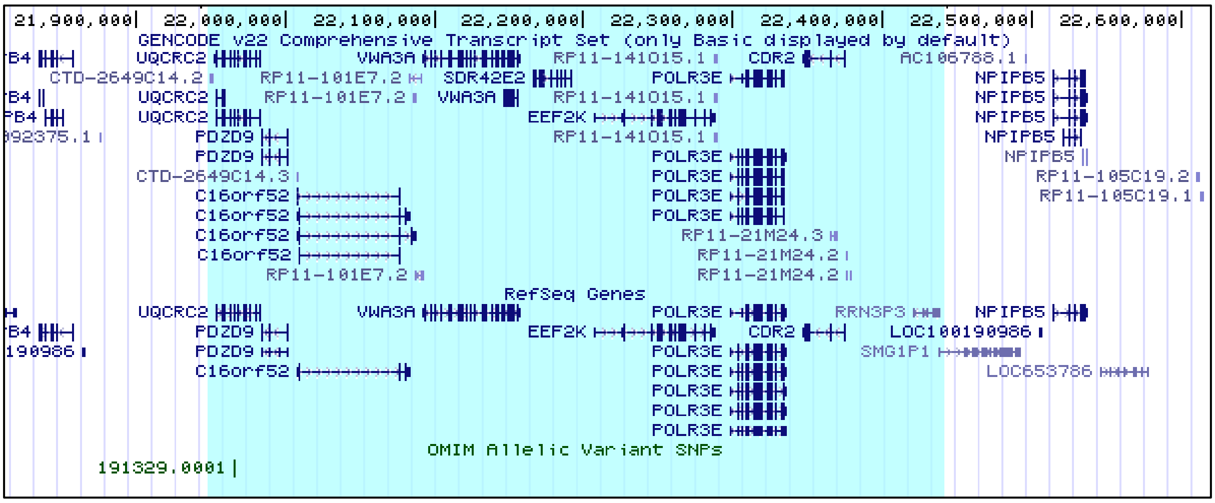


Figure 4. Known GENCODE genes within the 16p12.1 TV watching associated deletion

5.5 Discussion

5.5.1 Four CNVs are significantly associated with continuous traits in 120,286 UK Biobank individuals

We detected 2 deletions and 2 duplications in the UK Biobank that were significantly associated with multiple traits. All four of these CNVs have been previously characterised as being part of a morbidity map of developmental delay [2,4]. The deletion at 16p11.2 is negatively genome-wide significantly associated with multiple traits, whereas the 16p12.1 deletion, 17p12 duplication and the 15q13.2-13.3 are all genome-wide significant for single traits, but nominally associated with a number of related traits. This supports published evidence [11] that both deletions and duplications are pleiotropic, albeit rare in general populations at varying levels of penetrance. Although these CNVs are pleotropic, they tend to be associated with specific phenotype groups, which would be expected when several genes are altered in dosage. We discuss this further in the subsections below.

5.5.1.1 The 16p11.2 deletion remains highly pleiotropic

The 16p11.2 deletion is the most highly significant variant in this study. Both deletions and duplications at this locus have been associated with many traits including cognitive disorders [2,4] and BMI [9]. Specifically, these studies have indicated that the deletions and duplications have separate effects in phenotypes, for example 16p11.2 duplications have been associated with schizophrenia and deletions associated with Autism. Published findings have also indicated a decrease in BMI at this locus, potentially mirroring the effects of the deletion on BMI [3]. We detected 33 duplications at this locus in the UK Biobank (**Table S2A**) but found no association any traits, although we did detect 46 nominal associations (**Table S6**). The nominal associations for duplications at this locus showed a decreased bone mineral density, negative effect on mental health, and a decreased leg fat percentage, and an increase in height. No association with BMI was observed. These results show that deletions are likely to have an effect on increased BMI and related traits, whereas the duplication exhibits a more subtle link to diet and mental health related traits that could influence BMI in certain cases, but not in this UK Biobank cohort. A larger sample set is required to further ascertain the role of this duplication in BMI and to fully compare its effects to the deletion.

5.5.1.2 The 16p12.1 deletion is significantly associated with increased sedentary time

The 16p12.1 deletion is genome-wide significantly associated with television watching time and nominally associated with traits associated with decreased cognitive function, decreased lung function and increased sedentary time (**Table S3**). The deletion intersects with approximately 7 genes including *UQCRC2* (**Fig. 4**). Deleterious *UQCRC2* variants are present in individuals with mitochondrial complex III deficiency nuclear type 5 [12]. Individuals with this complex have variable symptoms including severe metabolic acidosis associated with hyperammonemia (excess of ammonia in the blood), hypoglycaemia, tachypnea (rapid breathing) and mild developmental delay. Previous publications have associated deletions in this region with childhood developmental delay and also neuropsychiatric phenotypes [13] which would fit with increased less severe phenotypes associated with decreased physical activity, lung function and cognitive function. Further replication would be required in other large control populations to further characterise the prevalence and penetrance of the effect of this deletion on these phenotypes in the general population.

5.5.1.3 The 17p12 CMT1A duplication is associated with traits surrogate to neuropathy

Thirty-three UK Biobank individuals have a duplication known to have an association with Charcot-Marie-Tooth disease type 1A (CMT1A) and hereditary neuropathy with liability to pressure palsy (HNPP) [14,15]. The duplication is genome-wide associated with a decreased hand grip strength and is also nominally associated with an increased number of falls and a decreased bone mineral density (BMD) (**Table S7**). These are surrogate traits of individuals with CMT1A and HNPP with a duplication encompassing the *PMP22* gene and are characterised by decreased muscle strength and bone mineral density. A follow up of diagnoses and/or access to medical records for each of these individuals would be required to ascertain if these findings are all syndromic of CMT1A or HNPP or if there is a prevalence in the general population for individuals with the duplication that exhibit a similar but less severe phenotype.

5.5.1.4 The 15q13.2-13.3 duplication is associated with decreased socio-economic status through household ownership of fewer vehicles

The duplication has been associated with various rare disorders including developmental delay and epilepsy [16]. An insufficient number of individuals carried out the cognitive ability test (fluid intelligence) to look for association between this duplication and cognitive function. We did however find an association present with ownership of fewer vehicles and other measures of socioeconomic status such as higher Townsend Deprivation Index and lower income, potential population-level surrogate measures of cognitive impairment. The duplication is nominally associated with a total of 69 traits (**Table S5**) including increased red blood cell distribution width and decreased bone mineral density, indicators of poor health/diet and increased mental health problems. Further phenotypic information would be required to further understand how these findings relate to the more severe published diagnoses exhibited by this duplication.

5.5.2 Limitations of this study

We have shown that it is possible to accurately characterise and score large (≥50Kb) biallellic heterozygous copy number variants and associate them with complex traits in the general population, however, few novel CNVs were detected in this study. There are three possible reasons, breakpoint heterogeneity, sample size and detection technology.

5.5.2.1 Breakpoint heterogeneity

Deletions and duplications with different breakpoints may represent the same genetic variant, or intersect common genes. We used a 90% reciprocal intersect to combine duplications or deletions with breakpoints that vary between individuals, a highly conservative threshold. There are many potential instances whereby two CNVs outside this threshold have the same effect on a given phenotype, for example, a very large deletion in an individual and a deletion in another individual that is, say 10% if its size, but intersecting the same genetic region that has influence on a given trait. A solution to the problem of multi-breakpoint CNVs is to assign quality scores to all SNP probes in each called region and then associate all of these probe quality scores with the phenotype [8]. The region in each CNV that is associated with the trait will then be reflected more strongly in the association statistic.

5.5.2.2 Sample size

Novel associations may in part be due to the sample size. We analysed 120,286 white-British individuals in total with CNV deletion counts ranging from 10 to 6,523 (a maximum 2.7% deletion allele frequency) and 10 to 11,774 (4.9% maximum duplication allele frequency). Sixty-eight percent of total deletions and 71% of duplications were present in ≤ 100 individuals. An increased sample size is needed to allow for a more adequate statistical power. At a CNV frequency of 0.000083% for example (10 in 120,286 individuals), we have 43% power to detect an association with an effect size (Beta) of 0.4. With a sample size of 400,000, the predicted full UK

Biobank population after correcting for population stratification, this power increases to 90% with the same effect size (**Table S8**).

5.5.2.3 Detection Technology

A lack of novel variation is likely due to limitations in the technology. The Affymetrix Axiom array used in this study generally required at least 50 genotyping probes and a size greater than 50Kb for a CNV call to appear convincing when plotting Log R Ratio and B Allele Frequency.

There are therefore three main limitations of the technology used in this study: (a) the study does not have the resolution to confidently detect smaller CNVs in the range of 100bp to 50Kb in size. CNVs generally become more prevalent as they become smaller [17] and it is not yet possible to detect CNVs at smaller sizes genome-wide without high-density signature Array CGH or whole genome sequencing techniques in the sample sizes large enough to detect associations with complex traits. (b) Smaller CNVs are generally more structurally complex and include inversions, translocations and multi-allelic profiles, but have been associated with complex traits [18,19]. SNP arrays simply do not have the resolution to detect these events and would be too obscured by probe intensity noise. (c) SNP arrays are designed around a specific genome build and it is difficult to assay regions using microarray technology in repetitive regions [20]. In the case of this study, the UK Biobank Axiom array is designed around genome build HG19 which comprises of comprises of many genome build gaps, and regions of the genome that haven’t fully been characterised at base-pair resolution level [21]. Many regions of the genome still have an unknown impact on complex traits, and more advanced technologies and methods are required to enable a more comprehensive understanding of these regions.

5.6 References

1. McCarroll SA, Altshuler DM (2007) Copy-number variation and association studies of human disease. Nat Genet 39: S37-42.
2. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, et al. (2011) A copy number variation morbidity map of developmental delay. Nat Genet 43: 838-846.
3. Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, et al. (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature 478: 97-102.
4. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, et al. (2014) Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet 46: 1063-1071.
5. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 42: D986-992.
6. Feuk L, Marshall CR, Wintle RF, Scherer SW (2006) Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 15 Spec No 1: R57-66.
7. Handsaker RE (2015) Large multiallelic copy number variations in humans. Nat Genet 47: 296-303.
8. Macé A, Beckmann JS, Jacquemont S, Reymond A, Kutalik Z (forthcoming 2016) New quality measure for SNP array based CNV detection. Bioinformatics.

9. Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, et al. (2010) A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature 463: 671-675.

10. Roa BB, Garcia CA, Suter U, Kulpa DA, Wise CA, et al. (1993) Charcot-Marie-Tooth disease type 1A. Association with a spontaneous point mutation in the PMP22 gene. N Engl J Med 329: 96-101.

11. O'Donovan MC, Kirov G, Owen MJ (2008) Phenotypic variations on the theme of CNVs. Nat Genet 40: 1392-1393.

12. Miyake N, Yano S, Sakai C, Hatakeyama H, Matsushima Y, et al. (2013) Mitochondrial complex III deficiency caused by a homozygous UQCRC2 mutation presenting with neonatal-onset recurrent metabolic decompensation. Hum Mutat 34: 446-452.

13. Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, et al. (2010) A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat Genet 42: 203-209.

14. Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, et al. (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell 66: 219-232.

15. Saporta AS, Sottile SL, Miller LJ, Feely SM, Siskind CE, et al. (2011) Charcot-Marie-Tooth disease subtypes and genetic testing strategies. Ann Neurol 69: 22-33.

16. van Bon BW, Mefford HC, Menten B, Koolen DA, Sharp AJ, et al. (2009) Further delineation of the 15q13 microdeletion and duplication syndromes: a clinical spectrum varying from non-pathogenic to a severe outcome. J Med Genet 46: 511-523.

17. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, et al. (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517: 608-611.

18. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, et al. (2016) Schizophrenia risk from complex variation of complement component 4. Nature 530: 177-183.

19. Boettger LM, Salem RM, Handsaker RE, Peloso GM, Kathiresan S, et al. (2016) Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat Genet 48: 359-366.

20. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40: 1166-1174.

21. International Human Genome Sequencing C (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-945.

5.7 Supplementary Information

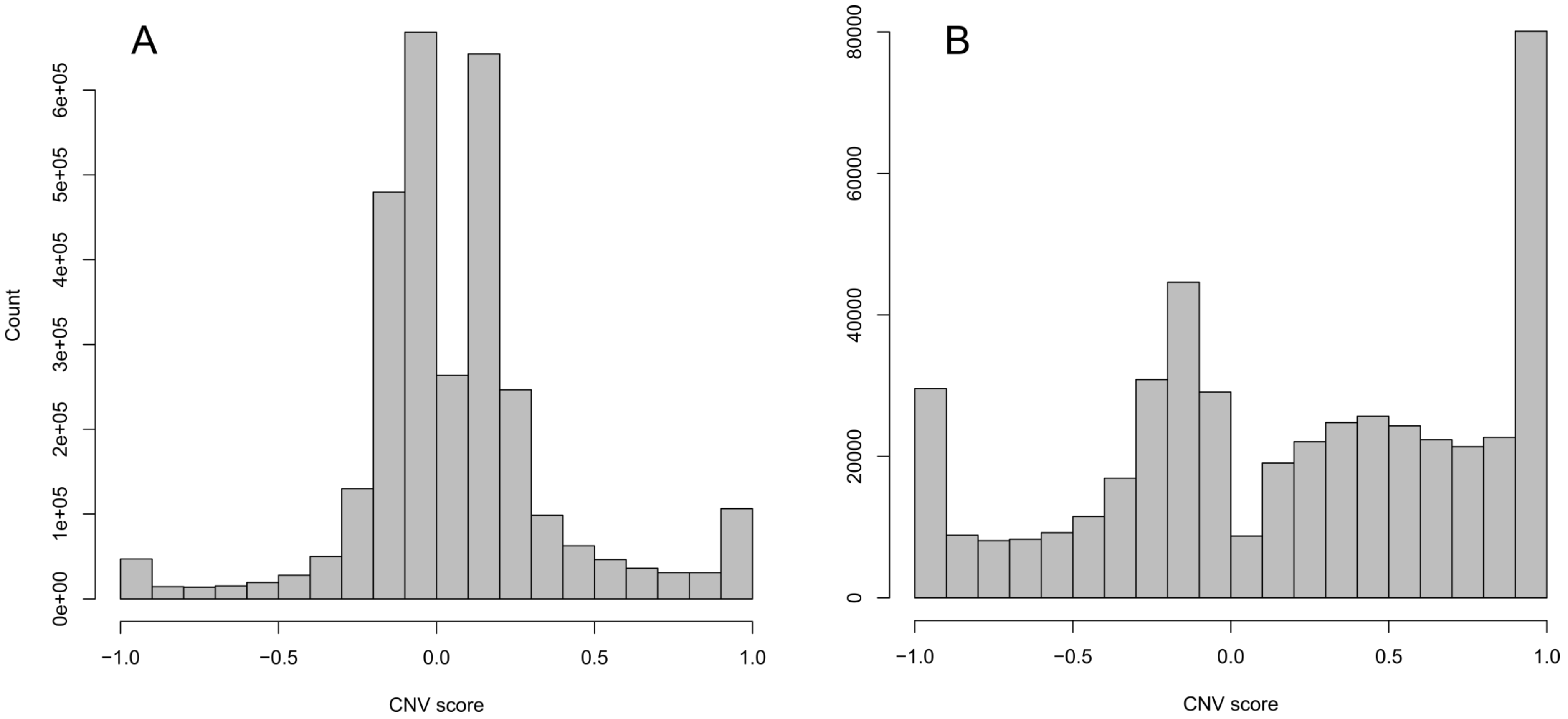


Figure S1. Distribution of CNV scores, raw **(A)** and filtered **(B)** negative scores indicate a deletion event, positive scores indicate duplications. Duplication and deletion are stratified during association analyses

Table S1. A list of all 204 continuous, single inverse quantile normalised phenotypes tested against each deletion and duplication

Trait	Category
BMI lean body size at age 10	Anthropometric
BMI normal body size at age 10	Anthropometric
BMI large body size at age 10	Anthropometric
Height in those short at age 10	Anthropometric
Height in those average height age 10	Anthropometric
Height in those tall at age 10	Anthropometric
Whole body fat mass	Anthropometric
Whole body fat free mass	Anthropometric
Impedance of whole body	Anthropometric
Whole body water mass	Anthropometric
Height	Anthropometric
Sitting to Standing Ratio	Anthropometric
BMI all	Anthropometric
BMI in women only	Anthropometric
BMI in never smokers only	Anthropometric
BMI in urban	Anthropometric
BMI in rural	Anthropometric
Hi circumference	Anthropometric
Waist circumference	Anthropometric
WHR	Anthropometric
WHR adjusted for BMI	Anthropometric
Skeletal Mass Index	Anthropometric
Arterial stiffness index	Arterial stiffness
Red blood cell distribution width	Biomarkers
Northing	Birth location
Easting	Birth location
Birth Weight	Birth Weight
First Child Birth Weight	Birth Weight
DBP	Blood Pressure
SBP	Blood Pressure
Arm fat percentage	Body Fat
Leg fat %	Body Fat
Trunk fat%	Body Fat
Trunk fat % to leg fat % ratio	Body Fat
Trunk fat % to limb fat % ratio	Body Fat
Bodyfat %	Body Fat
Bodyfat % in women only	Body Fat
BMD	Bone mineral density
BMD of heel based on T score	Bone mineral density
Pairs test	Cognitive function
Fluid intelligence test score	Cognitive function

Reaction time	Cognitive function
Numeric memory	Cognitive function
Month of birth	Demographics
Participants age at death	Demographics
Diabetes liability score in controls	Diabetes
Diabetes liability score in all	Diabetes
Diet variation from week to week	Diet
Caffeine consumption caffeinated only	Diet
Caffeine consumption	Diet
Decaf coffee intake	Diet
Tea Intake	Diet
Energy intake	Diet
Estimated intake of protein, based on food and beverage consumption yesterday	Diet
Healthy diet	Diet
Prudent diet	Diet
Low calorie drinks	Diet
Fat intake	Diet
Carbohydrate intake	Diet
Sat fat intake	Diet
Polyunsaturated fat intake	Diet
Sugar intake	Diet
Fibre intake	Diet
Iron intake	Diet
Vitamin B6	Diet
Vitamin B12	Diet
Vitamin C	Diet
Potassium intake	Diet
Magnesium intake	Diet
Retinol intake	Diet
Carotene intake	Diet
Vitamin D	Diet
Starch intake	Diet
Vitamin E	Diet
Water intake	Diet
Fizzy drinks	Diet
Yoghurt consumption - any	Diet
Yoghurt consumption - full fat	Diet
Fried potato intake	Diet
Fried chicken intake	Diet
Folate intake	Diet
Fried foods	Diet
Brassica	Diet
Yoghurt consumption – low fat	Diet
Bread intake (white)	Diet

Bread intake (brown)	Diet
Salt consumption	Diet
Percentage of protein in total calorie intake	Diet
Percentage of carbohydrates in total calorie intake	Diet
Percentage of fat in total calorie intake	Diet
A single Diet variable? "western diet"	Diet
Calcium intake	Diet
Vegetable intake	Diet
Body size at 10 years	Early life
Height at 10 years	Early life
Mobile phone use (years used at least once)	Electronic devices
Computer games	Electronic devices
Father current age	Family
Father age of death	Family
Mother current age	Family
Mothers age at death	Family
Mean age at death of both parents	Family
Number of brothers	Family
Number of sisters	Family
Number of older siblings	Family
All pregnancies	Female specific
Number of term pregnancies	Female specific
Age contraceptive use stopped	Female specific
Length menstrual	Female specific
Number of live births	Female specific
Age of primiparous women at birth of child	Female specific
Age at first live birth	Female specific
Age at last live birth	Female specific
Number of stillbirths	Female specific
Number of spontaneous miscarriages	Female specific
Number of pregnancy terminations	Female specific
Age at oophorectomy	Female specific
Age at hysterectomy	Female specific
participant BW females	Female specific
Menarche	Female specific
Menopause reprogen	Female specific
Menopause continuous	Female specific
Waist circumference women only	Female specific
WHR in women only	Female specific
WHR adjusted for BMI in women only	Female specific
BMI Pre Menopause	Female specific
BMI Post Menopause	Female specific
Reproductive lifespan	Female specific
Cigarettes per day	Health
Units per day	Health

FFQ - estimated alcohol consumption	Health
Time to first cigarette	Health
Difficulty not smoking	Health
Number of stop smoking attempts	Health
Maximal hand grip	Health
Pulse rate	Health
Unhealthy phenotype	Health
N cancer diseases	Health
N non cancer diseases	Health
N treatments taken	Health
N operations	Health
Overall health rating	Health
Tinnitus severity	Health
Hearing left ear	Health
Hearing right ear	Health
Falls in last year	Health
Age wore glasses if started at 40 years old plus	Health
Age glasses	Health
Loud music exposure	Health
PAD - leg pain action taken	Health
Visual acuity right	Health
Visual acuity left	Health
Years since last mammogram screen	Health
Hand Grip Corrected for Height	Health
Basal Metabolic Rate	Health
Smoking pack years	Health
Mental health	Health
Time since bowel cancer screen	Health
Years since cervical smear	Health
Noisy workplace	Health
Job type	Occupation
IPAQ category	Physical activity
Stair climbing	Physical activity
TV watching	Physical activity
Time spent on computer	Physical activity
Time spent driving	Physical activity
Sedentary variable	Physical activity
A single Phys activity variable ?IPAQ	Physical activity
Happiness	Psychosocial factors
Eysenck personality questionnaire	Psychosocial factors
Subjective wellbeing	Psychosocial factors
Longest depression spell	Psychosocial factors
Longest period of disinterest	Psychosocial factors
Number of depression episodes	Psychosocial factors
Frequency of friends/family visits	Psychosocial factors

Confide	Psychosocial factors
Drive faster than motorway speed limit	Random
Townsend deprivation index	SES
Annual household income	SES
Age completed full time education	SES
Number in household	SES
Number of vehicles in household	SES
Years in job	Occupation
Hours per week worked	Occupation
Age when 1st had sex	Sexual factors
Number of sexual partners	Sexual factors
Number of same sex sexual partners	Sexual factors
Insomnia	Sleep
Ease of getting up	Sleep
Nap	Sleep
Narcolepsy	Sleep
Hours Slept	Sleep
Morning or evening category	Sleep
FEV1	Spirometry
FVC	Spirometry
PEF	Spirometry
Time spent outdoors in winter	Sun exposure
Time spent outdoors in summer	Sun exposure
Skin colour	Sun exposure
Ease of tanning	Sun exposure
Childhood sunburn occasions	Sun exposure
Facial ageing	Sun exposure
Sunscreen	Sun exposure
Solarium use	Sun exposure

Table S2. Accessible from: https://goo.gl/AUbFLr shows all CNVs that were detected with probe and frequency metrics for 794 deletions **(A)** and 1,464 duplications **(B)**. The metrics listed are: CNV locus, CNV count and the mean values of CNV size, CNV score, LRR, LRR SD, BAF, BAF SD, BAF drift, waviness factor, N CNVs per genome, PennCNV confidence and probe count.

Table S3. All nominal phenotype associations for 16p12.1 TV Watching associated deletion

Phenotype	**Normalised Beta**	***P***	**Lower 95% CI**	**Upper 95% CI**
TV watching	0.8585542	4.30E-11	0.6033926	1.1137
Time spent on computer	-0.4880257	2.30E-04	-0.7480658	-0.22798
FVC	-0.4863822	3.50E-04	-0.7532044	-0.21955
Mobile phone use (years used at least once)	-0.4639631	5.90E-04	-0.7287193	-0.1992
FEV1	-0.4631963	6.20E-04	-0.7285216	-0.19787
Sedentary variable	0.4286291	1.00E-03	0.1733301	0.68392
Annual household income	-0.4816897	1.10E-03	-0.7698644	-0.1935
PEF	-0.4425917	1.20E-03	-0.7112616	-0.17392
SBP	0.404441	1.90E-03	0.148946	0.6599
Maximal hand grip	-0.3923652	2.90E-03	-0.6501609	-0.13456
Father current age	-0.7600937	4.40E-03	-1.282785	-0.2374
Mean age at death of both parents	-0.5326593	4.60E-03	-0.9007277	-0.16459
Age wore glasses if started at 40 years old plus	0.6299151	6.10E-03	0.1800672	1.0797
Arm fat percentage	0.3596143	6.70E-03	0.0998804	0.61934
Trunk fat%	0.3581075	6.90E-03	0.0983099	0.6179
Bodyfat %	0.352788	7.80E-03	0.0930662	0.6125
Hand Grip Corrected for Height	-0.346695	8.40E-03	-0.6044938	-0.08889
DBP	0.3319426	1.10E-02	0.0764024	0.58748
Reaction time	0.3365049	1.10E-02	0.0771972	0.59581
Number of vehicles in household	-0.3282073	1.20E-02	-0.583618	-0.07279
Prudent diet	-0.346098	1.40E-02	-0.6232618	-0.06893
Bodyfat % in women only	0.4360192	1.60E-02	0.0823873	0.7896
Mother current age	-0.4920842	1.60E-02	-0.8938538	-0.09031
Overall health rating	0.3113922	1.70E-02	0.056138	0.56664
Skin colour	-0.3146711	1.70E-02	-0.5724723	-0.05686
Stair climbing	-0.3138151	1.70E-02	-0.5714966	-0.05613
Skeletal Mass Index corrected for height	-0.30884	2.00E-02	-0.5688376	-0.04884
Whole body fat free mass	-0.2996969	2.40E-02	-0.5597607	-0.03963
Fluid intelligence test score	-0.6284031	2.50E-02	-1.176001	-0.0808
Leg fat%	0.2946105	2.60E-02	0.0350008	0.55422
Whole body water mass	-0.2939619	2.70E-02	-0.5540258	-0.03389
Trunk fat% to leg fat % ratio	0.2913033	2.80E-02	0.0312443	0.55136
Eysneck personality questionnaire	0.3132946	3.00E-02	0.0309434	0.59564
Number of brothers	0.2785603	3.30E-02	0.022996	0.53412
Red blood cell distribution width	-0.3965896	3.40E-02	-0.7636223	-0.02955
Impedance of whole body	0.2750501	3.80E-02	0.0150559	0.53504
Skeletal Mass Index	-0.2734454	3.90E-02	-0.5334085	-0.01348
Full fat yoghurt consumption	-0.5860922	4.20E-02	-1.150401	-0.02178
Maximal hand grip in females	-0.3599201	4.30E-02	-0.7083641	-0.01147
Number of sisters	0.2664572	4.30E-02	0.0086999	0.52421

Height	-0.2576342	4.80E-02	-0.5129958	-0.00227
N treatments taken	0.2559862	4.90E-02	0.0007079	0.51126

Table S4. All nominal phenotype associations for 16p11.2 pleiotropic deletion

Phenotype	Normalised Beta	*P*	Low 95% CI	High 95% CI
Leg fat %	1.612803	9.70E-14	1.188223	2.0373
Waist circumference	1.448171	3.10E-13	1.058776	1.8375
Bodyfat %	1.513584	9.00E-13	1.098471	1.9286
BMI all	1.336544	1.80E-11	0.9469122	1.7261
Arm fat percentage	1.416062	2.30E-11	1.000919	1.8312
FVC	-1.553319	2.30E-11	-2.008664	-1.0979
BMI in rural	1.323336	2.80E-11	0.9337429	1.7129
Townsend deprivation index	1.316318	3.50E-11	0.9269578	1.7056
WHR	1.265148	1.90E-10	0.8760565	1.6542
Number of vehicles in household	-1.339631	2.10E-10	-1.75262	-0.92664
Height	-1.255512	2.80E-10	-1.645325	-0.86569
Height in those average height age 10	-1.685404	5.80E-10	-2.218496	-1.1523
Trunk fat %	1.306288	7.00E-10	0.8910309	1.7215
FEV1	-1.399517	1.40E-09	-1.852321	-0.94671
Whole body fat mass	1.233629	5.80E-09	0.8183725	1.6488
Maximal hand grip	-1.135978	2.20E-08	-1.534138	-0.73781
PEF	-1.296347	3.00E-08	-1.75487	-0.83782
Annual household income	-1.260254	8.50E-08	-1.721299	-0.7992
Hip circumference	1.018165	3.10E-07	0.628116	1.4082
Bodyfat % in women only	1.620937	1.10E-06	0.9684851	2.2733
BMI normal body size at age 10	1.552542	1.20E-06	0.9254065	2.1796
BMI in current smokers	1.344693	4.00E-06	0.7731275	1.9162
Waist circumference women only	1.439326	5.10E-06	0.8210241	2.0576
BMI in never smokers only	1.475795	5.40E-06	0.8398585	2.1117
Time spent driving	-0.9580499	6.30E-06	-1.373764	-0.54233
Diabetes liability score in controls	-0.9791062	1.30E-05	-1.419857	-0.53835
BMI in women only	1.370447	1.40E-05	0.7516056	1.9892
BMI in women only	1.370447	1.40E-05	0.7516056	1.9892
Hand Grip Corrected for Height	-0.8661466	2.00E-05	-1.26433	-0.46796
Height in those tall at age 10	-2.441827	2.30E-05	-3.57213	-1.3115
Sunscreen	-0.814709	1.00E-04	-1.22556	-0.40385
Stair climbing	-0.7797053	1.40E-04	-1.181782	-0.37762
Time spent on computer	-0.8157083	2.00E-04	-1.246067	-0.385
Overall health rating	0.7642789	2.30E-04	0.3581525	1.1704
Vegetable intake	-0.7676014	2.90E-04	-1.183155	-0.35204
BMI all	1.088305	4.20E-04	0.4837272	1.6928
Number in household	-0.7056317	6.20E-04	-1.109847	-0.30141
BMI lean body size at age 10	1.26099	8.40E-04	0.5211844	2.0007
Time spent outdoors in winter	0.70056	9.60E-04	0.2848643	1.1162
Healthy diet	-0.8485359	1.20E-03	-1.363297	-0.33377
All pregnancies	-0.9897589	1.70E-03	-1.60956	-0.36995

Menarche	-0.9782351	2.00E-03	-1.598035	-0.3584
Drive faster than motorway speed limit	-0.7567054	2.50E-03	-1.246728	-0.26668
Number of term pregnancies	-0.9151596	3.80E-03	-1.534988	-0.29533
Number of live births	-0.8940787	4.70E-03	-1.513911	-0.27424
Diabetes liability score in all	-0.5887976	5.50E-03	-1.004339	-0.17325
Narcolepsy	0.5558702	5.70E-03	0.1618197	0.94992
Height in females	-0.864335	6.20E-03	-1.483798	-0.24487
Skeletal Mass Index	0.5793441	6.30E-03	0.163774	0.99491
N treatments taken	0.5411149	6.50E-03	0.1513761	0.93085
Government health guidelines ignored	0.7584168	6.60E-03	0.2107963	1.306
Percentage of protein in total calorie intake	-2.700304	6.90E-03	-4.659859	-0.7407
N non cancer diseases	0.5225594	8.60E-03	0.1327637	0.91235
Hearing left ear	1.07057	8.70E-03	0.2705644	1.8705
Job type	0.9270088	8.70E-03	0.2348013	1.6192
BMI in female current smokers	1.280221	9.70E-03	0.3098191	2.2506
Hearing right ear	1.018385	1.30E-02	0.2184293	1.818
Confide	-0.5011031	1.40E-02	-0.899316	-0.10289
Reaction time	0.5049318	1.40E-02	0.10046	0.9094
Sitting to Standing Ratio	0.4787355	1.60E-02	0.0885307	0.86894
Sedentary variable	-0.4758039	1.70E-02	-0.8655884	-0.08601
Pulse rate	0.5269356	1.80E-02	0.0913378	0.96253
Visual acuity left	1.654295	1.90E-02	0.2690783	3.0395
BMD of heel based on T score	0.5496394	2.30E-02	0.0775115	1.0217
TV watching	0.4827105	2.30E-02	0.0676119	0.8978
WHR in women only	0.7103096	2.40E-02	0.0921605	1.3284
Estimated intake of protein	-2.24272	2.50E-02	-4.202325	-0.28311
IPAQ category	-0.4822107	2.60E-02	-0.9075319	-0.05688
Years since last mammogram screen	0.7758923	2.80E-02	0.0829901	1.4687
Age contraceptive use stopped	0.9779384	2.90E-02	0.1021329	1.8537
Prudent diet	-0.5683344	3.00E-02	-1.08303	-0.05363
Impedance of whole body	-0.4511843	3.30E-02	-0.8668095	-0.03555
Participants age at death	1.344131	3.40E-02	0.1029541	2.5853
Units per day	-0.5189651	3.40E-02	-0.9980267	-0.0399
Skeletal Mass Index corrected for height	0.4446149	3.60E-02	0.0289822	0.86024
BMI Pre Menopause	1.173993	4.10E-02	0.0453928	2.3025
Frequency of friends/family visits	0.4065196	4.50E-02	0.0083101	0.80472
Father current age	0.6643692	4.60E-02	0.0115539	1.3171
IPAQ	-0.429547	4.80E-02	-0.8549615	-0.00413

Table S5. All nominal phenotype associations for the 15q13.2-13.3 duplication associated with decreased socio-economic status through household ownership of fewer vehicles

Phenotype	Normalised Beta	*P*	Lower 95% CI	Upper 95% CI
Number of vehicles in household	-0.7934671	2.00E-08	-1.070443	-0.51649
Townsend deprivation index	0.7504716	1.10E-07	0.473865	1.027
BMD	-0.5886072	3.20E-05	-0.8657661	-0.31144
Job type	0.8455831	1.50E-04	0.4077676	1.2833
Bodyfat % in women only	0.7351078	1.00E-03	0.2974144	1.1728
BMD of heel based on T score	-0.4816648	1.10E-03	-0.77063	-0.19269
Bread intake (brown)	-0.4669787	1.10E-03	-0.7467287	-0.18722
Time to first cigarette	-1.053852	1.30E-03	-1.697133	-0.41057
BMI in women only	0.6808006	1.80E-03	0.2537589	1.1078
BMI in women only	0.6808006	1.80E-03	0.2537589	1.1078
Age completed full time education	0.605733	3.10E-03	0.2042946	1.0071
SBP	0.5543254	3.10E-03	0.1866929	0.92195
Hearing left ear	1.462938	3.40E-03	0.4831897	2.4426
Bread intake (white)	0.425191	4.30E-03	0.1331534	0.71722
Percentage of protein in total calorie intake	-1.233661	5.80E-03	-2.110017	-0.3573
Facial ageing	0.5213115	6.30E-03	0.1473479	0.89527
BMI Post Menopause	0.6736369	7.00E-03	0.1843827	1.1628
Pairs test	0.3700102	8.90E-03	0.0928008	0.64721
Arm fat percentage	0.3689933	9.70E-03	0.0893173	0.64866
Age glasses	-0.8311139	9.80E-03	-1.461818	-0.20041
Age contraceptive use stopped	0.7411537	1.00E-02	0.1758056	1.3065
Age wore glasses if started at 40 years old plus	0.6432823	1.00E-02	0.1533174	1.1332
Annual household income	0.5045515	1.10E-02	0.1164479	0.89265
Red blood cell distribution width	1.109665	1.30E-02	0.2337598	1.9855
BMI in urban	-1.162337	1.60E-02	-2.103984	-0.22069
Caffeine consumption	0.4669648	1.80E-02	0.0786385	0.85529
Drive faster than motorway speed limit	-0.3790809	1.80E-02	-0.692934	-0.06522
Time spent on computer	0.4433739	1.80E-02	0.0756033	0.81114
WHR in women only	1.052149	1.80E-02	0.1781279	1.9261
Water intake	0.4223711	2.40E-02	0.0548374	0.7899
TV watching	0.622166	2.50E-02	0.0794453	1.1648
Years since cervical smear	-0.6757794	2.50E-02	-1.266637	-0.08492
Age at last live birth	0.6203414	2.80E-02	0.066686	1.1739
Brassica	1.265118	2.80E-02	0.133727	2.3965
PEF	-0.3357399	2.80E-02	-0.6343218	-0.03715
Difficulty not smoking	-1.531572	2.90E-02	-2.902295	-0.16084
Cups of coffee per day	0.4043535	3.10E-02	0.0366297	0.77207
Noisy workplace	1.079165	3.10E-02	0.0996478	2.0586
Pulse rate	0.5968048	3.10E-02	0.0534383	1.1401

Western diet	0.4328889	3.40E-02	0.0336412	0.83213
Waist circumference women only	0.9419749	3.50E-02	0.067621	1.8163
Time spent outdoors in winter	0.320338	3.60E-02	0.0214366	0.61923
Age of primiparous women at birth of child	0.8467382	3.70E-02	0.0493649	1.6441
Number in household	0.5719082	3.90E-02	0.0286506	1.1151
Maximal hand grip	-0.2909439	4.00E-02	-0.5681582	-0.01372
Morning or evening category	0.5898381	4.10E-02	0.0242035	1.1554
Number of live births	0.4417065	4.30E-02	0.0139988	0.86941
Number of stop smoking attempts	-0.6342306	4.30E-02	-1.248359	-0.0201
Fried chicken intake	-1.154334	4.50E-02	-2.285528	-0.02313
Skin colour	-0.3026172	4.50E-02	-0.598121	-0.00711
Mental health	-0.3774734	4.60E-02	-0.7478098	-0.00713
Time spent driving	-0.5748952	4.60E-02	-1.140667	-0.00912
Healthy diet	-0.3398528	4.70E-02	-0.6758847	-0.00382
bilateral oophorectomy age	-1.975119	4.80E-02	-3.93189	-0.01834
Number of term pregnancies	0.4306655	4.80E-02	0.0029579	0.85837

Table S6. All nominal phenotype associations for 16p11.2 pleiotropic duplication

Phenotype	**Normalised Beta**	***P***	**Lower 95% CI**	**Upper 95% CI**
BMD	-0.7297269	1.60E-04	-1.107994	-0.35145
BMD of heel based on T score	-0.7629354	1.90E-04	-1.164276	-0.36159
TV watching	0.7096424	2.30E-04	0.3319108	1.0873
Visual acuity right	1.331176	4.30E-04	0.590731	2.0716
Reaction time	0.6682498	5.40E-04	0.2899228	1.0465
Visual acuity left	1.267845	7.90E-04	0.5274409	2.0082
Sitting to Standing Ratio	-0.6113276	1.50E-03	-0.9896245	-0.23303
Mental health	0.5916561	2.60E-03	0.2062193	0.97709
SBP	-0.5770981	2.80E-03	-0.955277	-0.19891
Townsend deprivation index	0.5592894	3.70E-03	0.1817486	0.93683
Fibre intake	-1.020727	3.90E-03	-1.713247	-0.3282
Leg fat%	-0.5642107	4.10E-03	-0.9491236	-0.17929
Bread intake (brown)	-0.5330464	5.70E-03	-0.9110149	-0.15507
Subjective wellbeing	0.9768155	5.90E-03	0.2819863	1.6716
Menarche	0.710187	7.90E-03	0.1858528	1.2345
Yoghurt consumption	-0.937842	8.00E-03	-1.630753	-0.24493
Low fat yoghurt consumption	-0.9149886	9.70E-03	-1.607921	-0.22205
Fluid intelligence test score	-0.856544	1.00E-02	-1.511461	-0.20162
Height in those tall at age 10	0.8437827	1.70E-02	0.1479983	1.5395
Job type	0.7818702	1.90E-02	0.126303	1.4374
Number of depression episodes	1.035792	2.10E-02	0.1575277	1.914
Magnesium intake	-0.8073547	2.20E-02	-1.499721	-0.11498
Height in females	0.5522728	2.30E-02	0.0753757	1.029
Hand Grip Corrected for Height	-0.4342822	2.40E-02	-0.8126209	-0.05594
Birth Weight	0.6745091	2.50E-02	0.0829331	1.266
Annual household income	-0.4739807	2.60E-02	-0.8925484	-0.0554
Vitamin D	-0.78734	2.60E-02	-1.480264	-0.09441
Height	0.4236833	2.80E-02	0.0457043	0.80166
Maximal hand grip in females	-0.5347832	2.80E-02	-1.011968	-0.05759
Units per day	-0.4944835	2.80E-02	-0.9342627	-0.0547
N treatments taken	-0.4211802	2.90E-02	-0.7990359	-0.04332
PEF	-0.4627831	3.00E-02	-0.8814797	-0.04408
Number of stillbirths	-0.5247835	3.10E-02	-1.001968	-0.04759
DBP	-0.4133509	3.20E-02	-0.7915987	-0.0351
FEV1	-0.4510913	3.20E-02	-0.8645783	-0.0376
FVC	-0.4546639	3.20E-02	-0.8704857	-0.03884
Happiness	0.7123597	3.30E-02	0.0573362	1.3673
Time spent outdoors in summer	0.4313999	3.50E-02	0.0303581	0.83244
Calcium intake	-0.7421281	3.60E-02	-1.434846	-0.04941
Number of pregnancy terminations	0.5226419	3.70E-02	0.0306767	1.0146
Eysenck personality questionnaire	0.441243	3.90E-02	0.0221855	0.8603

Tea Intake	0.3981386	3.90E-02	0.0197948	0.77648
Birth Weight	0.5688798	4.00E-02	0.0247884	1.1129
Time spent driving	-0.3959318	4.00E-02	-0.7742727	-0.01759
Iron intake	-0.7103292	4.40E-02	-1.402736	-0.01792
Number of vehicles in household	-0.3872427	4.50E-02	-0.7652992	-0.00918

Table S7. The 17p12 CMT1A duplication is associated with traits surrogate to neuropathy

Phenotype	Normalised Beta	*P*	Lower 95% CI	Upper 95% CI
Maximal hand grip	-1.09673	1.40E-09	-1.451734	-0.74172
Hand Grip Corrected for Height	-1.032203	3.60E-08	-1.399472	-0.66493
Falls in last year	0.7977108	1.10E-05	0.4427189	1.1527
Skeletal Mass Index corrected for height	-0.6512328	6.40E-04	-1.025061	-0.2774
Skeletal Mass Index	-0.6375947	8.30E-04	-1.011372	-0.2638
Maximal hand grip in females	-0.8030102	1.90E-03	-1.30904	-0.29698
Impedance of whole body	0.579188	2.40E-03	0.2053623	0.95301
BMD	-0.5409992	3.90E-03	-0.908246	-0.17375
Time spent outdoors in winter	-0.5398127	4.70E-03	-0.9137125	-0.16591
IPAQ	-0.5588921	6.80E-03	-0.9634133	-0.1543
Townsend deprivation index	0.4840943	7.40E-03	0.1297887	0.83839
Body size at 10 years	-0.4824637	7.70E-03	-0.8373388	-0.12758
BMD of heel based on T score	-0.5136336	8.20E-03	-0.8945039	-0.13276
Age at first live birth	-0.7226751	1.20E-02	-1.285575	-0.15977
Overall health rating	0.4532904	1.20E-02	0.0987242	0.80785
Water intake	0.449465	1.30E-02	0.0946593	0.80427
Whole body water mass	-0.4699738	1.40E-02	-0.8439056	-0.096
Number of older siblings	-0.994492	1.50E-02	-1.794434	-0.19455
Tea Intake	-0.428767	1.80E-02	-0.7838223	-0.07371
DBP	-0.4232368	1.90E-02	-0.7782028	-0.06827
Narcolepsy	0.416498	2.10E-02	0.0614582	0.77153
Annual household income	-0.5070161	2.20E-02	-0.939293	-0.07473
Age at last live birth	-0.6457073	2.50E-02	-1.210047	-0.08136
Leg fat%	0.4405913	2.60E-02	0.052937	0.82824
Mental health	0.4234256	2.60E-02	0.0496363	0.79721
Bread intake (white)	0.4201922	2.80E-02	0.0464717	0.79391
IPAQ category	-0.4433139	2.80E-02	-0.8394057	-0.0472
Number of stop smoking attempts	0.6527785	2.90E-02	0.0672577	1.2382
Whole body fat free mass	-0.41454	3.00E-02	-0.7884741	-0.0406
WHR in women only	0.5361939	3.20E-02	0.0475527	1.0248
Mobile phone use (years used at least once)	-0.3794031	3.60E-02	-0.7344609	-0.02434
N operations	0.3792775	3.60E-02	0.02425	0.7343

Table S8. Power calculations under the additive model for Beta values < 1 at a CNV frequency equal to 0.000083 (~10 in 120,283 individuals) in the current number of individuals and a predicted approximate full UK Biobank sample size when accounting for population stratification of 400,000

Effect size (Beta)	**N**	**Power**
0.2	120,286	14.5%
	400,000	37.1%
0.4	120,286	43.2%
	400,000	90.3%
0.6	120,286	76.5%
	400,000	99.8%
0.8	120,286	94.7%
	400,000	100.0%

Box S1. Details of the CNV scoring algorithm showing the truth dataset, metrics used, and the equation (**Eq. 1**) used to define the score

A truth dataset was defined in ~1000 Illumina Omni 2.5 individuals as those CNVs intersecting across 3 calling algorithms. And used 8 CNV call metrics to score:

Individual Level

- # CNVs / sample
- Waviness Factor
- BAF drift
- BAF SD & Mean
- LRR SD & Mean

CNV call level

- CNV length
- # of probes
- Confidence Score

$$QS_{cnv} = \frac{1}{1 + \exp(-(\beta_0 + \sum_{i=1}^{N} \beta_i V_i))} \qquad (1)$$

β = Residuals of logistic regression model for the above 8 metrics with ‘true’ or ‘false’ as the outcome variables.

V = Actual value in the UK Biobank call dataset

CHAPTER 6

Discussion and Future Work

6.1 Overview

Genome-wide association studies have given us the ability to identify many genetic markers associated with complex traits and disease risk. There are now over 10,000 known SNP-trait associations, giving us a powerful overview of the heritability of these traits [1]. Estimations of narrow-sense heritability (h^2) however, tell a different story as to how much of this known genetic variation explains the heritability of complex traits. We described a number of hypotheses that have been proposed to explain missing h^2 including the possibility that these estimates were over-estimated, or could be accounted for by smaller effect variation that would become apparent when using larger sample sizes. Another hypothesis was that rare and low frequency and/or structural forms of 'still-missing' variation contribute towards explaining this missing h^2. It has, and continues to be a challenging task to detect and characterise rare, low-frequency and structural variation. This thesis has presented some approaches to explaining missing h^2 by analysing and interpreting the additive effect of low frequency variants and structural variation on complex traits and disease risk. By analysing low-frequency and structural variation, we hoped to shed some light on this dark-matter and further pinpoint the most appropriate next steps in fully understanding the heritability of complex traits and common disease risk.

6.2 Low frequency - large effect variants detected using WGS may explain more of the variance explained for some complex traits – but only fractionally more than imputing the 1000 genomes reference panel into SNP chip data

Before the advent of whole genome sequencing, low frequency variants were difficult to detect and it has therefore been difficult to determine their contribution to complex trait effects and disease risk. In chapter 2 we sequenced 680 whole human genomes at 7X median read depth. Variants were called filtered and missing data were imputed missing using our extensive processing pipeline [2]. Variants were quality controlled by checking concordance with SNP chip and targeted high-depth sequencing data. We then carried out association analyses against 11,132 cis-eQTLs and 93 circulating biomarkers.

Of the 11.7 million variants detected, we found there to be 1,314 *cis*-eQTL associations at P<1x10^{-06} and 8 biomarker associations at P<8x10^{-10}. Seven percent of our detectable *cis*-gene expression traits (89 of 1,314) and 1/8 biomarker associations were low frequency and large effect. All nominal evidence of association was lost for 13 of 1,232 common variant associations when conditioning on the strongest low frequency variant suggesting that few common variants are driven by low frequency variants. Conversely, all nominal significance was lost for 11 of 90 low frequency variant associations when adjusting for the strongest common variant in the region, suggesting that ~12% of significant low frequency variation can be explained by common variation. Sixty two of the 90 low frequency variant associations (69%) detected using sequencing were found at P<0.0001 in 1000 genomes imputed data. Furthermore, associated variants were reduced to 33 (37%)

when based on the same statistical thresholds. These results illustrate that imputation can capture most low frequency variation, albeit less accurately.

Our data show that low frequency variants contribute to the heritability of complex traits. This study confirms that whole genome sequencing can identify low frequency variants not discovered by genotyping based approaches when sample sizes are sufficiently large to detect substantial numbers of common variant associations, and that common variant associations are rarely explained by single low frequency variants of large effect. These findings show that searching for low frequency variation associated with complex traits and disease risk is not a futile effort. The study highlights the fact there fewer low frequency variants associated with complex traits, but they do exist, however, it may be more resource and time effective to detect these using imputation reference panels rather than whole genome sequence data.

6.3 No evidence that a complex region encompassing the *AMY1* locus previously implicated as associated with obesity has an effect on BMI

Regions of the genome that harbour complex, multi-allelic copy number variants or repetitive regions confound genome-wide association studies. These regions are generally masked from SNP chip design due to their unpredictable nature, and are of a low quality when attempting to map reads in that region to the reference genome using next generation sequencing [3]. In chapter 3 we attempted to characterise one of these complex regions harbouring the *AMY1* and *AMY2* genes. A previous study found that the highly copy-number polymorphic *AMY1* gene that transcribes the salivary Amylase enzyme responsible for converting dietary starch into sugar exhibited an additive association with BMI with each additional copy decreasing risk of obesity with an odds ratio of 1.19 [4].

We characterised absolute copy number in 657 individuals in the *AMY* region using two computational tools that use read depth accounting for GC bias, repetitive regions and only loci that are triplicated to account for the three AMY1 homologs present in the reference genome. Genome STRiP then continues to classify calculated absolute copy number for each individual into a constrained Gaussian mixture model using expectation-maximization. When incrementing mrCaNaVaR calculated absolute copy number by one, we found that mrCaNaVaR and Genome STRiP were correlated when accounting for differences expected when not using population-level data. We found after carrying out ddPCR in 54 samples that copy number was almost perfectly concordant with the Genome STRiP copy number calls.

We found that copy number detected using both methods showed no evidence of nominal association with BMI in the 657 individuals even though the individuals' *FTO* variant and BMI polygenic score was significantly associated with BMI. The results were replicated in 2 additional European cohorts: GoT2D and the Estonian biobank. These results show that although complex structural variants can now be characterised, extreme care is needed when attempting to do so. Although it is important to understand the causes of missing heritability in complex traits such as obesity, it is also important to ensure that the methods used and the results obtained are rigorously tested. That being said, the *AMY* region is just one of many complex, multi-allelic structurally variable regions in the genome. More work is now needed to characterise these regions and attempt to understand their impact on complex traits and disease risk [5-8].

6.4 Large, rare pathogenic deletions and duplications detected using custom SNP chip intensity data in 120,286 individuals are highly pleiotropic and increase in pathogenicity with several exceptions

Structural variation with the most striking impact on complex traits and disease risk are large (≥50Kb), rare single-event deletions and duplications. Coe and colleagues [9] characterised a morbidity map of developmental delay that comprises of the majority of large deletions and duplications that are known to be pathogenic, but not full chromosomal/congenital abnormalities. In chapter 4 we wanted to characterise these known pathogenic copy number variants in 120,286 UK Biobank individuals because they were extremely rare or completely lacking prevalence in the original study that used ~20,000 controls. We wanted to find out if prevalence of controls changed in a 6-fold larger sample size. These variants generally have highly deleterious effects and if they were to be prevalent in control populations, they may be less penetrant in developmental delay and may account for some of the additive heritability in complex traits.

We found that most of the deletion and duplication associations with developmental delay were strengthened when using the UK Biobank as a control population. This is reflected by the lower prevalence, or lack of CNVs in the control population strengthening their pathogenic role in developmental delay. We found four notable exceptions, one deletion and three duplications that are no longer significantly associated with developmental delay. These consisted of a 320Kb deletion at 3p11.2, a 1.7 Mb duplication at 2q13, a 290 Kb duplication at 15q11.2 and a 790 Kb duplication at 16p13.11. The deletion at 3p11.2 was newly discovered in the previous study using the ~20,000 controls so additional replication with large numbers may be required to ascertain its prevalence in control populations. The three duplications are previously discovered associations and also have less association with surrogate continuous traits. These results indicate that duplications are less likely to be associated with developmental delay, and also have a less noticeable effect in complex traits than deletions in general. These data may also suggest that deletions many have more of a residual polygenic effect than duplications and the duplications that are pathogenic are less frequent, but less pleiotropic. This study was limited by the differences in technology between cases and controls with the developmental delay case CNVs being detected using signature array CGH methods that may have different biases and levels of resolution compared to SNP chip.

6.5 A genome-wide scan of large, rare deletions and duplications reveals known CNVs to be associated with many continuous traits

In chapter 5 we explored the extent of large, rare duplications and deletions throughout the genome. Duplications and deletions larger than 50Kb are very rare in the general population. We wanted to harness the large sample size of UK Biobank to detect potentially novel CNVs associated with complex traits that may have been missed in smaller studies. We accounted for false positives by applying a novel CNV quality score to each deletion and duplication. The quality score was a function of population and individual-level metrics weighted for presence within a truth dataset that was based on CNVs called using 3 separate calling methods. Using this approach, we were able to scan the UK Biobank for known and novel copy number variants without needing to visually score the levels of Log R Ratio and B Allele frequency in each CNV call.

We found two deletions and two duplications to be genome-wide significantly associated with at least one of the 204 complex traits tested. The 16p11.2 deletion was associated with 18 traits including height, BMI, lung function, social economic status. The deletion is highly pleiotropic at genome-wide significance whereas the duplication was not genome-wide significant for any traits, but nominally significant for 46 of the traits tested. The results for 16p11.2 indicated that duplications and deletions at the same locus have different effect sizes on the phenotype. The 16p12.1 deletion is genome-wide associated with number of hours of television watched per day and nominally associated with increased sedentary lifestyle and decreased lung function, intelligence and longevity. We also found a genome-wide association between the 17p12 duplication and hand grip strength, a duplication known to be associated with Charcot-Marie neuropathy. Finally, we found a genome-wide association between the 15q13.2-13.3 duplication and a decreased vehicle ownership with further nominal negative associations with social economic status, bone mineral density and red blood cell count distribution width.

Our results show that large-effect, copy number variants encompassing multiple genes that are too rare to be in linkage disequilibrium with common SNPs variants are highly pleiotropic, but a small proportion (4 out of 2,257 deletions and duplications tested) are genome-wide associated with the complex traits tested. The study was limited however, in the following ways: (a) array design – SNP probes are less abundant in structurally complex regions, (b) resolving and classifying CNVs with heterogeneous breakpoints is a complex task and (c) even larger sample sizes being required for studies involving very rare variants such as CNVs in order to detect association. Larger and more comprehensive study designs, along with more sophisticated approaches to resolving breakpoint heterogeneity will alleviate these limitations in future studies.

6.6 Missing heritability explained by low frequency and structural variation or overestimated?

We have shown that there are both low frequency and structural variant signals associated with complex traits and disease risk. These findings by no means explain the full extent of missing heritability, and there are many avenues of research left to explore. There are now more accurate methods for estimating narrow sense heritability (h^2) in complex traits. New methods such as restricted maximum likelihood (REML) [10] and another method that uses a Monte-Carlo algorithm based around it [11] have emerged that account for many small effects from all genetic variants. It is important to continue to explore both larger sample sizes and also structural and low frequency variants much further, but to also us these results to recalibrate our estimations of h^2 with ‘gold standard’ datasets with the ultimate goal of these paradigms converging and h^2 being fully explained.

6.7 Future directions

As previously mentioned, the findings from this thesis are the tip of the iceberg when exploring the full extent of complex trait genetics. There is a whole spectrum of rare, low frequency and structural variation yet to be explored. These areas will need to be adequately studied before we can come to any firm quantifications of the presence of additional heritability over what we've already found using GWAS. Once h^2 has been fully explained, the field must also understand the full extent of broad-sense heritability (H^2) and its implications. Genotyping increasingly larger sample sizes and developing more refined and comprehensive imputation reference panels that include rare, low-frequency and structural variation including structurally complex and repetitive regions will be of much benefit to complex trait and common disease genetics. Below we explain some potential directions leading on from this thesis.

6.7.1 More complex, accurate and comprehensive reference panels

The 1000 genomes reference panel as proved to be an invaluable resource for discovering new signals associated with complex traits [12,13] and includes a large amount of information on both low-frequency, rare and structural variation. The next generation of reference panel is now coming into fruition with projects such as the Haplotype Reference Consortium (refer to http://www.haplotype-reference-consortium.org/) which includes whole genome sequence data from over 38,000 individuals in 20 cohorts from around the world, including 1000 genomes. Imputing massive reference panels such as these into existing studies will help us to potentially uncover more signals associated of complex traits and disease risk and help us gain a further estimation of their heritability using emerging methods such as REML.

6.7.2 Larger sample sizes

Larger sample sizes will allow us to more fully understand both hiding heritability and also impute more of the rare variation from these more sophisticated reference panels into larger sample sizes, increasing their confidence. We will have power to detect those rarer variants that have smaller effects. The UK Biobank will be releasing its second phase of genetic data in late 2016, this will prove to be a huge leap forward when combined with more comprehensive reference panels. It could indeed be said that larger sample sizes and more detailed reference panels have a symbiotic relationship when it comes to drawing a more complete picture of the heritable components of complex traits and disease risk.

6.7.3 Droplet digital PCR and whole genome sequencing to understand structurally complex regions

Much work is being dedicated to understanding the structurally complex regions of the genome from a variety of different angles. Some successful work is now beginning to surface showing that complex structural variants do play a role in complex traits and disease risk [6,7]. These methods use droplet digital PCR (ddPCR) and whole-genome sequencing as their primary methods to accurately characterise these complex regions. We explored these methods in chapter 3, but it has been found that when these complex structural variants are accurately defined in hundreds of individuals, each of their structural configurations can be in linkage disequilibrium with groups of known SNPs. In other words, large numbers of common and low frequency SNPs can, when compounded, allow us to accurately infer the structural haplotypes of a complex region (**Fig. 1**). Having these regions accurately characterised in reference genomes will allow them to be routinely included in future GWAS. These variants could shed light on causality/aetiology – as these complex regions may intersect with candidate disease regions. These structural variants can also create a new structural haplotype not seen when looking at individual SNPs.

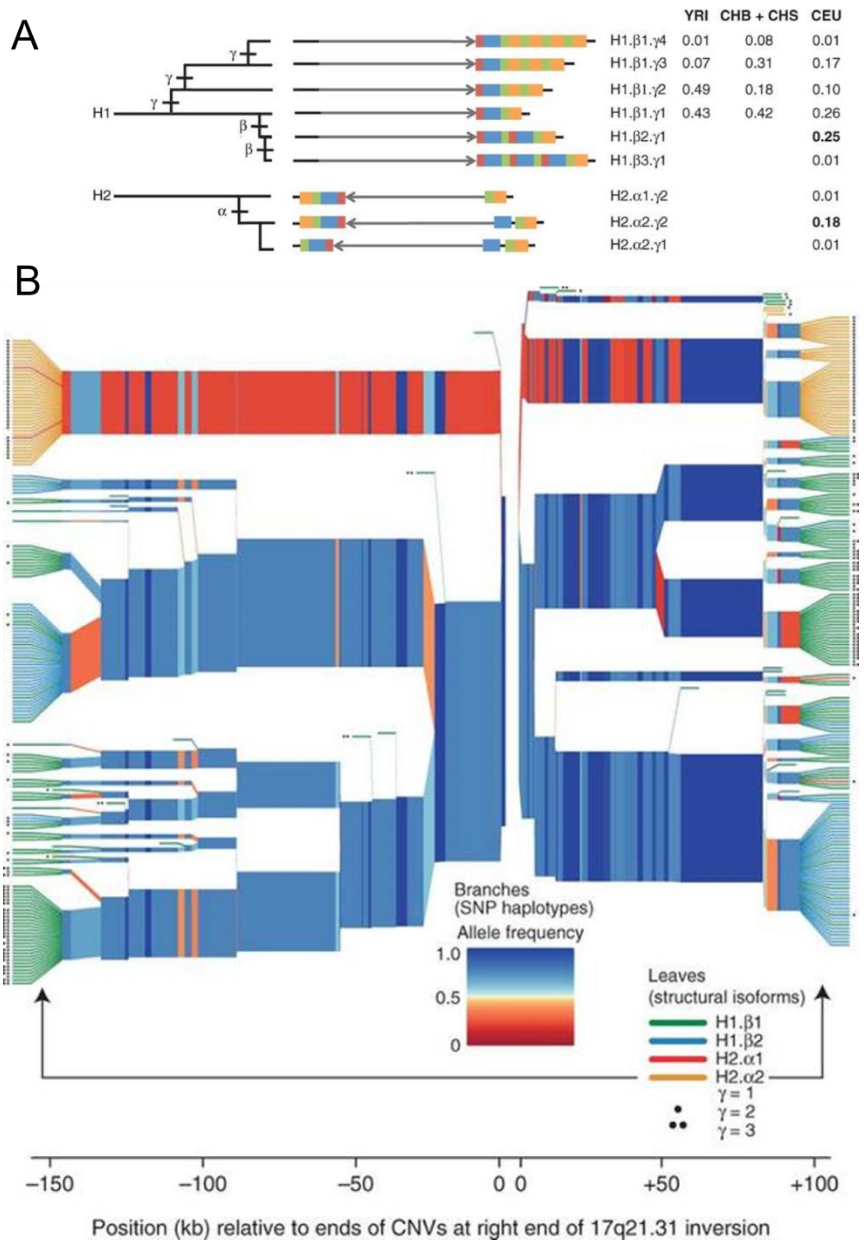


Figure 1. The 17q21.31 locus is an example of a structurally complex region. **(A)** Shows the nine most common structural forms of this locus with its frequency in HapMap 3. **(B)** Represents the haplotypes of the structural forms of this locus reconstructed using SNPs distal to the structural variant. The branches represent allele frequency of SNPs relative to the structural variant shown on the x-axis showing how many neighbouring SNPs can be used to infer haplotypes of complex structural regions. Taken from Boettger et al. 2012.

6.7.4 Single molecule read technology – widening the goal posts?

Much of the human reference genome is highly repetitive or contains gaps due to complex regions, rendering it difficult or impossible to align short 100-250bp reads in these regions. Single-molecule read technology (SMRT) such as the PacBio Sequel System will allow us to analyse repetitive and structural regions and their architecture because of their extremely long sequencing reads that span upwards of 10Kb. Studies of these complex structural and repetitive regions require very accurate data for them to be useful in imputation. The quality of the sequencing is still evolving, but work is already in progress to develop a new gold standard human reference genome that accounts for complex structural & difficult to decipher repetitive regions that comprise so much of the genome [14]. This approach could in theory open a new frontier research in complex trait genetics, with imputation reference panels based on these technologies, but is still very much science fiction for the near future.

6.8 Conclusion

In this thesis I have shown how understanding low-frequency and structural genetic variation will help us understand more about the heritability of complex traits and disease risk. These approaches make for a promising and exciting future for the field of human genetics. A great deal of care needs to be taken however, in how we go about moving forward in these areas of the field with respect to both methodology and study design. More work needs to be done towards these existing a reference panel that reflects accurate structural variation, repetitive regions and rare/low frequency variation, along with GWAS with very large sample sizes to resolve the complexity of the heritable component of complex traits and disease risk.

6.9 References

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42: D1001-1006.
2. Wood AR, Tuke MA, Nalls M, Hernandez D, Gibbs JR, et al. (2015) Whole-genome sequencing to understand the genetic architecture of common gene expression and biomarker phenotypes. Hum Mol Genet 24: 1504-1512.
3. Sudmant PH (2010) Diversity of human copy number variation and multicopy genes. Science 330: 641-646.
4. Falchi M (2014) Low copy number of the salivary amylase gene predisposes to obesity. Nat Genet 46: 492-497.
5. Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, et al. (2015) Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. Nat Genet 47: 921-925.
6. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, et al. (2016) Schizophrenia risk from complex variation of complement component 4. Nature 530: 177-183.
7. Boettger LM, Salem RM, Handsaker RE, Peloso GM, Kathiresan S, et al. (2016) Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat Genet 48: 359-366.
8. Boettger LM, Handsaker RE, Zody MC, McCarroll SA (2012) Structural haplotypes and recent evolution of the human 17q21.31 region. Nat Genet 44: 881-885.

9. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BW, Vulto-van Silfhout AT, et al. (2014) Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet 46: 1063-1071.

10. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88: 76-82.

11. Loh PR, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, et al. (2015) Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet 47: 1385-1392.

12. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

13. Wood AR, Perry JR, Tanaka T, Hernandez DG, Zheng HF, et al. (2013) Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation. PLoS One 8: e64343.

14. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, et al. (2015) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517: 608-611.

APPENDIX I

Source code used to bulk quantile normalise raw probeset intensities in 152,729 UK Biobank individuals

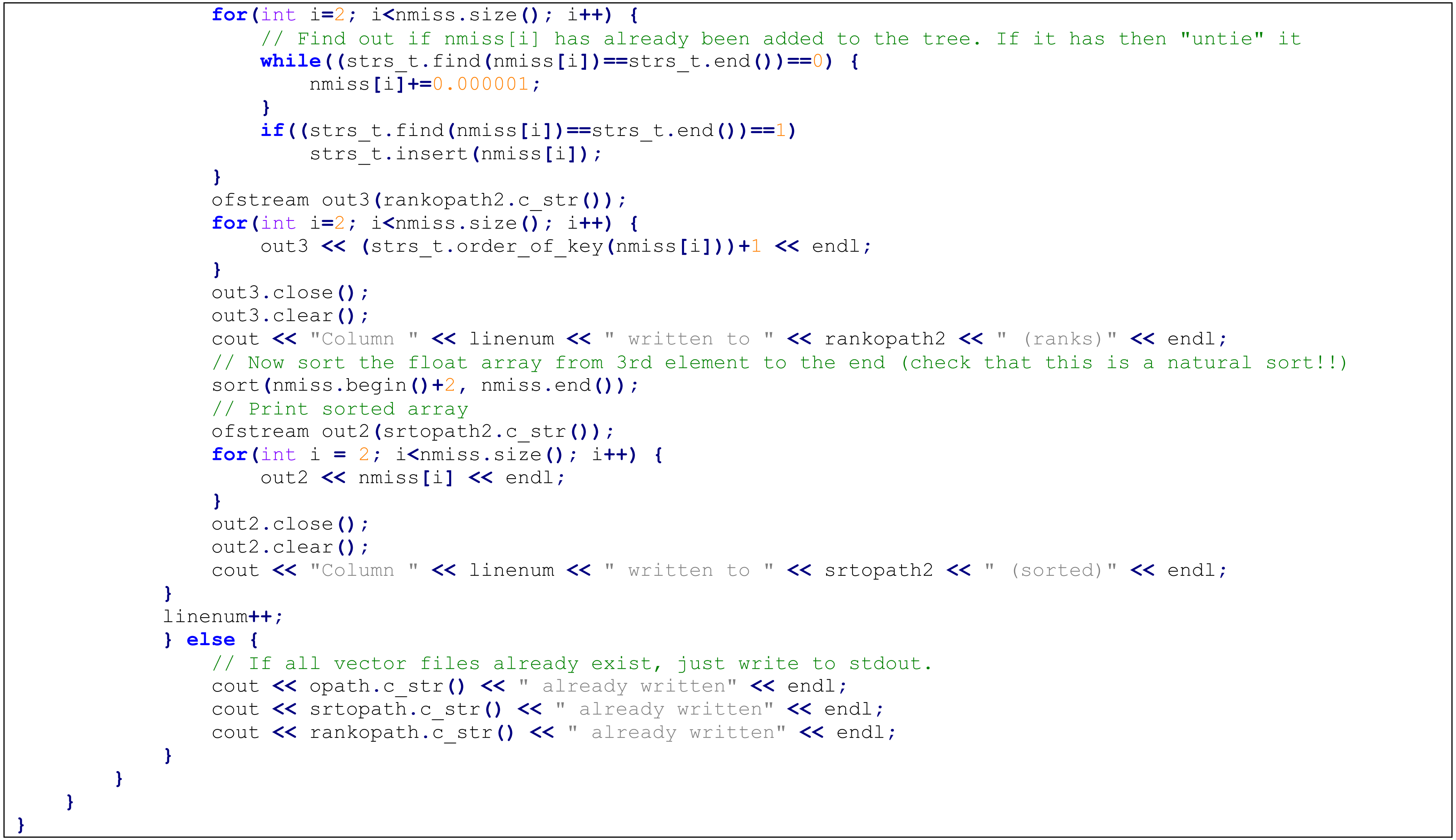
Full source code, binaries, example workflows and README file is accessible from:

http://goo.gl/iKm196

Box 1. quantileNorm1 sorts and ranks each individual’s raw probeset for processing by quantileNorm2 (**Box 2**)

```
/* Author: Marcus Tuke.
 * Date: 26th Oct 2015
 * Generate input vector temp files for running quantileNorm2 (raw, sorted & rank intensity vector files for each
sample)
 * Usage: ./quantileNorm1 <UK Biobank intensity file> <path to tmp data>
 */

#include <thread>
#include <iostream>
#include <sstream>
#include <fstream>
#include <string>
#include <vector>
#include <algorithm>
#include <boost/algorithm/string.hpp>
#include <unordered_map>
#include <cassert>
#include <ext/pb_ds/assoc_container.hpp>
#include <ext/pb_ds/tree_policy.hpp>
#include <iomanip>

// Define namespaces

using namespace std;
using namespace __gnu_pbds;
using namespace boost;

// Define Order Statistic Tree - data structure to store intensity rankings

typedef
    tree<
    long double,
    null_type,
    less<long double>,
    rb_tree_tag,
    tree_order_statistics_node_update>
set_t;
```

```
// Execute main code importing arguments: [ UKBB_file , temp_data_path ]

int main (int argc, char *argv[]) {
    cout << setprecision(12);
    cout << "Input files are:" << endl << endl;
    for(int arg_list = 1; arg_list < argc; arg_list++) {
        cout << argv[arg_list] << endl;
    }
    cout << endl;
    for(int arg_list = 1; arg_list < argc; arg_list++) {
    ifstream infile(argv[arg_list]);
    string line;
    int linenum=1;
    while (getline(infile, line)) {
        // Create an order statistic tree so value ranks can be detected more quickly
        set_t strs_t;
        vector<string> strs;
        split(strs, line, is_any_of(" "));
        const int NMISS_SIZE=strs.size();
        vector<long double> nmiss(NMISS_SIZE);
        long double median=0;
        if(strs[0]!="Person") {
            // Loop split array and get N that are not lines 1 or 2 and are not == NA
            // Get size of float-only array
            int num_float=0;
            for(int i = 2; i<strs.size(); i++) {
                if(strs[i]!="NA") num_float++;
            }
            const int FLOAT_ARRAY_SIZE=num_float;
            vector<long double> float_array(FLOAT_ARRAY_SIZE);
            // Now create a float array to get median value of non-missing data
            num_float=0;
            for(int i = 2; i<strs.size(); i++) {
                if(strs[i]!="NA") {
                    float_array[num_float]=stold(strs[i]);
                    num_float++;
                }
            }
            // Sort the float array numerically
            sort(float_array.begin(), float_array.end());
            median=float_array[ceil(float_array.size()/2)];
            //Create float array (nmiss) where first 2 values are zero (subject and sex) and NAs converted to
medians
            nmiss[0]=0.00;
            nmiss[1]=0.00;
            for(int i = 2; i<strs.size(); i++) {
                if(strs[i]=="NA") {
                   nmiss[i]=median;
                } else {
                   nmiss[i]=stold(strs[i]);
                }
            }
            // If first line, these are the probe IDs, so create one standard file to contain these first
            // Declare header/unsorted file output stream
        }
        // Assign the two arguments
        string argv1=argv[1];
        string path=argv[2];
        // Set temp file prefix
        string arg_str = argv1.substr(0, argv1.size()-4);
        // Set raw vector temp file
        stringstream outstring;
        outstring << path << "." << arg_str << "." << strs[0] << ".temp";
        string opath = outstring.str();
        // Set sorted vector temp file
        stringstream outstringsrt;
        outstringsrt << path << "." << arg_str << "." << strs[0] << ".sorted.temp";
        string srtopath = outstringsrt.str();
        // Set rank vector temp file
        stringstream outstringrank;
        outstringrank << path << "." << arg_str << "." << strs[0] << ".ranks.temp";
        string rankopath = outstringrank.str();
        // Open input test streams for each file and if any do not exist then continue
        ifstream inout(opath.c_str());
        ifstream inout2(srtopath.c_str());
        ifstream inout3(rankopath.c_str());
        if((inout.good()==0 || inout2.good()==0 || inout3.good()==0 || inout.peek() ==
std::ifstream::traits_type::eof() || inout2.peek() == std::ifstream::traits_type::eof() || inout3.peek() ==
std::ifstream::traits_type::eof())) {
            // Close test streams
            inout.close();
            inout2.close();
            inout3.close();
            // Create output streams for raw, sorted & rank vector files
            stringstream outstring2;
            outstring2 << path << "." << arg_str << "." << strs[0] << ".temp";
            string opath2 = outstring2.str();
            stringstream outstringsrt2;
            outstringsrt2 << path << "." << arg_str << "." << strs[0] << ".sorted.temp";
            string srtopath2 = outstringsrt2.str();
            stringstream outstringrank2;
            outstringrank2 << path << "." << arg_str << "." << strs[0] << ".ranks.temp";
            string rankopath2 = outstringrank2.str();
            // First line of first file is the sample list
            if(strs[0]=="Person" && arg_list==1) {
                ofstream out(opath2.c_str());
                // Populate raw vector for sample list & then close file
                for(int i = 2; i<strs.size(); i++) {
                    out << strs[i] << endl;
                }
                out.close();
                out.clear();
                // Print confirmation to stdout
                cout << "Column " << linenum << " written to " << opath2 << " (header)" << endl;
            } else {
                // Print unsorted values to file (strs)
                ofstream out(opath2.c_str());
                for(int i = 2; i<strs.size(); i++) {
                    out << strs[i] << endl;
                }
                out.close();
                out.clear();
                cout << "Column " << linenum << " written to " << opath2 << " (unsorted)" << endl;
                // Print ranking file go through each unsorted line and find value in sorted, when found print to
rank file and break
                int itm = strs_t.order_of_key(median);
                // Loop to populate the tree
```



```
        for(int i=2; i<nmiss.size(); i++) {
          // Find out if nmiss[i] has already been added to the tree. If it has then "untie" it
          while((strs_t.find(nmiss[i])==strs_t.end())==0) {
            nmiss[i]+=0.000001;
          }
          if((strs_t.find(nmiss[i])==strs_t.end())==1)
            strs_t.insert(nmiss[i]);
        }
        ofstream out3(rankopath2.c_str());
        for(int i=2; i<nmiss.size(); i++) {
          out3 << (strs_t.order_of_key(nmiss[i]))+1 << endl;
        }
        out3.close();
        out3.clear();
        cout << "Column " << linenum << " written to " << rankopath2 << " (ranks)" << endl;
        // Now sort the float array from 3rd element to the end (check that this is a natural sort!!)
        sort(nmiss.begin()+2, nmiss.end());
        // Print sorted array
        ofstream out2(srtopath2.c_str());
        for(int i = 2; i<nmiss.size(); i++) {
          out2 << nmiss[i] << endl;
        }
        out2.close();
        out2.clear();
        cout << "Column " << linenum << " written to " << srtopath2 << " (sorted)" << endl;
      }
      linenum++;
      } else {
        // If all vector files already exist, just write to stdout.
        cout << opath.c_str() << " already written" << endl;
        cout << srtopath.c_str() << " already written" << endl;
        cout << rankopath.c_str() << " already written" << endl;
      }
    }
  }
}
```

Box 2. quantileNorm2 creates quantile normalised output data from ranked and sorted data

```
/* Author: Marcus Tuke.
 * Date: 26th Oct 2015
 * Usage: ./quantileNorm2 <UK Biobank intensity file> <path to tmp data>
 *
 * Process output and generate quantile normalised PennCNV-Affy input using Unix paste command. For example:
 *
 * # Create paste command
 * $ awk 'BEGIN {print "paste -d \0\4\7\t\0\4\7"} NR==1 {print "<PATH>/.test."$1".temp"} NR>1 {print 
"<PATH>/.test."$1".quant.temp"}' batch.txt | tr '\n' ' ' > tmp.sh
 *
 * # Join paste output with batch header
 * $ { tr '\n' '\t' < batch.txt | sed 's/$/\n/g' & sh tmp.sh; } > quantile_normalised_intensities.txt
 *
 * $ rm tmp.sh
 *
 */

#include <math.h>
#include <omp.h>
#include <thread>
#include <iostream>
#include <sstream>
#include <fstream>
#include <string>
#include <vector>
#include <algorithm>
#include <boost/algorithm/string.hpp>
#include <unordered_map>
#include <cassert>
#include <iomanip>

// Define namespaces

using namespace std;
using namespace boost;

// Execute main code importing arguments: [ UKBB_file , Genotyping_batch_ID, temp_data_path ]
```

```
int main (int argc, char *argv[]) {
    cout << setprecision(12);
    cout << "Input files are:" << endl << endl;
    string path=argv[3];
    string argv1=argv[1];
    string arg_str = argv1.substr(0, argv1.size()-4);
    for(int arg_list = 1; arg_list < argc; arg_list++) {
        cout << argv[arg_list] << endl;
    }
    cout << endl;
    int arg_list;

    for(arg_list = 1; arg_list < 2; arg_list++) {
        int linenum=1;
        ifstream infile(argv[arg_list]);
        string line, ln;
        // Output totals file:
        stringstream totstring;
        //outstringrank << path << "." << arg_str << "." << strs[0] << ".ranks.temp";
        totstring << path << "." << arg_str << "." << argv[2] << ".totals.temp";
        string totl = totstring.str();
        // Read in 'Person' column
        getline(infile, line);
        // Read in each line of samples list (after header)
        int z;
        while(getline(infile, line)) {
            // Loop here to read in totals file
            stringstream curstring;
            string cur2;
            curstring << path << "." << arg_str << "." << line << ".sorted.temp";
            // Read in sorted file twice. Once to count elements (1st loop) and once to populate array
            string cur = curstring.str();
            ifstream curfile(cur.c_str());
            // Set a group of empty streams for below 'if' blocks. These blocks are created to account for
occasional unreadable temp files due to corrupt disk blocks on NFS. Create a file postfixed with "2" of this is the
case.
            ifstream curfile0;
            ifstream curfile1;
```

```cpp
ifstream curfile2;
ifstream curfile12;
if(curfile.good()==0  || curfile.peek() == std::ifstream::traits_type::eof()) {
    curstring << "2";
    cur2 = curstring.str();
    curfile1.open(cur2.c_str());
    curfile12.open(cur2.c_str());
} else {
    curfile0.open(cur.c_str());
    curfile2.open(cur.c_str());
}
z=0;
// Determine vector size z depending on file type
if(curfile.good()==0  || curfile.peek() == std::ifstream::traits_type::eof()) {
    while (getline(curfile12, ln)) {
        z++;
    }
} else {
    while (getline(curfile2, ln)) {
        z++;
    }
}
curfile2.close();
// Allocate two vectors in memory: one for current sorted array, and one for running total
vector<long double> curarr(z);
vector<long double> totarr(z);
int y=0;
// Assign sorted data values to current vector 'currarr'
if(curfile.good()==0 || curfile.peek() == std::ifstream::traits_type::eof()) {
    while (getline(curfile1, ln)) {
        curarr[y]=stold(ln);
        y++;
    }
} else {
    while (getline(curfile0, ln)) {
        curarr[y]=stold(ln);
        y++;
    }
}
```

```
        y=0;
        // Read in totals file and assign sorted file to 'totals' array.
        ifstream totfile(totl.c_str());
        while (getline(totfile, ln)) {
            totarr[y]=stold(ln);
            y++;
        }
        totfile.close();
        // Sum current vector with total vector and re-write total vector file
        ofstream tot_out(totl.c_str());
        for(int j=0; j<totarr.size(); j++) {
            if(linenum==1) {
                tot_out << curarr[j] << endl;
            } else {
                tot_out << totarr[j]+curarr[j] << endl;
            }
        }
        tot_out.close();
        linenum++;
    }

    // Create 'averages' vector file
    stringstream avestring;
    avestring << path << "." << arg_str << "." << argv[2] << ".averages.temp";
    string ave = avestring.str();
    ofstream avefile(ave.c_str());
    // Allocate averages vector in memory
    vector<long double> averages(z);
    int a=0;
    string total;
    ifstream totfile(totl.c_str());
    // Calculate log2(mean) of total vector
    while(getline(totfile, total)) {
        long double tot = stold(total);
        avefile << log2(tot/linenum) << endl;
        averages[a]=log2(tot/linenum);
        a++;
    }
    // Loop through sample vectors again to allocate each log2(mean) to its equivalent rankings in the original
```

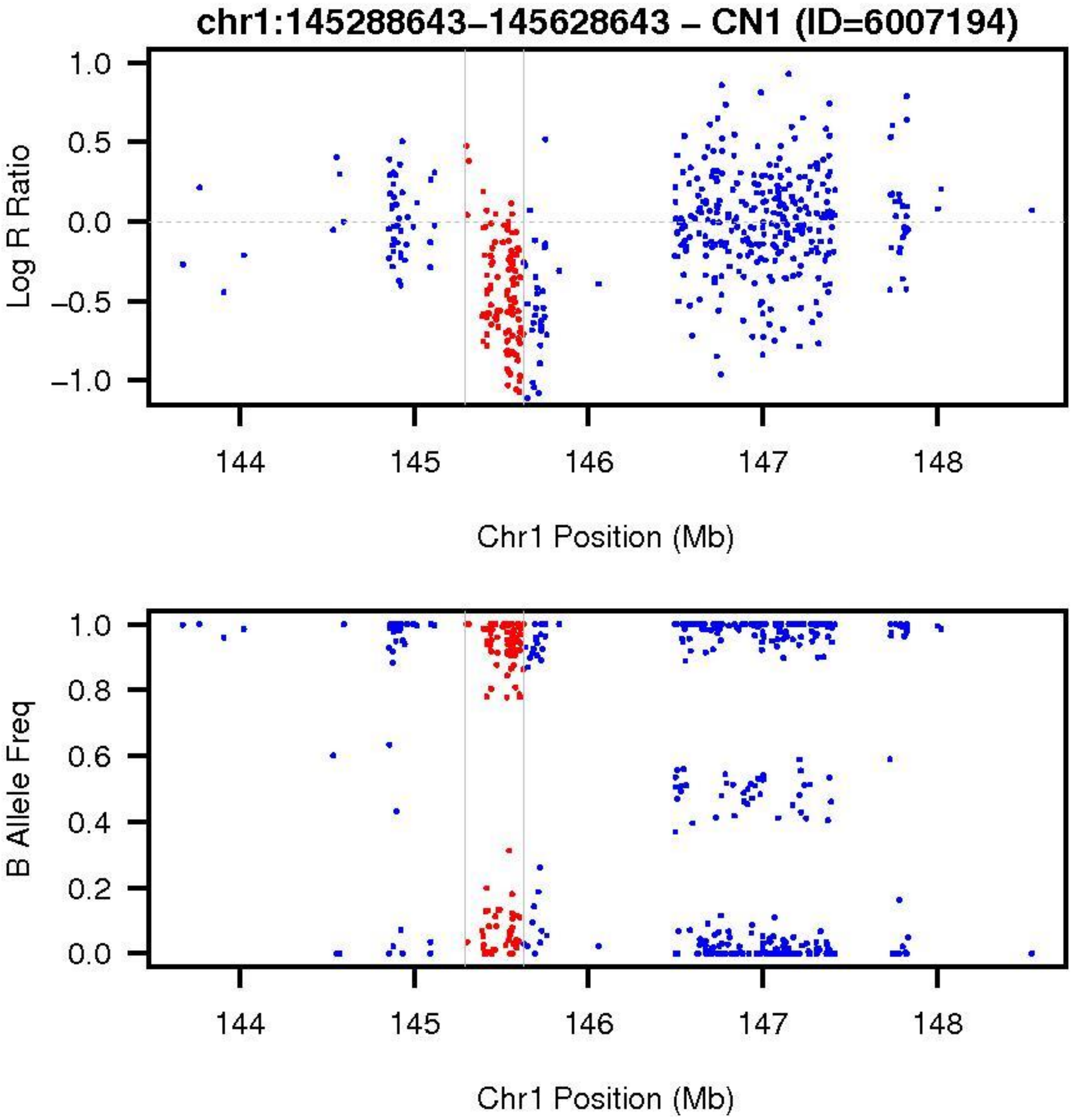
```
file
        // These are the quantile normalised values 'quant' for short
        ifstream infile2(argv[arg_list]);
        // Skip header line
        getline(infile2, line);
        while (getline(infile2, line)) {
            // Read raw (curstring), ranks (ranstring). Open 'quant' file to write to
            stringstream curstring;
            stringstream ranstring;
            stringstream quantstring;
            curstring << path << "." << arg_str << "." << line << ".temp";
            ranstring << path << "." << arg_str << "." << line << ".ranks.temp";
            quantstring << path << "." << arg_str << "." << line << ".quant.temp";
            string cur = curstring.str();
            string quant = quantstring.str();
            string ran = ranstring.str();
            //
            ifstream quantfile1(quant.c_str());
            string quanttest;
            getline(quantfile1, quanttest);
            // Open raw vector file
            ifstream curfile3;
            ifstream curfile4;
            ifstream curfile41;
            curfile3.open(cur.c_str());
            if(curfile3.good()==0 || curfile3.peek() == std::ifstream::traits_type::eof()) {
                curstring << "2";
                string cur2 = curstring.str();
               curfile41.open(cur2.c_str());
            } else {
               curfile4.open(cur.c_str());
            }
            // Open rank vector file
            ifstream ranfile;
            ifstream ranfile1;
            ifstream ranfile3;
            ranfile3.open(ran.c_str());
            if(ranfile3.good()==0 || ranfile3.peek() == std::ifstream::traits_type::eof()) {
                ranstring << "2";
                string ran2 = ranstring.str();
                ranfile1.open(ran2.c_str());
            } else {
                ranfile.open(ran.c_str());
            }
            ofstream quantfile(quant.c_str());
            // Allocate vector for 'ranks' file
            vector<int> ranks(z);
            a=0;
            // Put rank file contents into array
            string rank;
            if(ranfile3.good()==0 || ranfile3.peek() == std::ifstream::traits_type::eof())  {
                while(getline(ranfile1, rank)) {
                    ranks[a] = stoi(rank);
                    a++;
                }
            } else {
                while(getline(ranfile, rank)) {
                    ranks[a] = stoi(rank);
                    a++;
                }
            }
            // Now loop through the raw vector file and print average using rank as index
            string curln;
            a=0;
            int rankid;
            // 'averages' file is in an array. Loop through 'raw' intensity file. In each loop, load 'ranks' into
array.
            if(curfile3.good()==0 || curfile3.peek() == std::ifstream::traits_type::eof()) {
                while(getline(curfile41, curln)) {
                    if(curln=="NA") {
                        // Keep missing data points as missing
                        quantfile << "NA" << endl;
                    } else {
                        // Add average value indexed by rank item
                        rankid = ranks[a];
                        quantfile << averages[rankid-1] << endl;
                    }
                    a++;
                }
            } else {
                while(getline(curfile4, curln)) {
                    if(curln=="NA") {
                        // Keep missing data points as missing
                        quantfile << "NA" << endl;
                    } else {
                        // Add average value indexed by rank item
                        rankid = ranks[a];
                        quantfile << averages[rankid-1] << endl;
                    }
                    a++;
                }
            }
        // Print 'processed' message to stdout
        cout << "Quant file " << quant.c_str() << " processed." << endl;
        }
    }
}
```

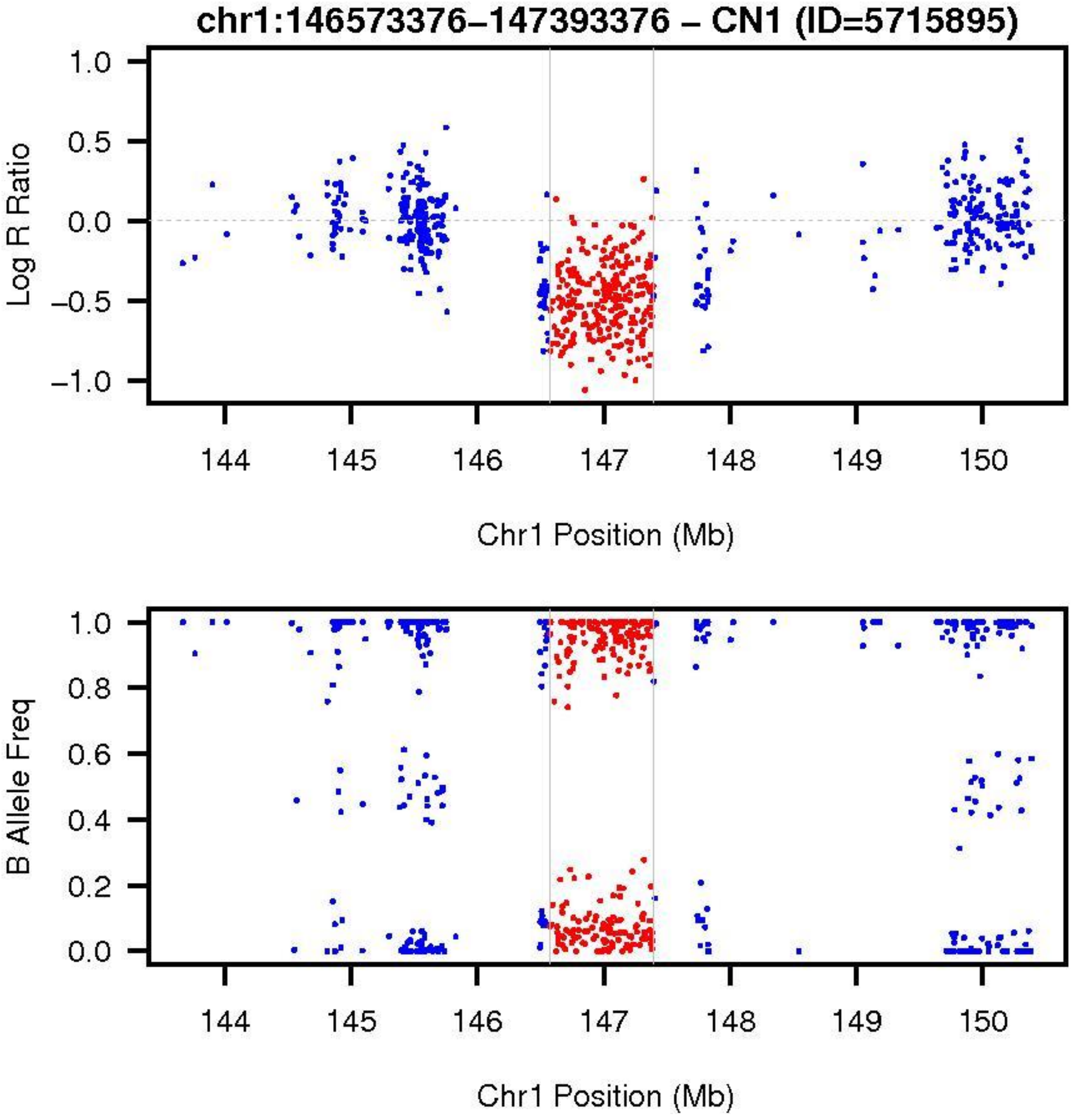
APPENDIX II

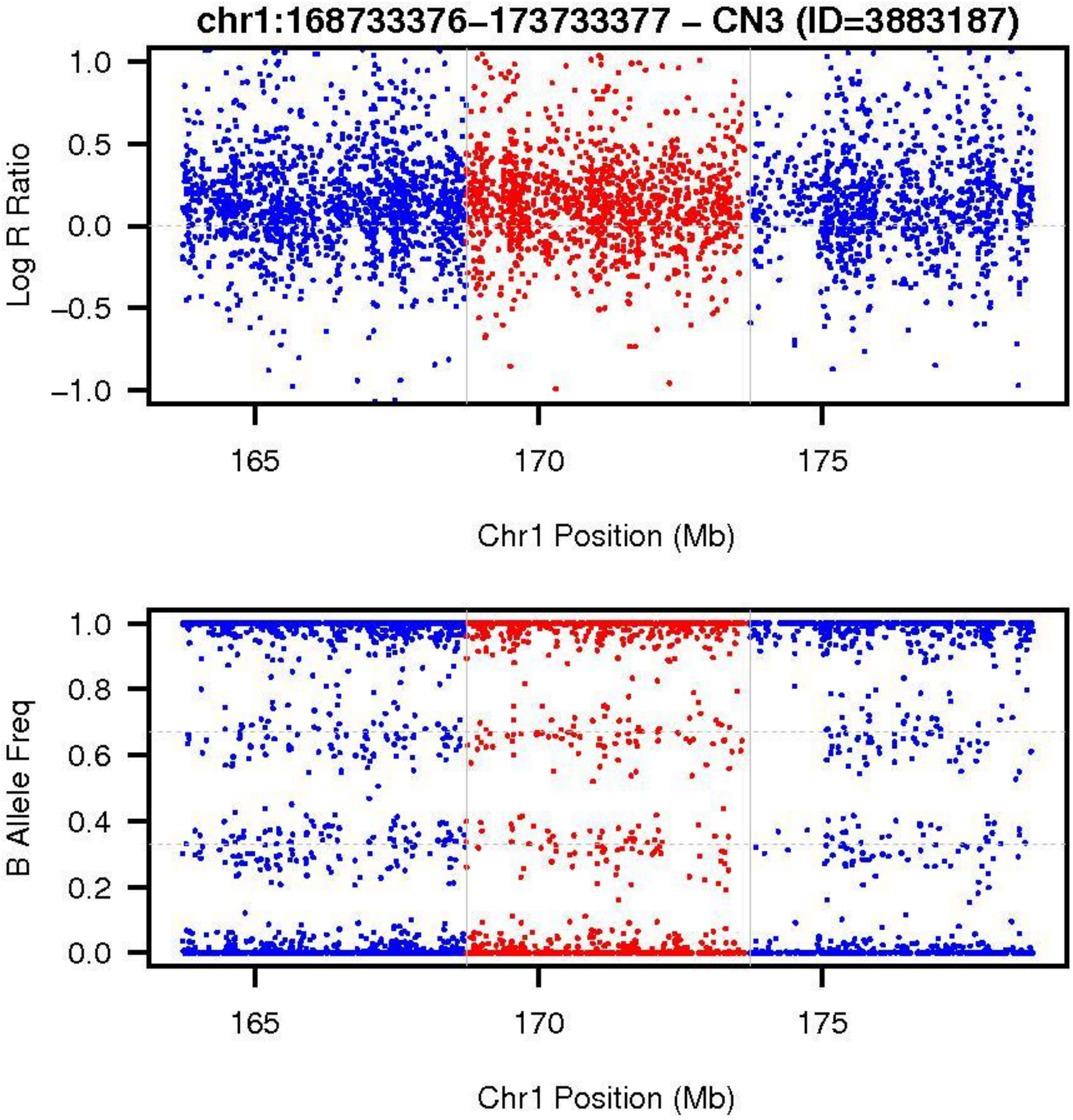
Exemplar probe intensity Log R Ratio (LRR) and B Allele Frequency (BAF) plots for the 49 unique deletion (n=25) and duplication (n=24) copy number variants that are present in greater than or equal to one individual in the UK Biobank

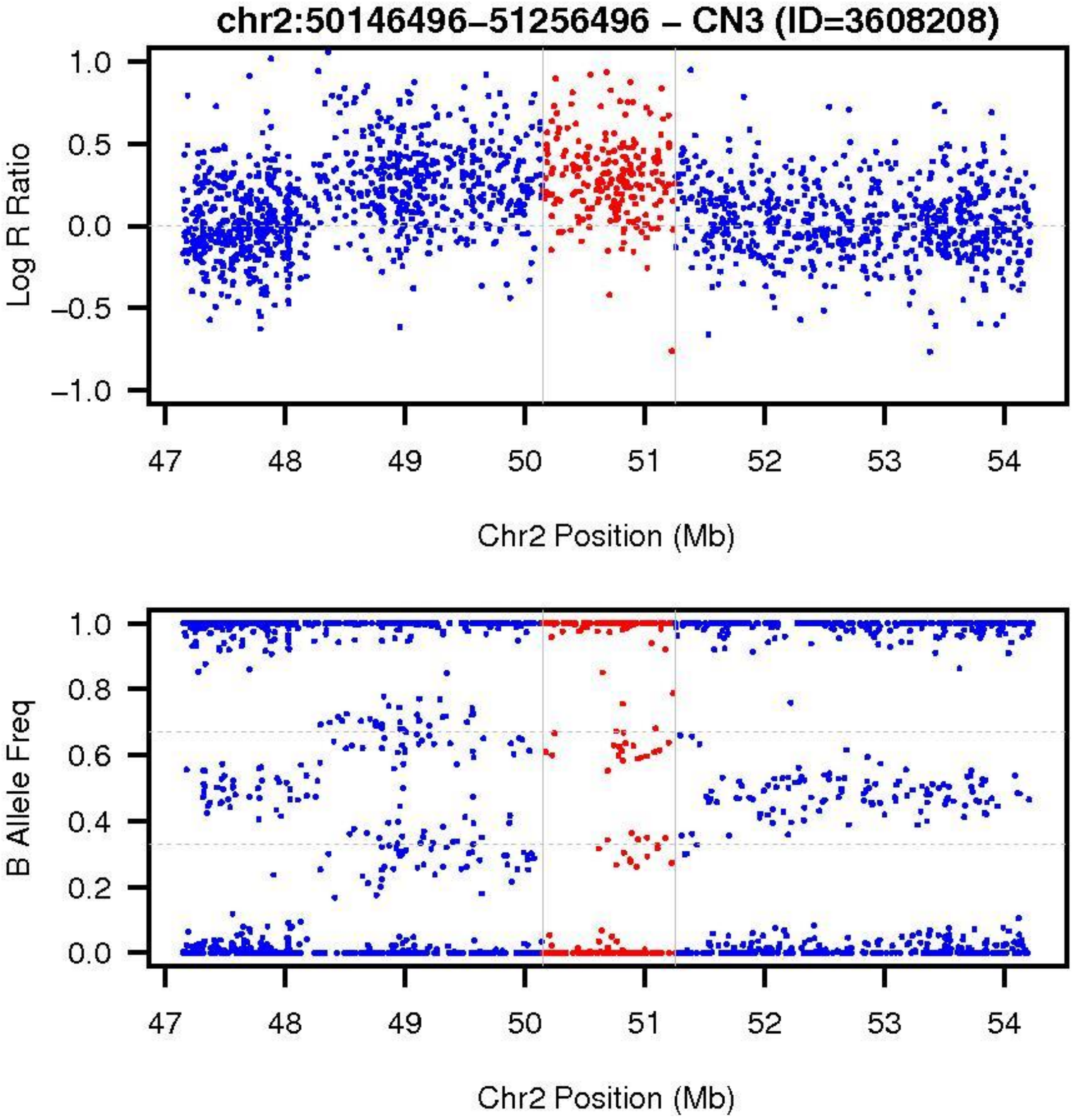
A full set of 2,168 LRR/BAF plots for all deletion (n=937) and duplication (n=1,231) calls plus 2,223 calls that were rejected from visual inspection (**Section 4.3.7**) is accessible from: https://goo.gl/X1J0Zv (491 MB)

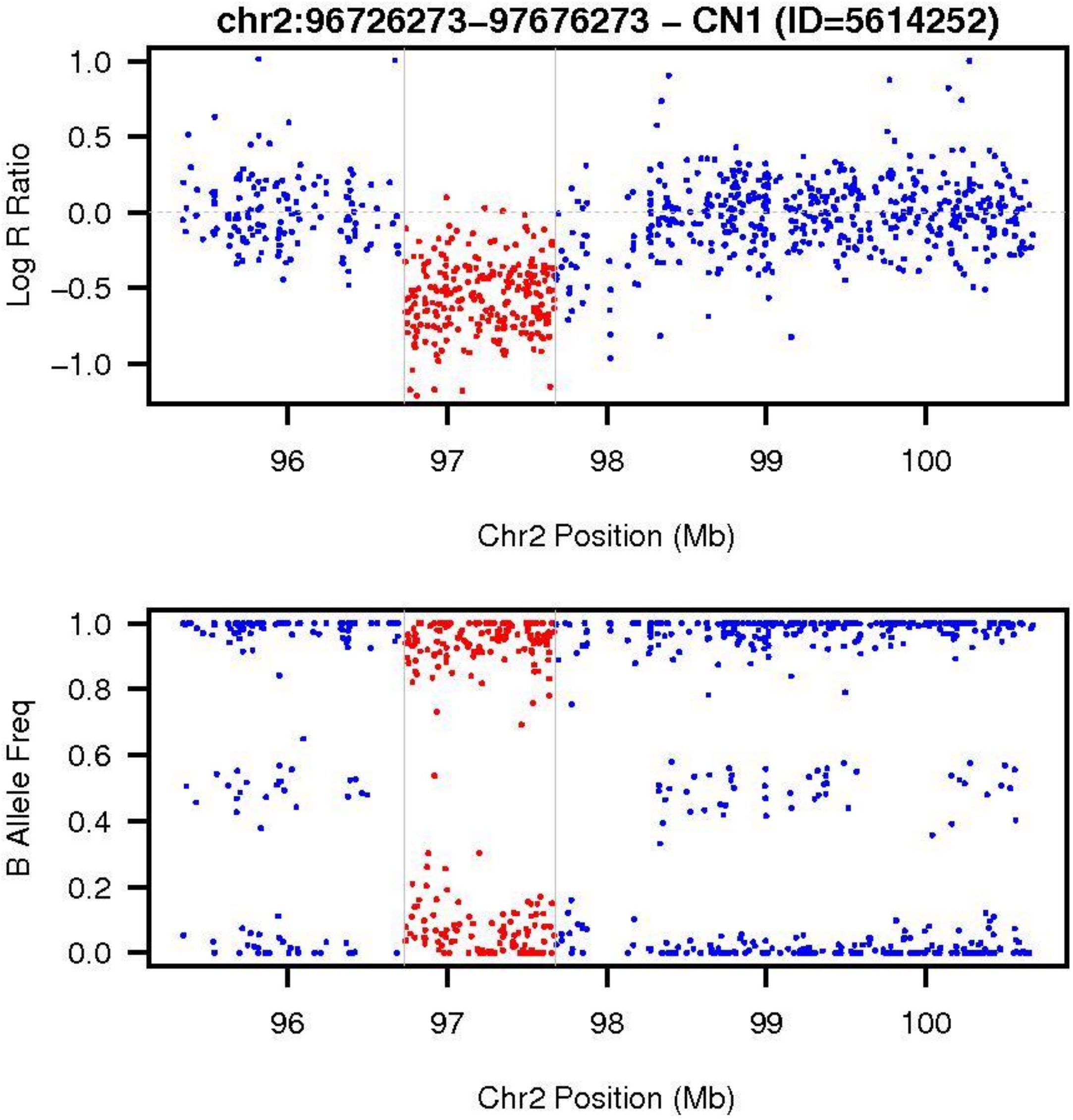
NOTE: interpretation of plot titles are ‘CN1’ = Deletion, ‘CN3’ = Duplication, ‘ID’ = UK Biobank subject ID

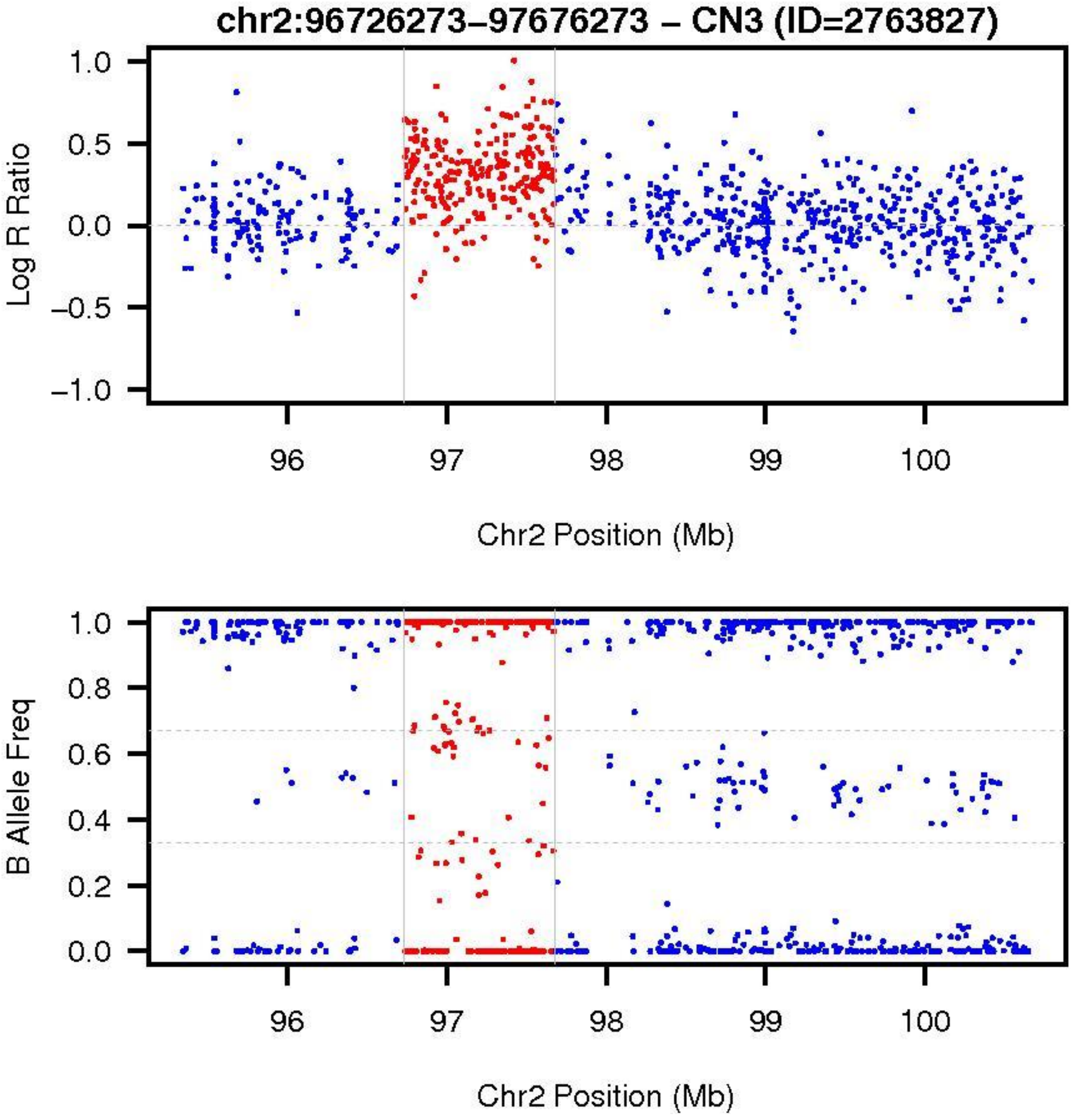

chr1:145288643–145628643 – CN1 (ID=6007194)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
144
145
146
147
148
Chr1 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
144
145
146
147
148
Chr1 Position (Mb)

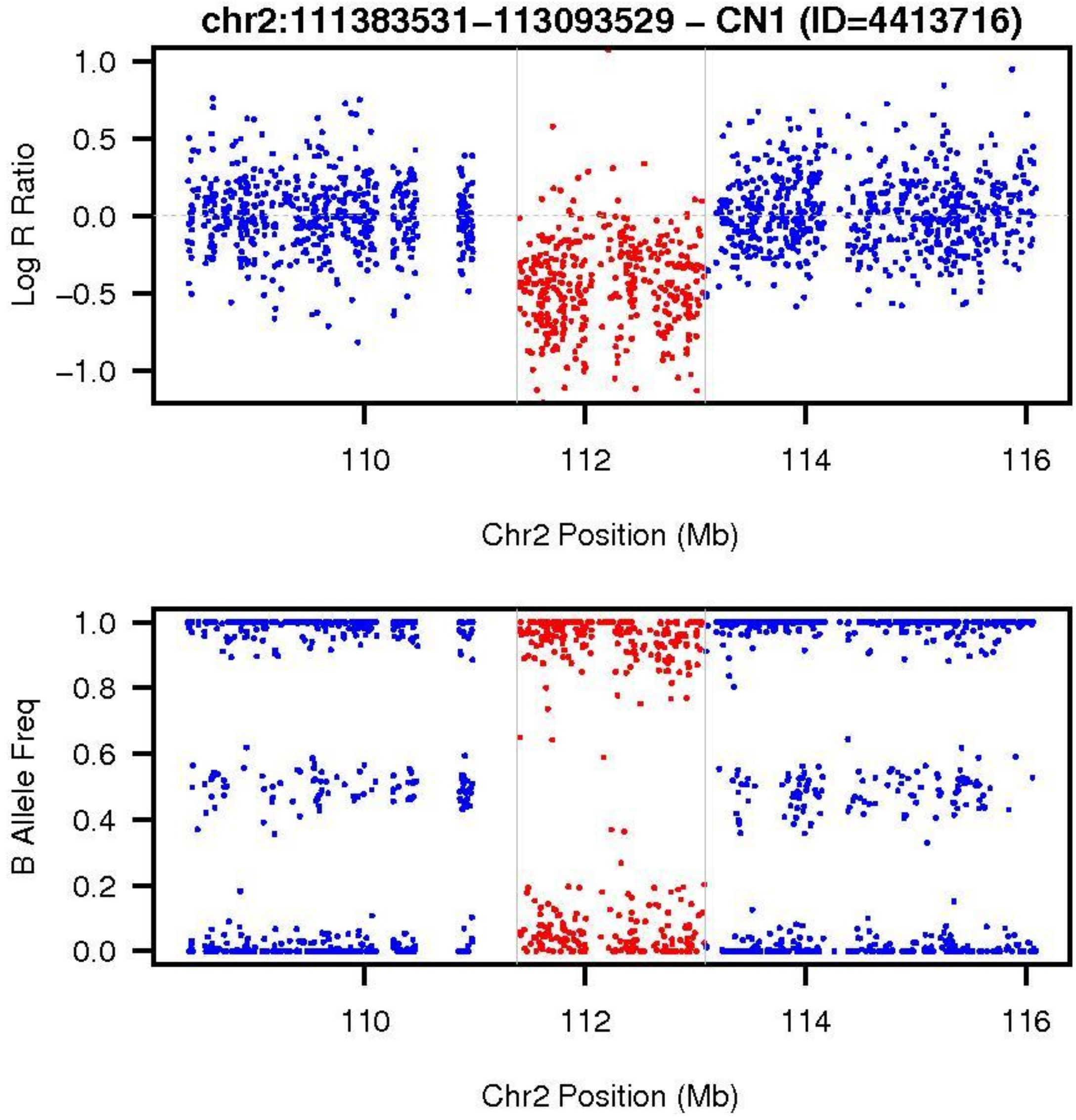

chr1:146573376–147393376 – CN1 (ID=5715895)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
144
145
146
147
148
149
150
Chr1 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
144
145
146
147
148
149
150
Chr1 Position (Mb)

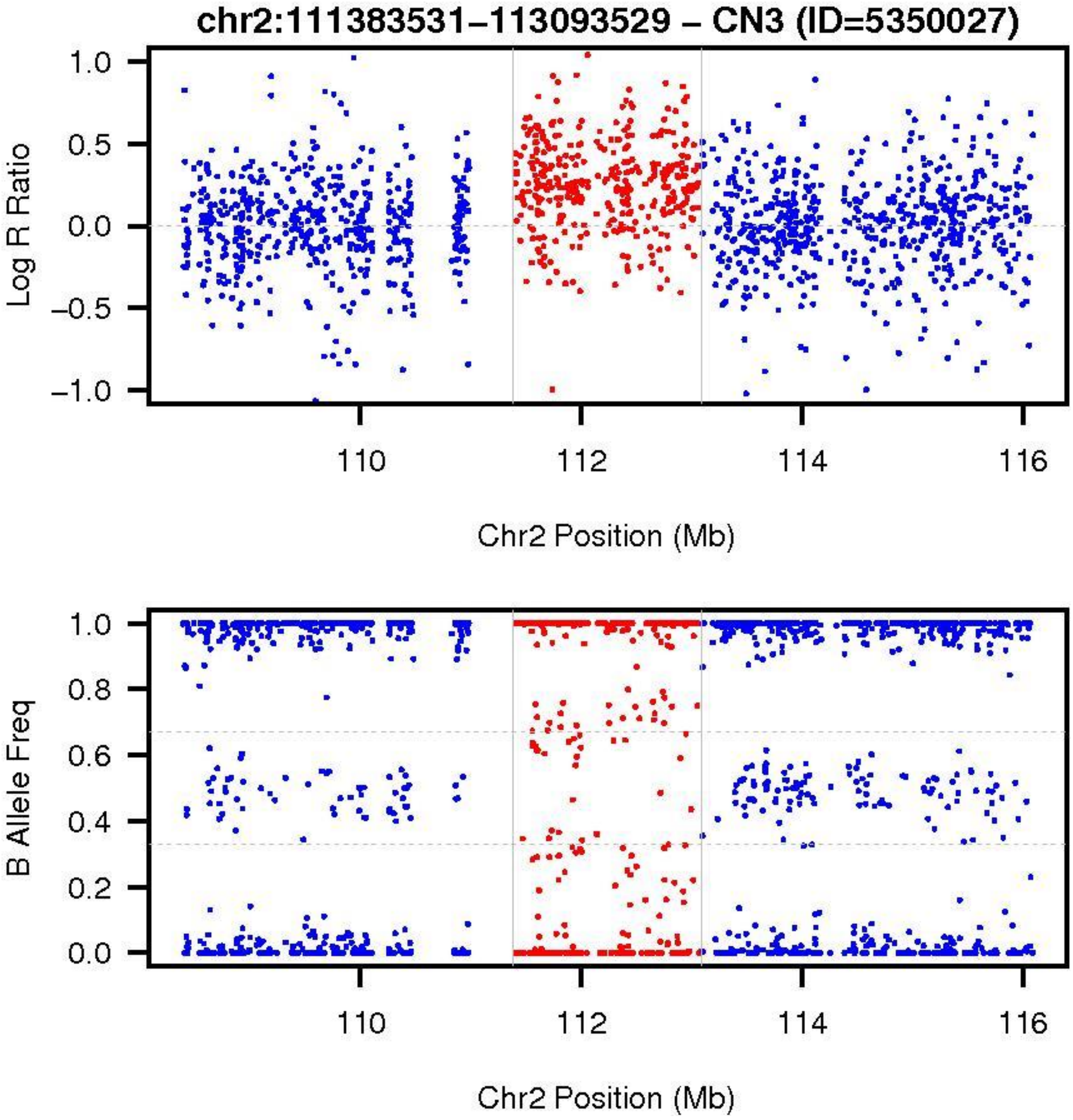

chr1:168733376–173733377 – CN3 (ID=3883187)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
165
170
175
Chr1 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
165
170
175
Chr1 Position (Mb)

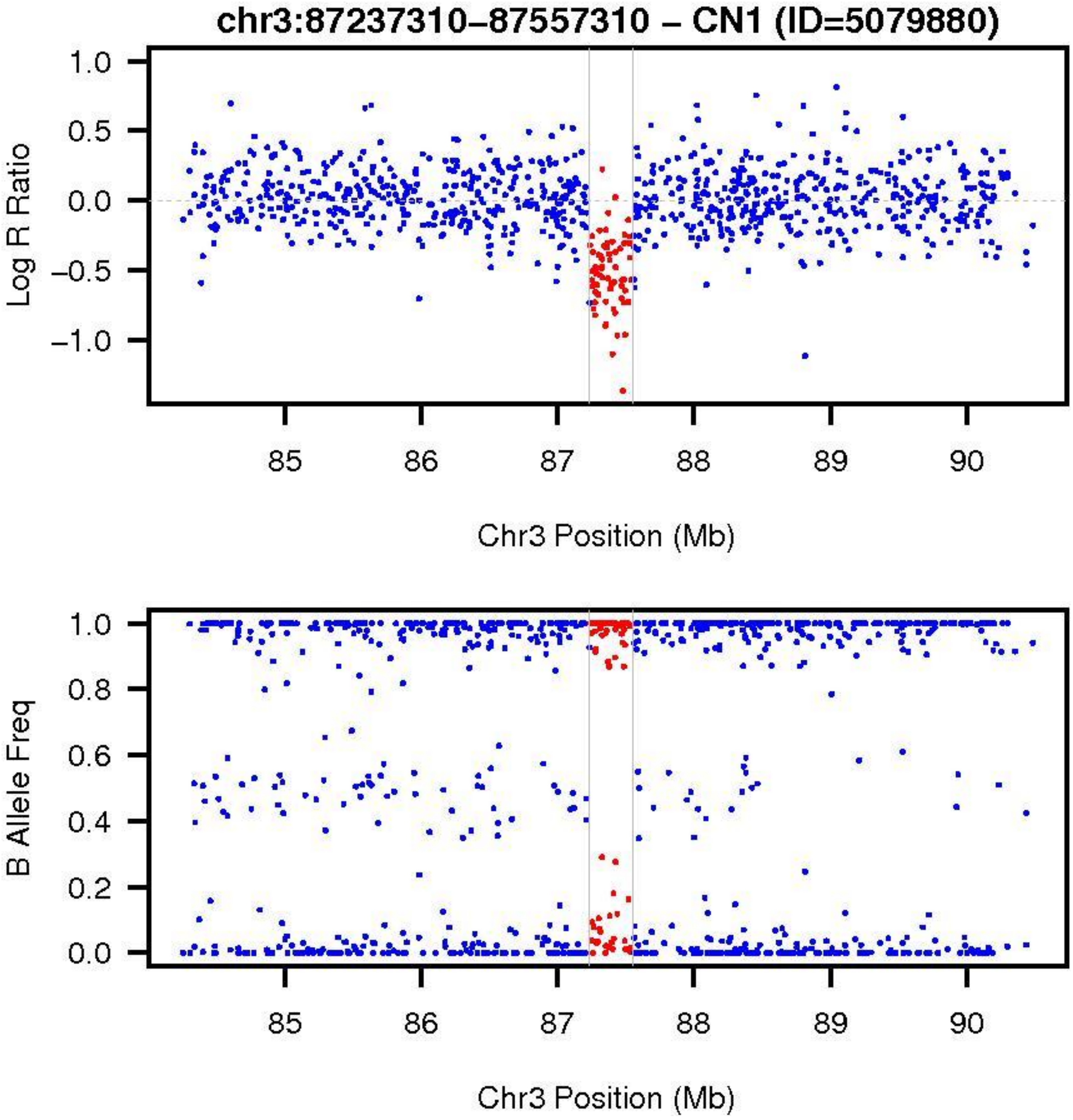

chr2:50146496–51256496 – CN3 (ID=3608208)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
47
48
49
50
51
52
53
54
Chr2 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
47
48
49
50
51
52
53
54
Chr2 Position (Mb)

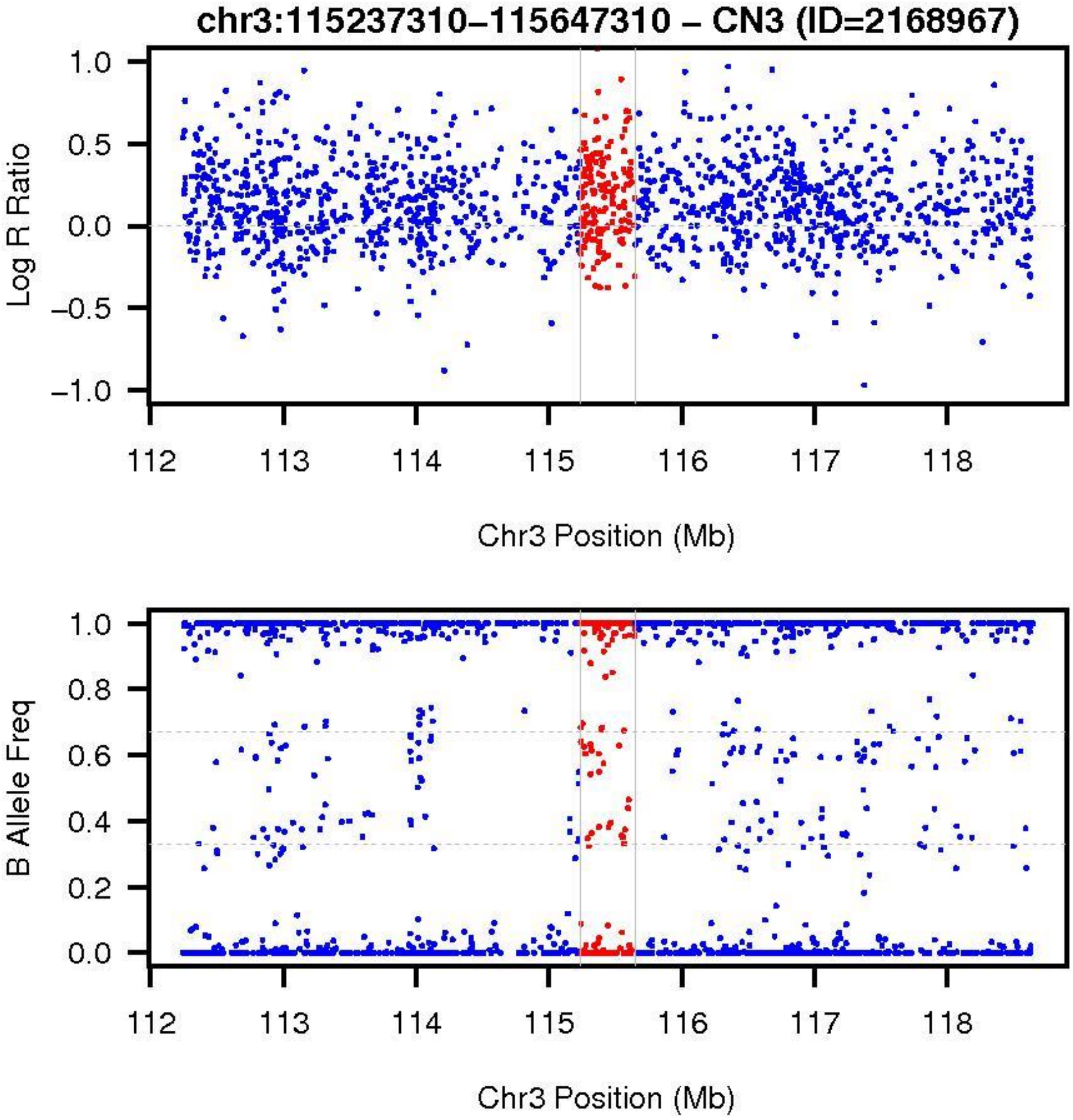

chr2:96726273–97676273 – CN1 (ID=5614252)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
96
97
98
99
100
Chr2 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
96
97
98
99
100
Chr2 Position (Mb)

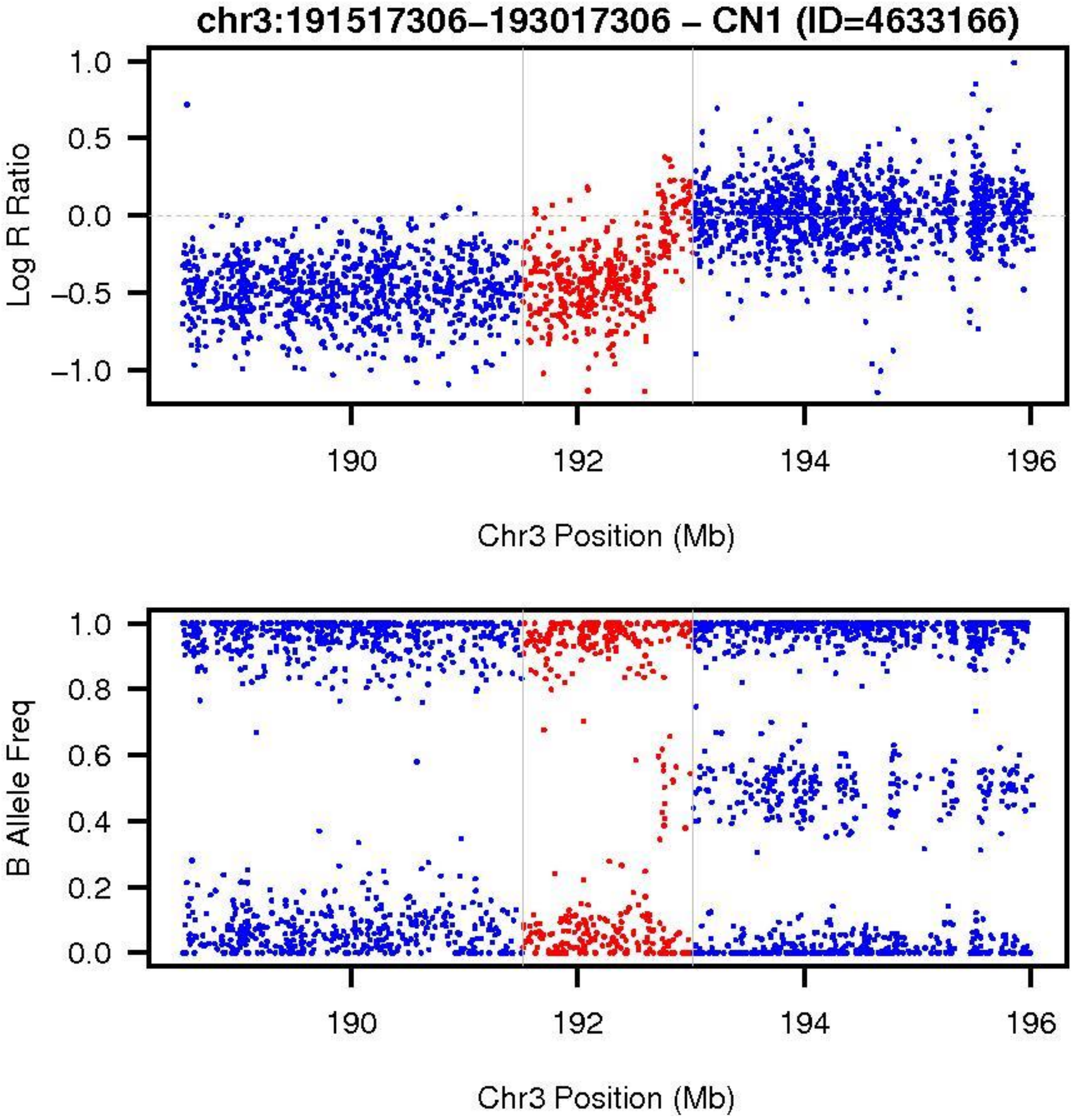

chr2:96726273–97676273 – CN3 (ID=2763827)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
96
97
98
99
100
Chr2 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
96
97
98
99
100
Chr2 Position (Mb)

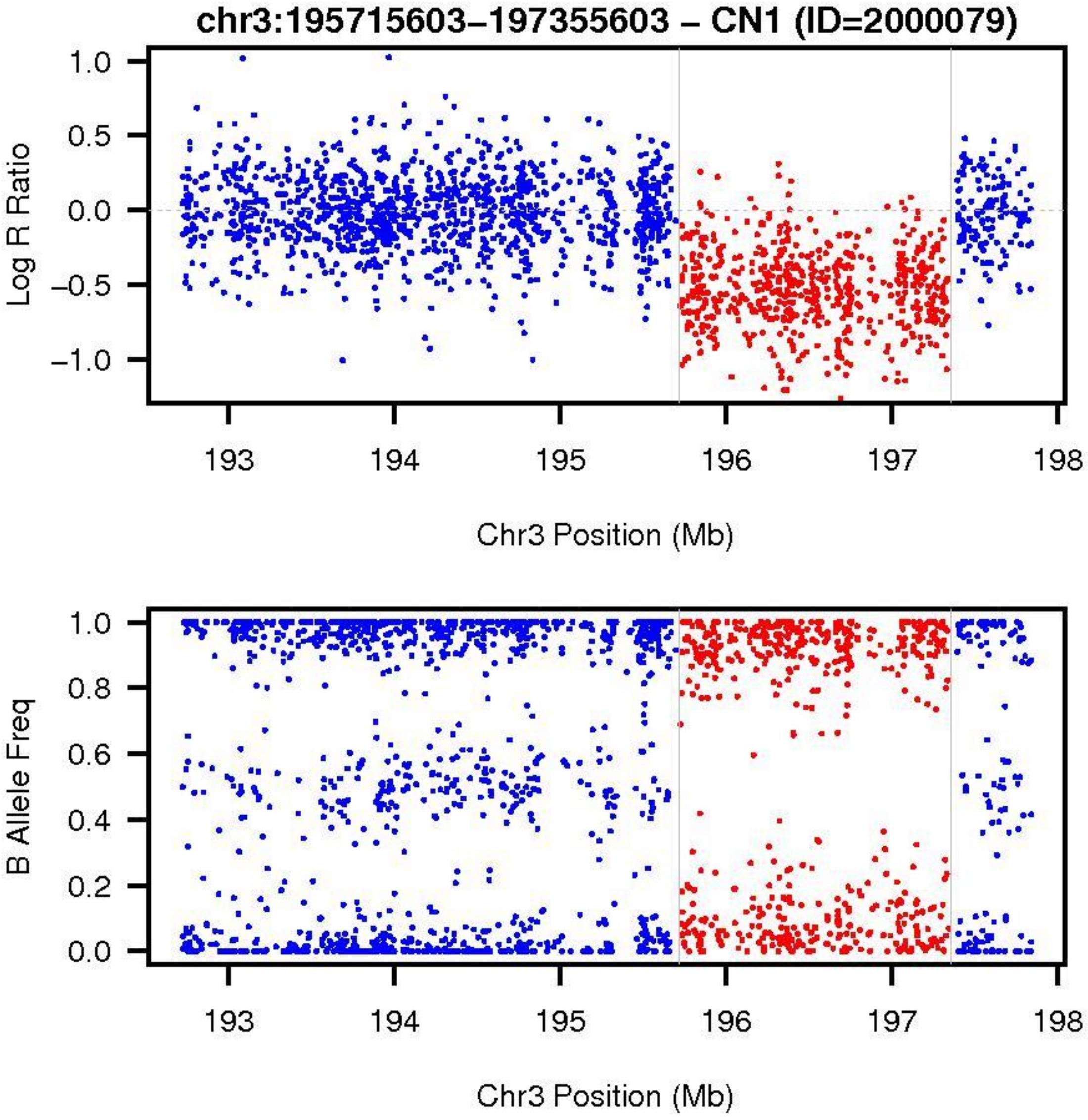

chr2:111383531–113093529 – CN1 (ID=4413716)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
110
112
114
116
Chr2 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
110
112
114
116
Chr2 Position (Mb)

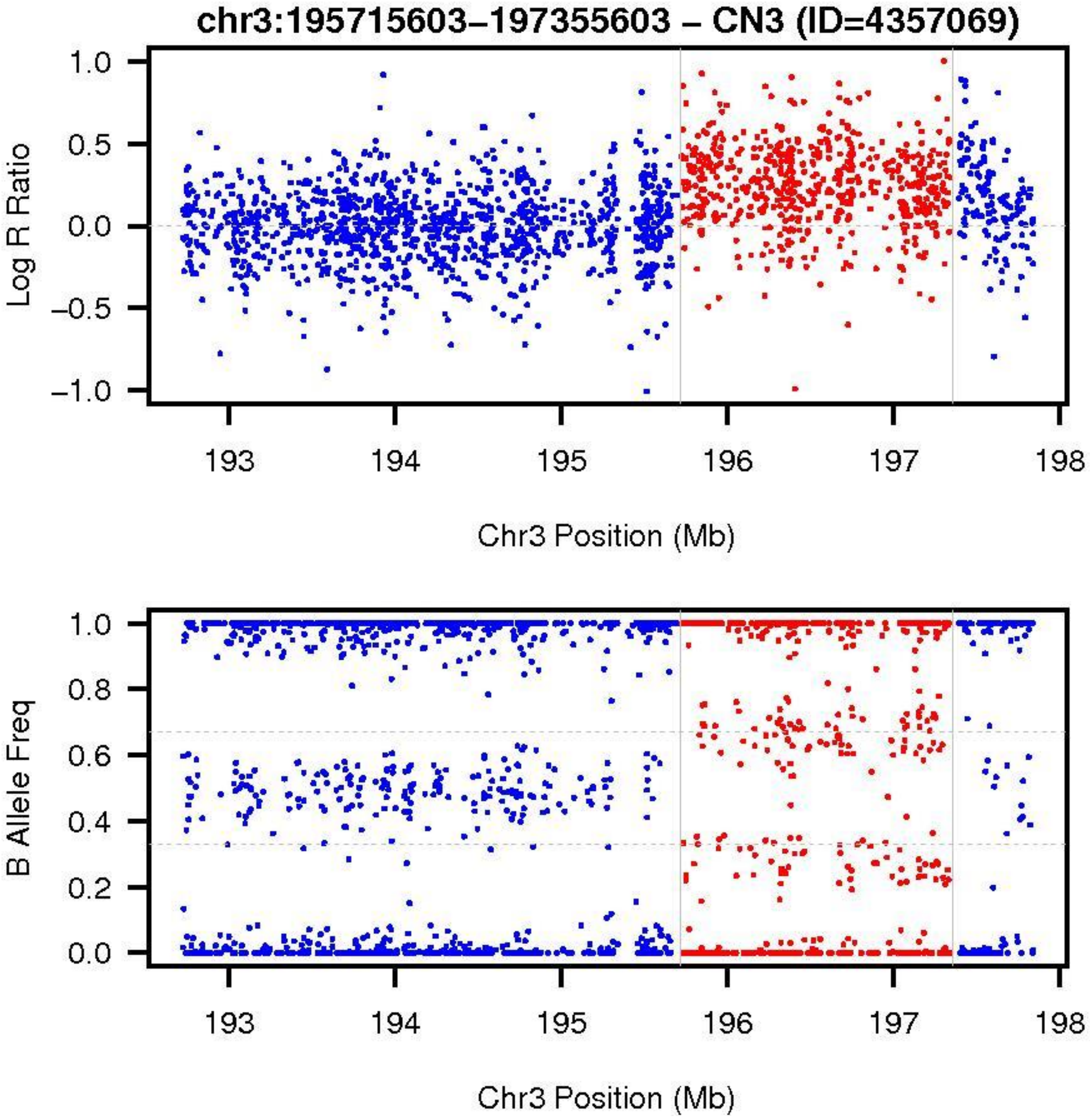

chr2:111383531–113093529 – CN3 (ID=5350027)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
110
112
114
116
Chr2 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
110
112
114
116
Chr2 Position (Mb)

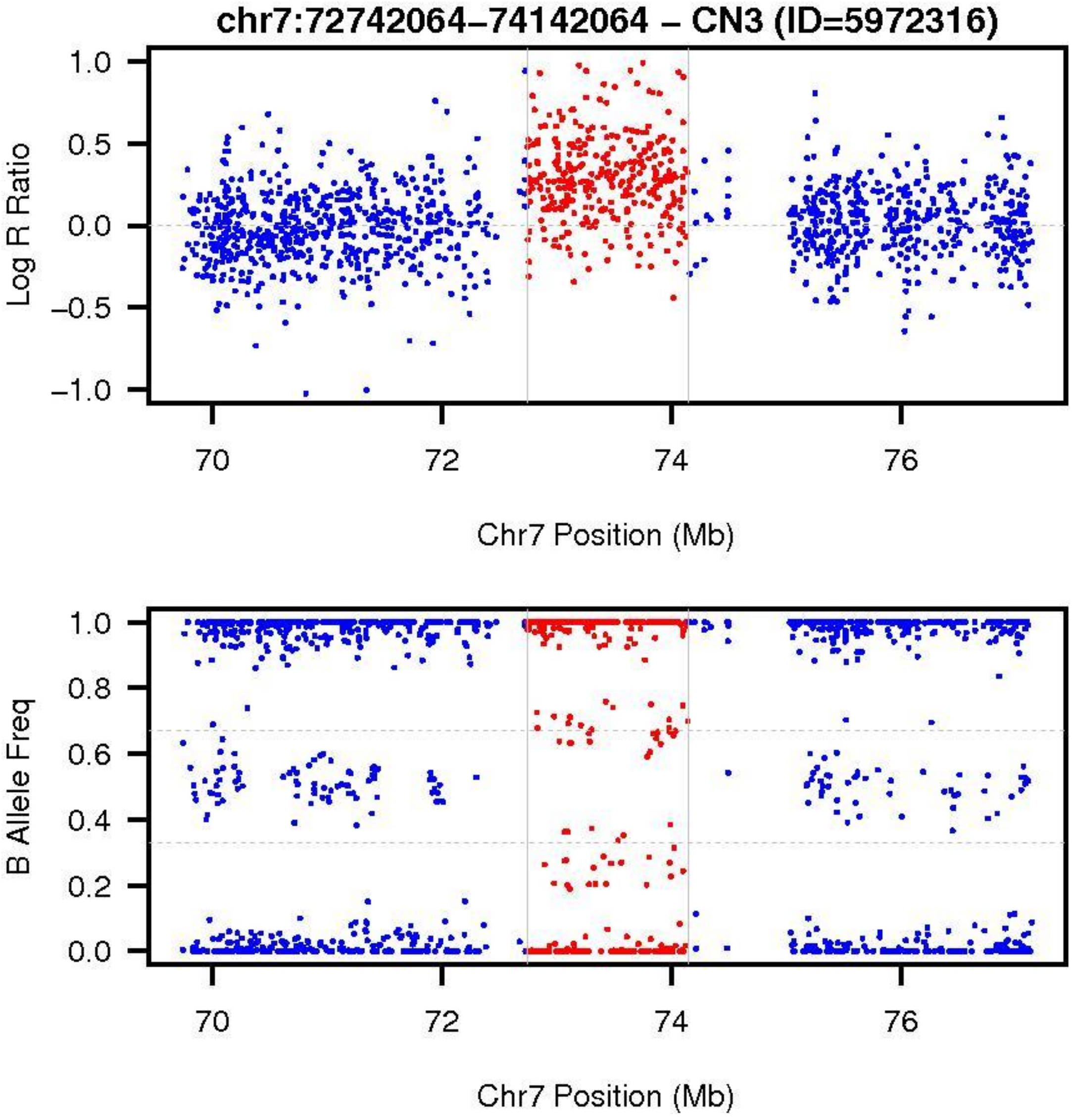

chr3:87237310–87557310 – CN1 (ID=5079880)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
85
86
87
88
89
90
Chr3 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
85
86
87
88
89
90
Chr3 Position (Mb)

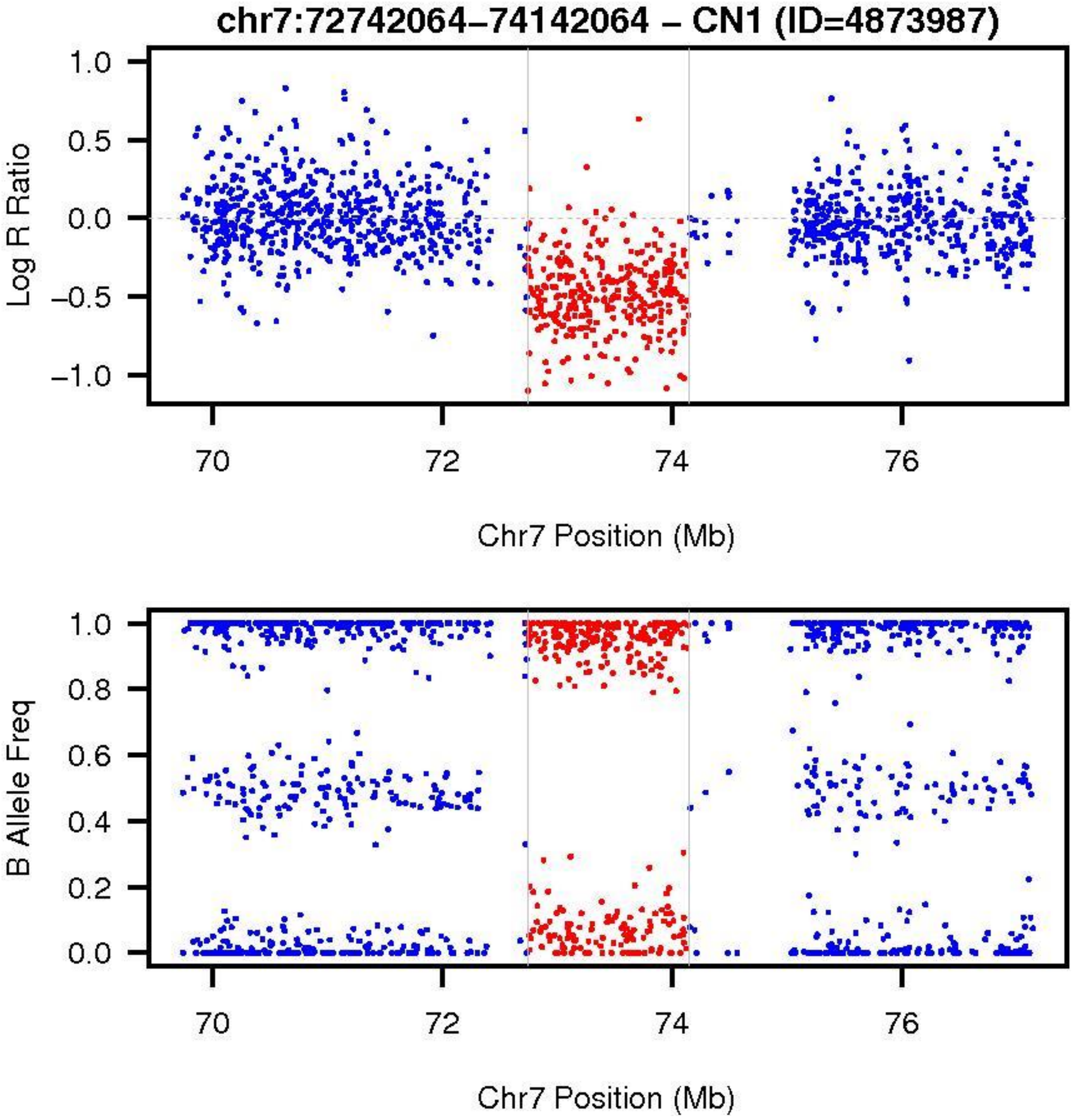

chr3:115237310–115647310 – CN3 (ID=2168967)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
112
113
114
115
116
117
118
Chr3 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
112
113
114
115
116
117
118
Chr3 Position (Mb)

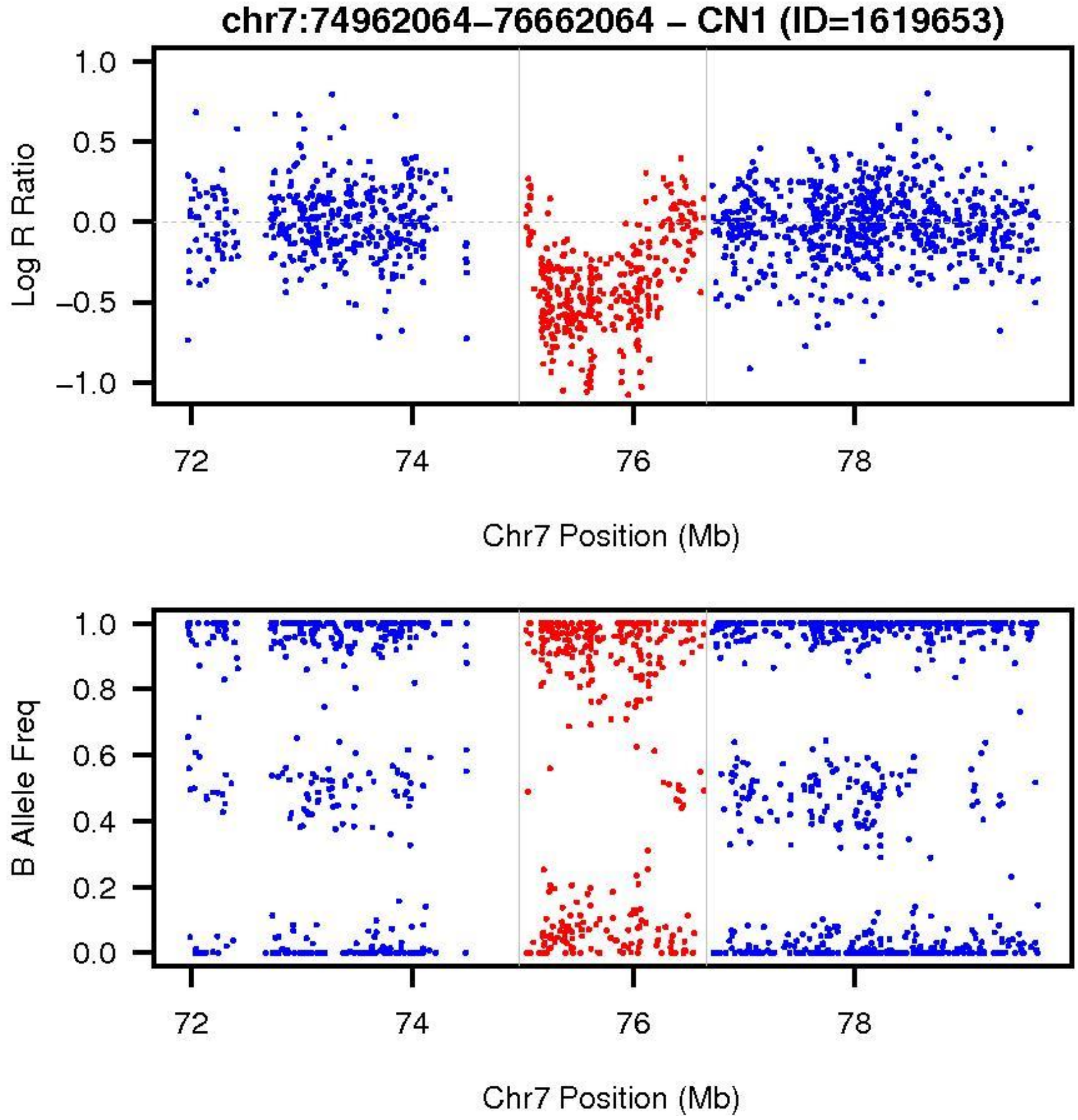

chr3:191517306–193017306 – CN1 (ID=4633166)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
190
192
194
196
Chr3 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
190
192
194
196
Chr3 Position (Mb)

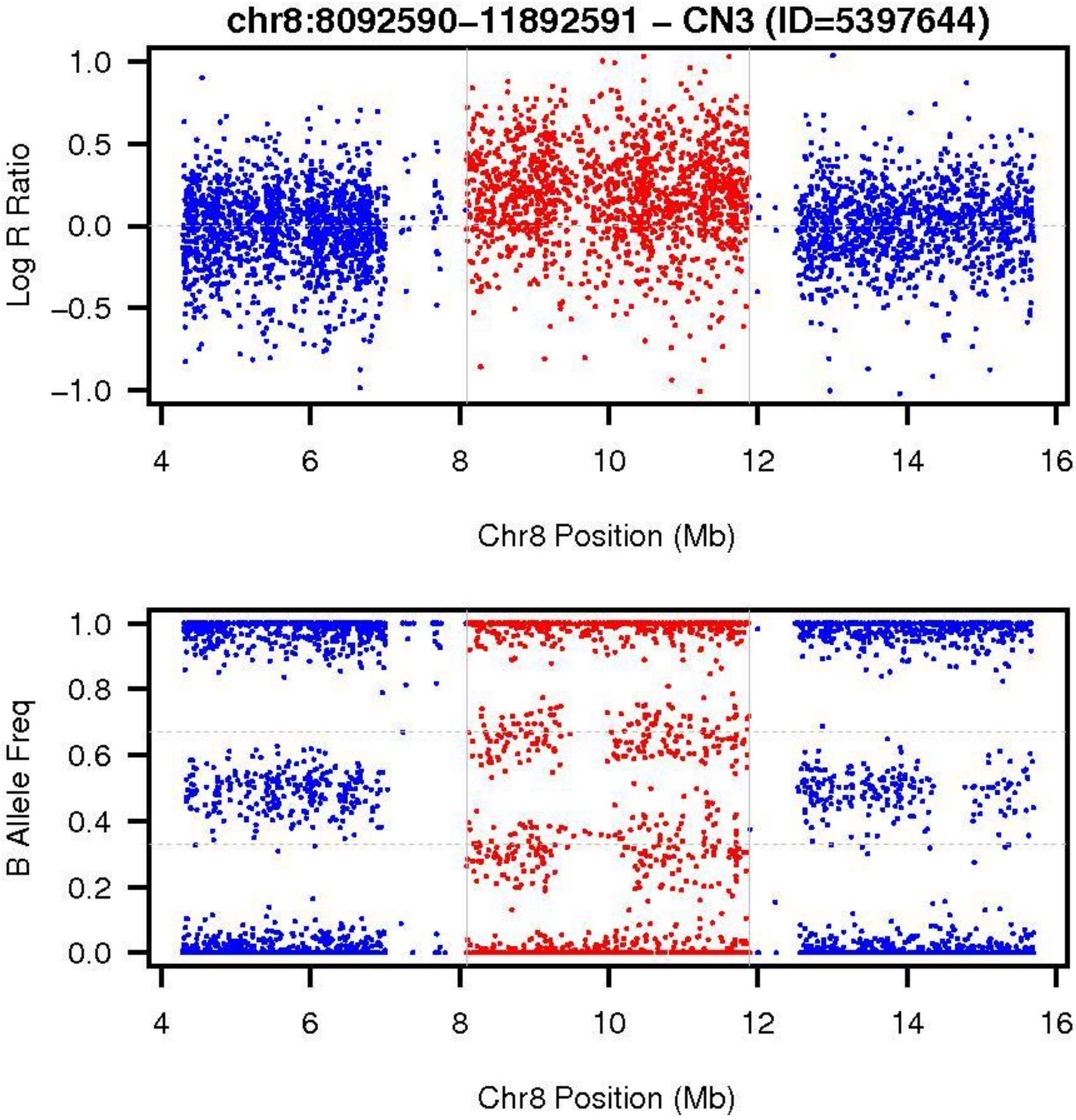

chr3:195715603–197355603 – CN1 (ID=2000079)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
193
194
195
196
197
198
Chr3 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
193
194
195
196
197
198
Chr3 Position (Mb)

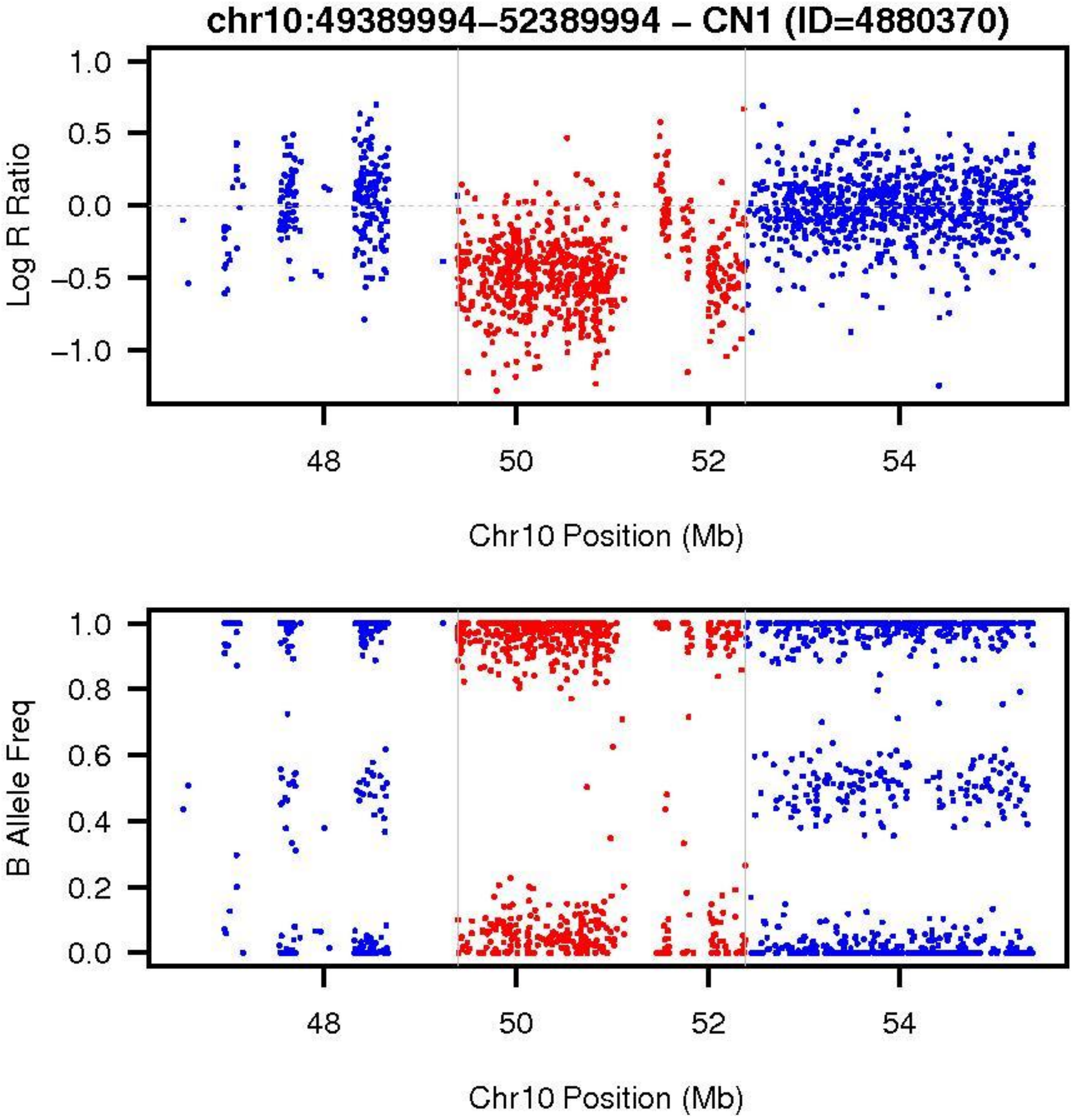

chr3:195715603–197355603 – CN3 (ID=4357069)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
193
194
195
196
197
198
Chr3 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
193
194
195
196
197
198
Chr3 Position (Mb)

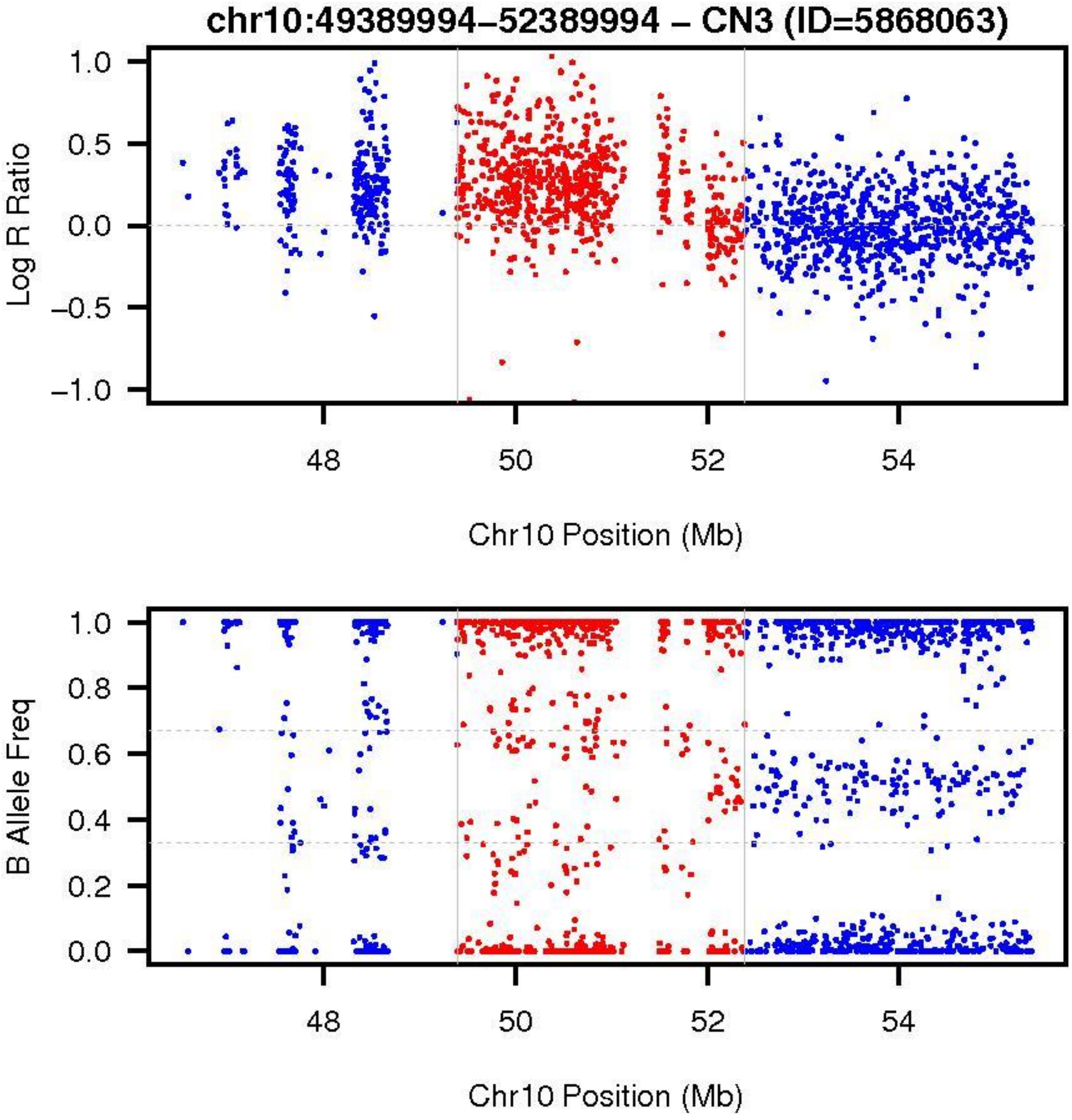

chr7:72742064–74142064 – CN3 (ID=5972316)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
70
72
74
76
Chr7 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
70
72
74
76
Chr7 Position (Mb)

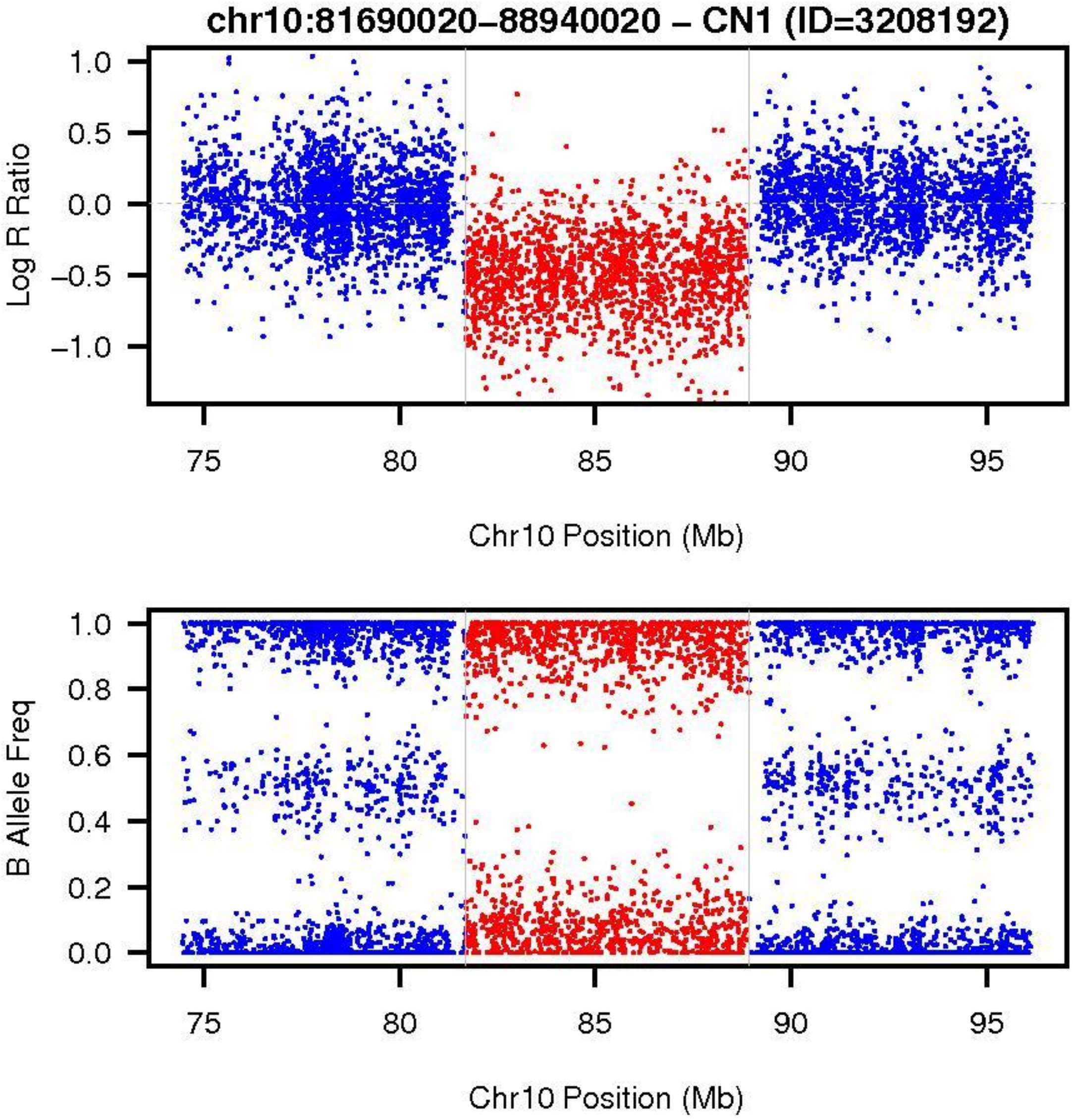

chr7:72742064–74142064 – CN1 (ID=4873987)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
70
72
74
76
Chr7 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
70
72
74
76
Chr7 Position (Mb)

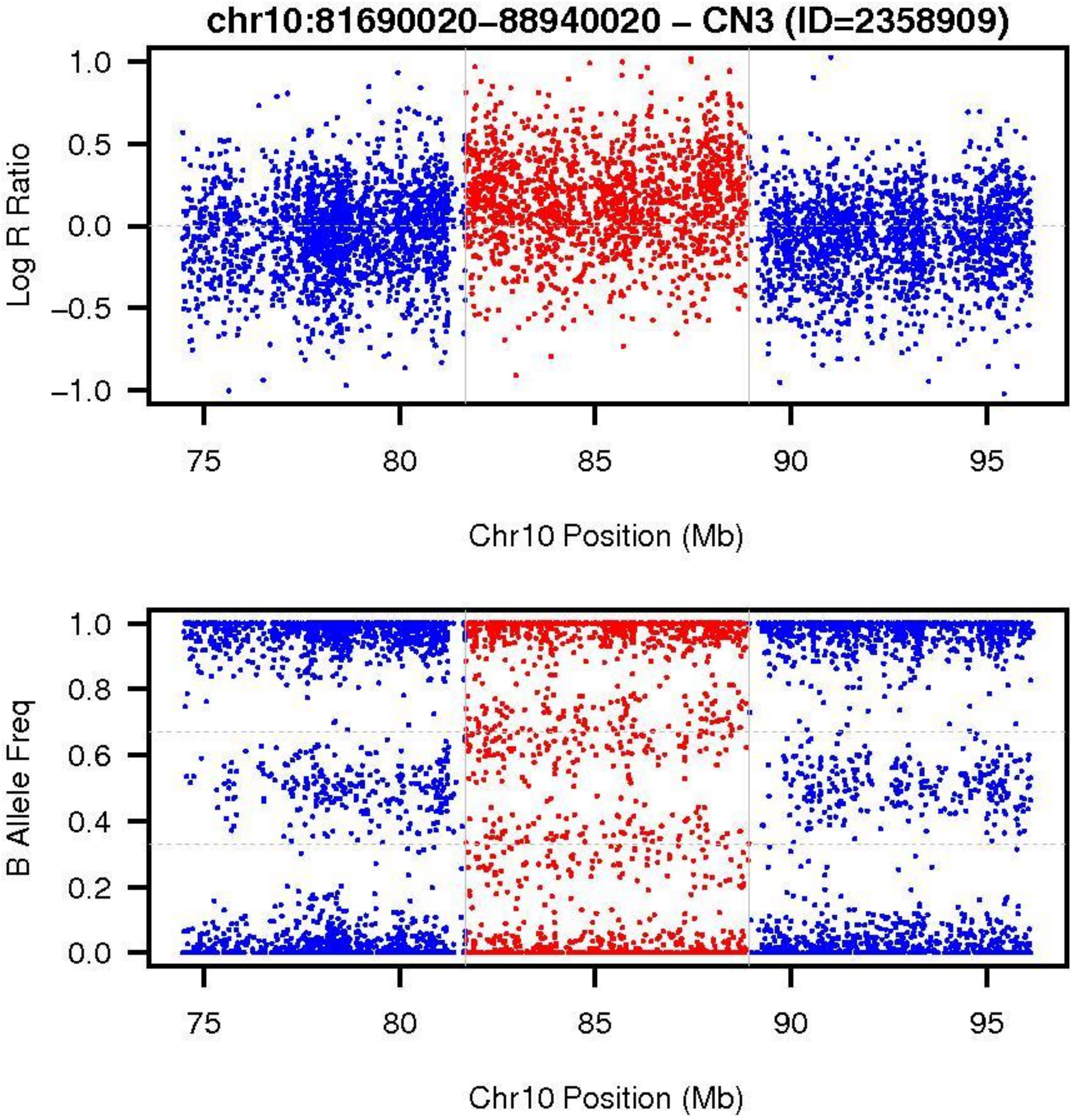

chr7:74962064–76662064 – CN1 (ID=1619653)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
72
74
76
78
Chr7 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
72
74
76
78
Chr7 Position (Mb)

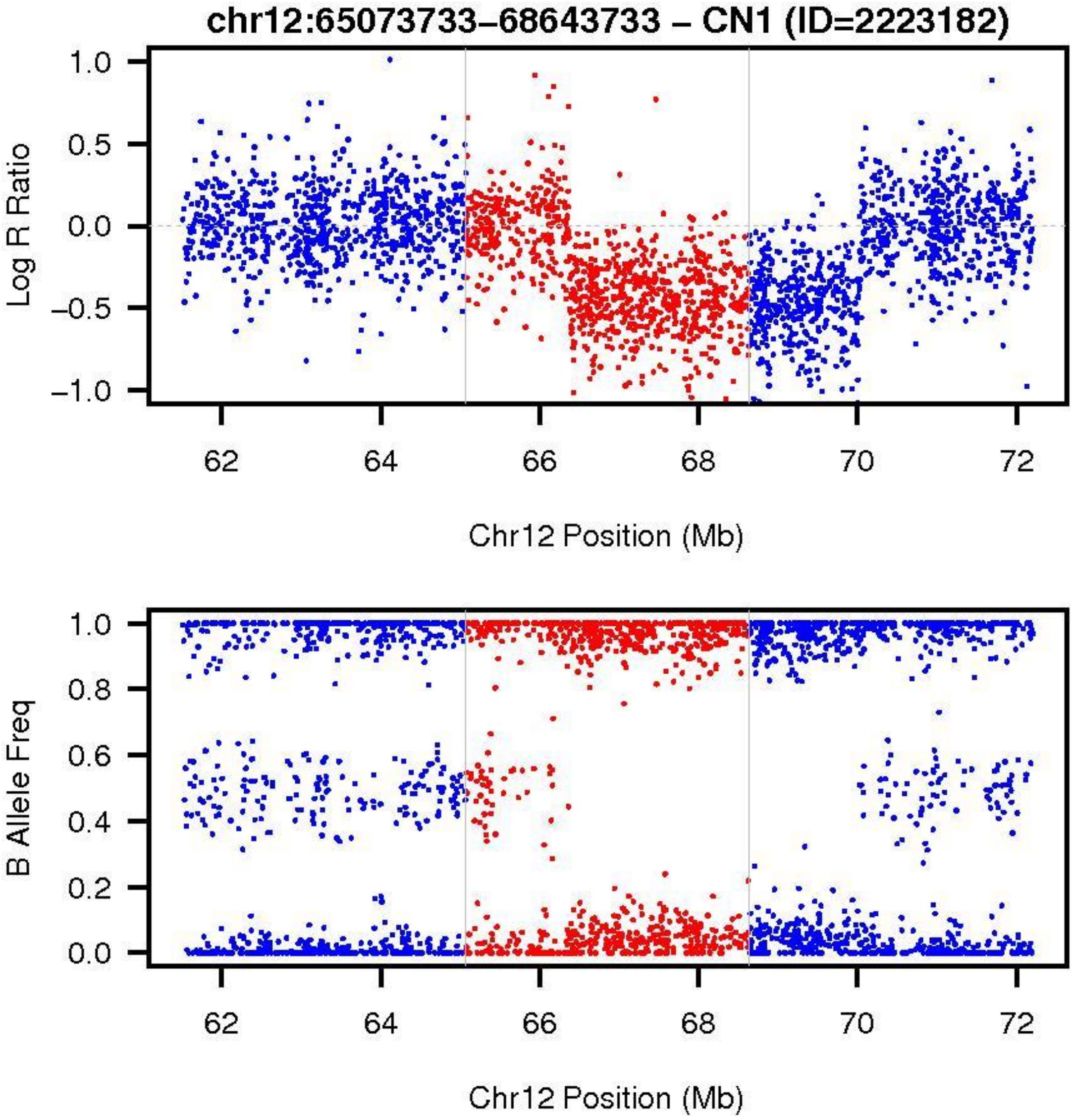

chr8:8092590−11892591 − CN3 (ID=5397644)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
4
6
8
10
12
14
16
Chr8 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
4
6
8
10
12
14
16
Chr8 Position (Mb)

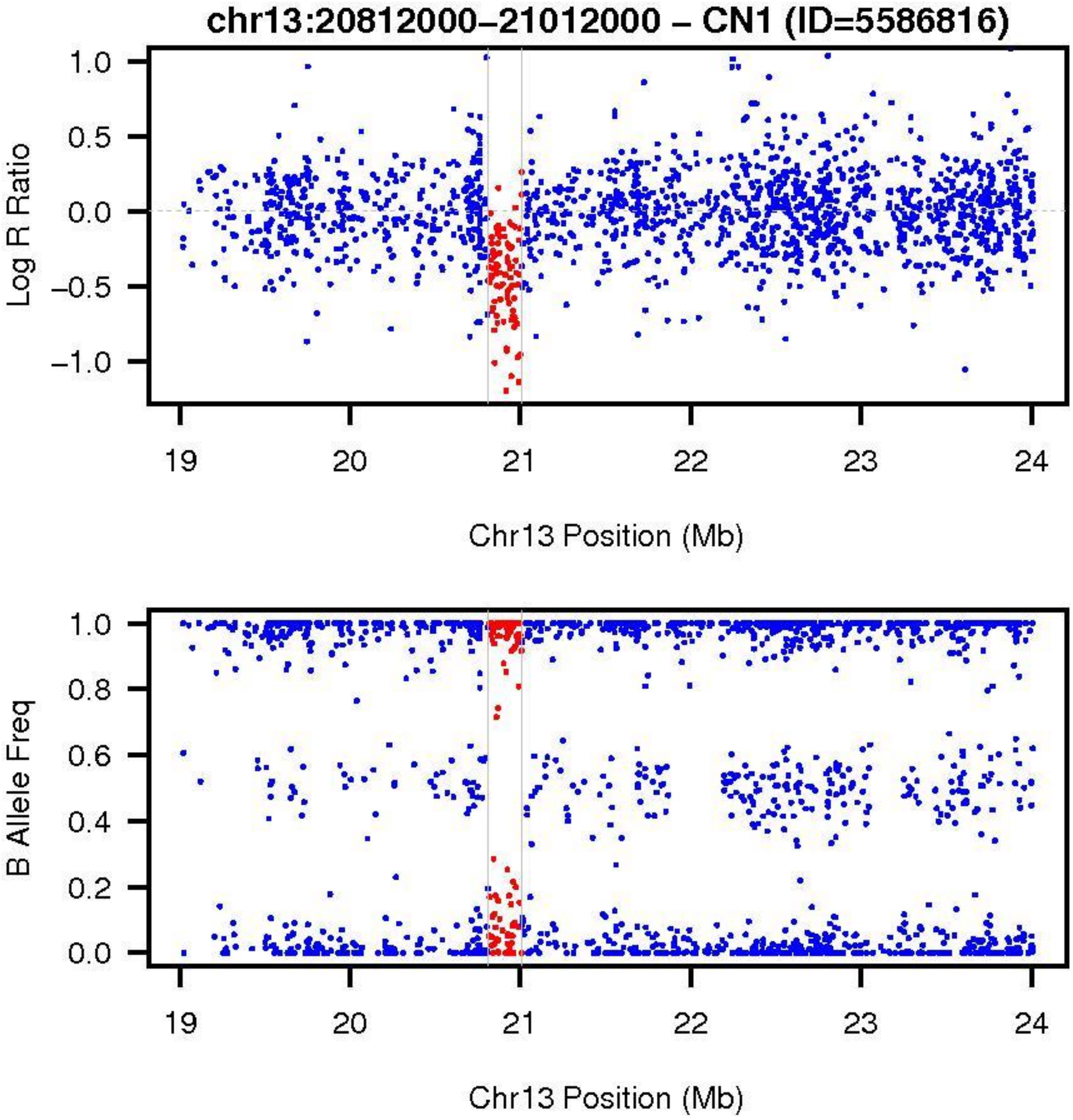

chr10:49389994–52389994 – CN1 (ID=4880370)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
48
50
52
54
Chr10 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
48
50
52
54
Chr10 Position (Mb)

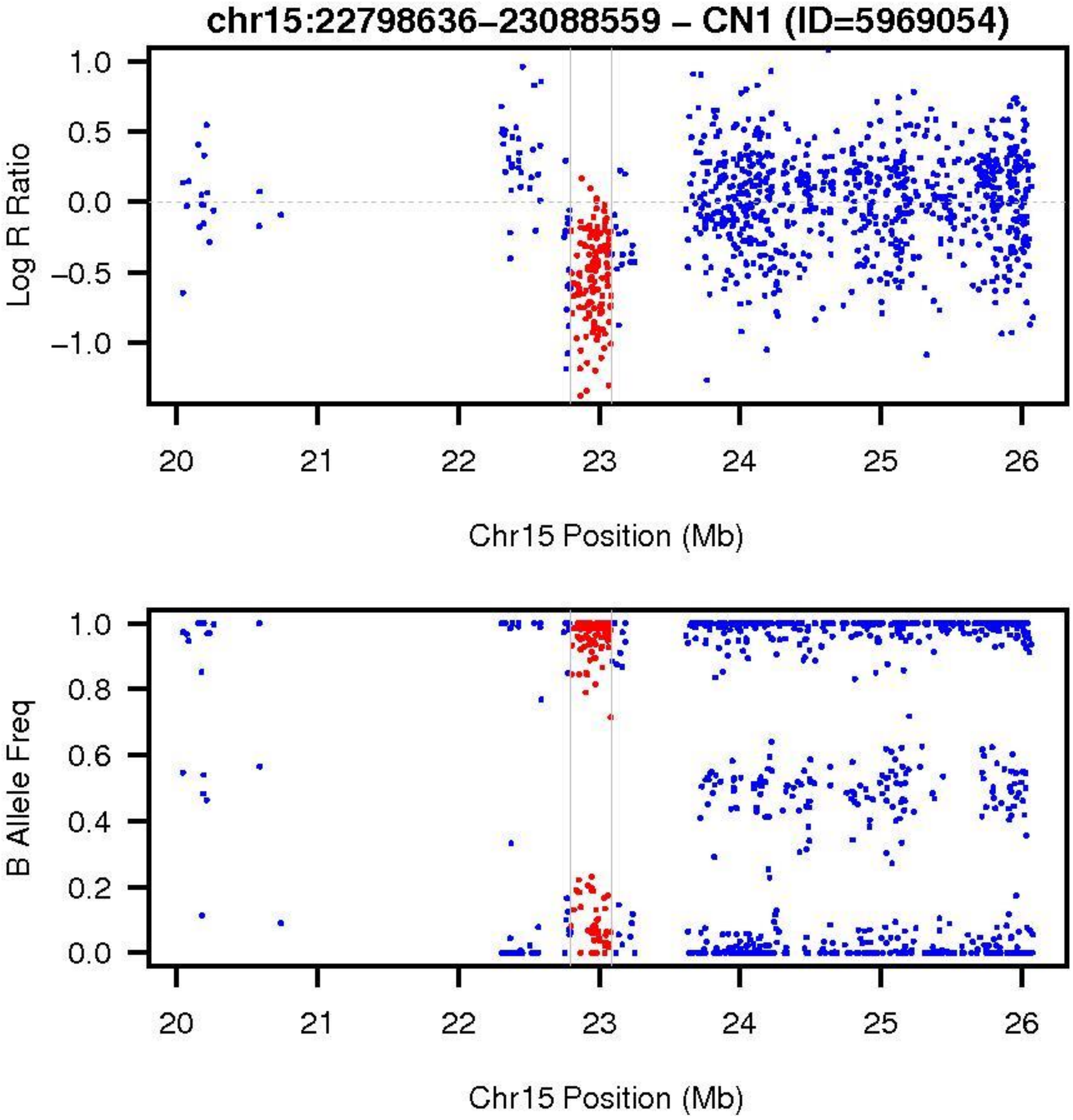

chr10:49389994–52389994 – CN3 (ID=5868063)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
48
50
52
54
Chr10 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
48
50
52
54
Chr10 Position (Mb)


chr10:81690020–88940020 – CN1 (ID=3208192)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
75
80
85
90
95
Chr10 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
75
80
85
90
95
Chr10 Position (Mb)

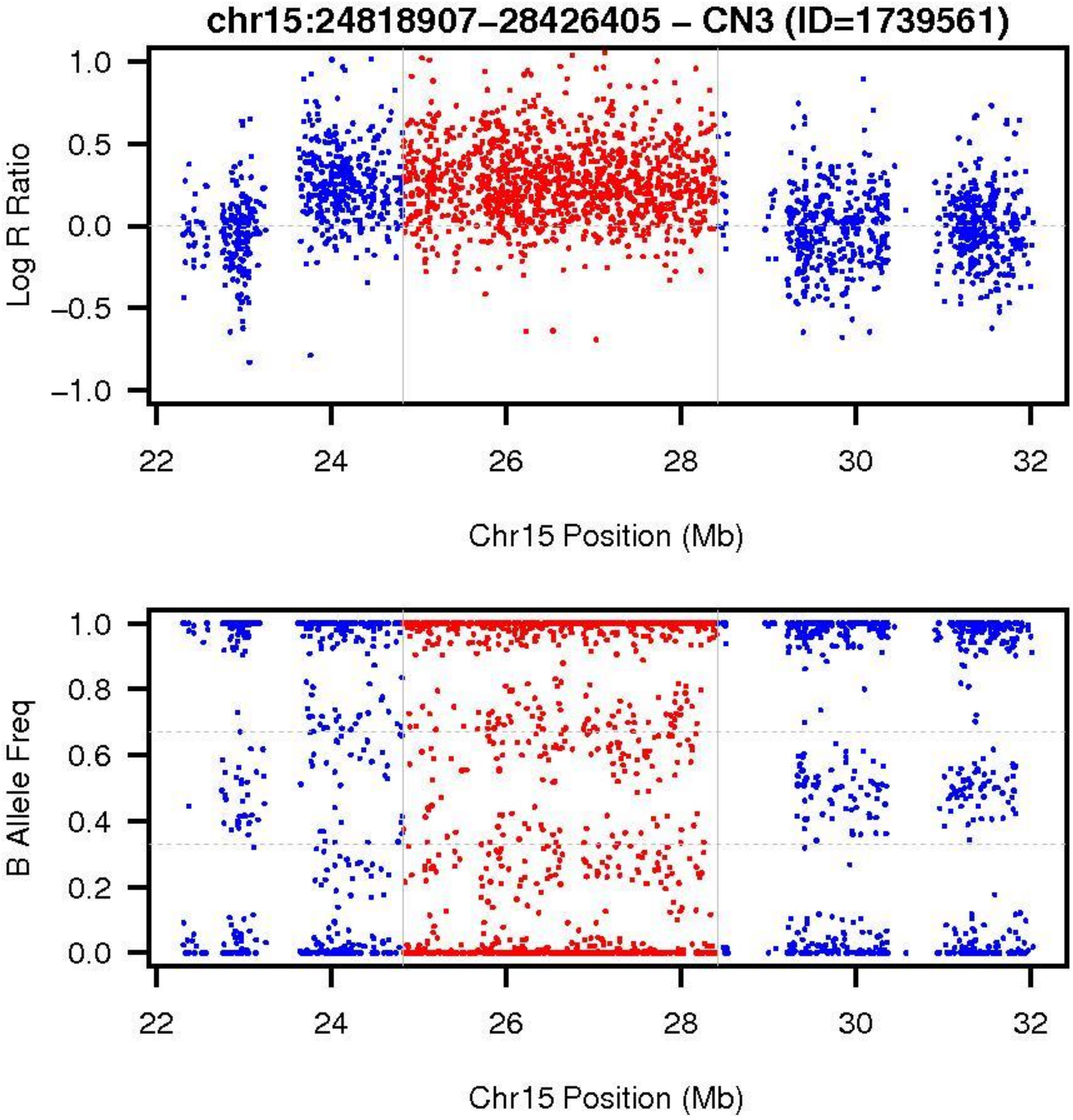

chr10:81690020–88940020 – CN3 (ID=2358909)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
75
80
85
90
95
Chr10 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
75
80
85
90
95
Chr10 Position (Mb)

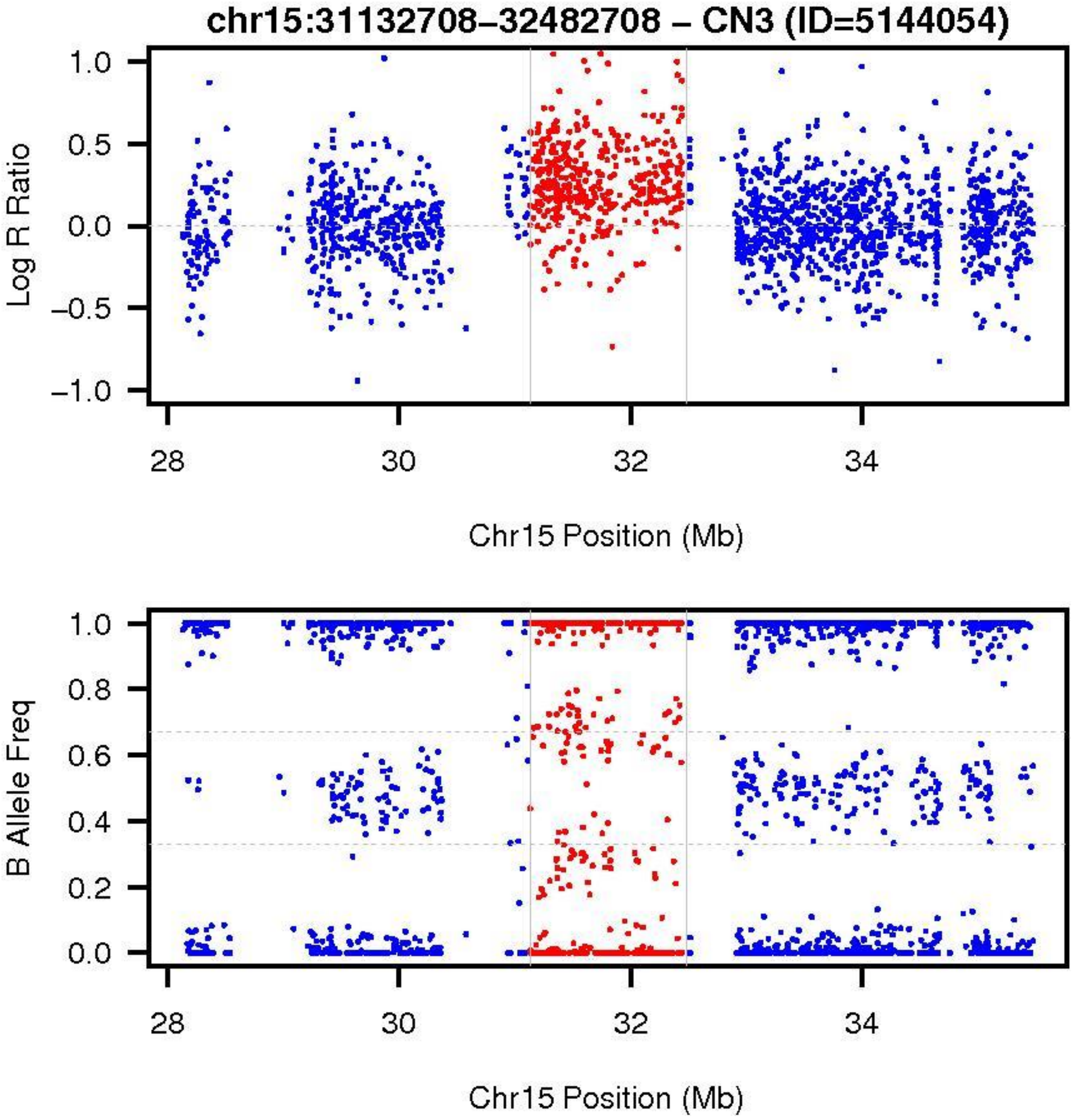

chr12:65073733–68643733 – CN1 (ID=2223182)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
62
64
66
68
70
72
Chr12 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
62
64
66
68
70
72
Chr12 Position (Mb)

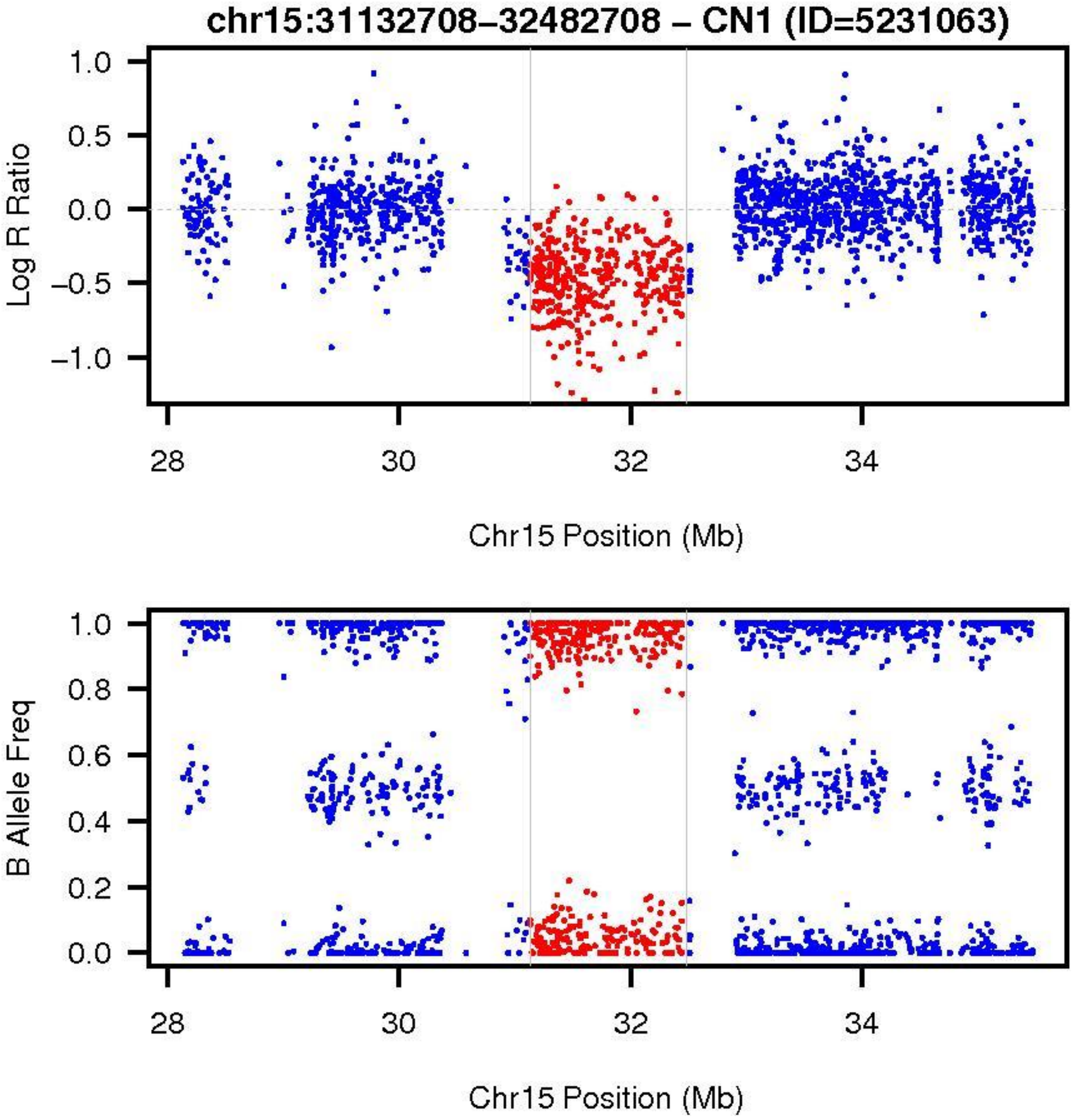

chr13:20812000–21012000 – CN1 (ID=5586816)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
19
20
21
22
23
24
Chr13 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
19
20
21
22
23
24
Chr13 Position (Mb)

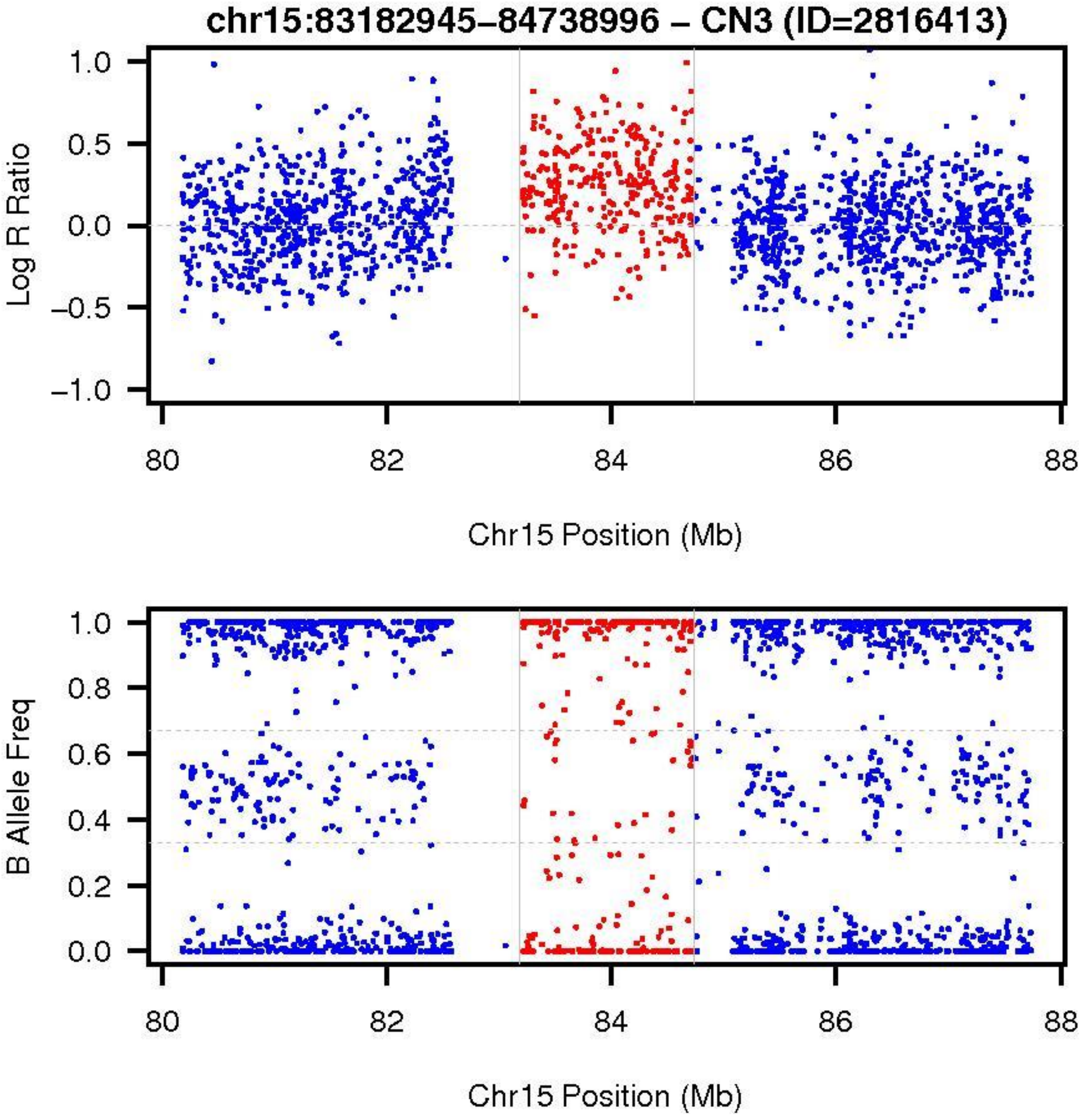


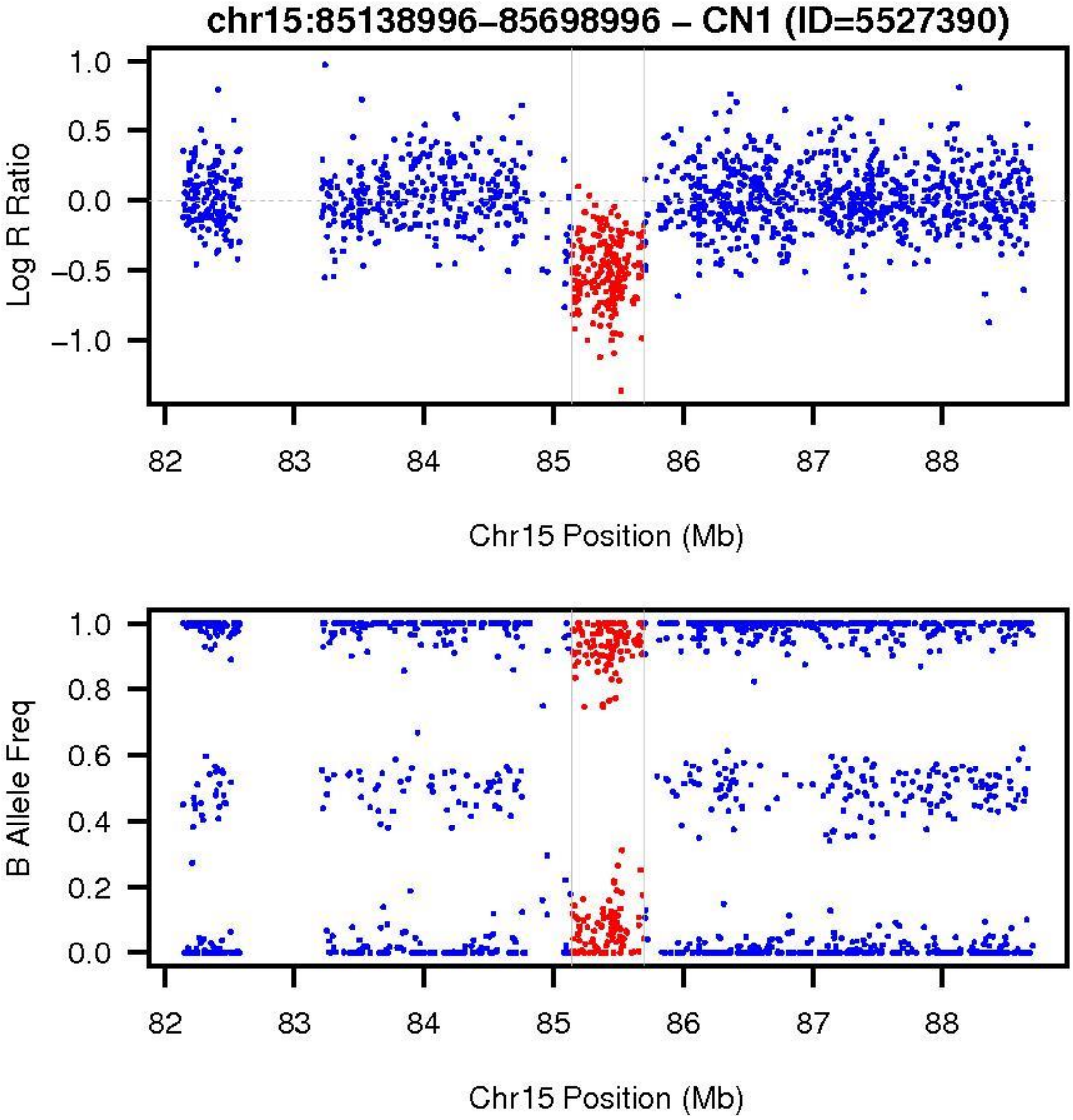
chr15:22798636–23088559 – CN1 (ID=5969054)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
20
21
22
23
24
25
26
Chr15 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
20
21
22
23
24
25
26
Chr15 Position (Mb)

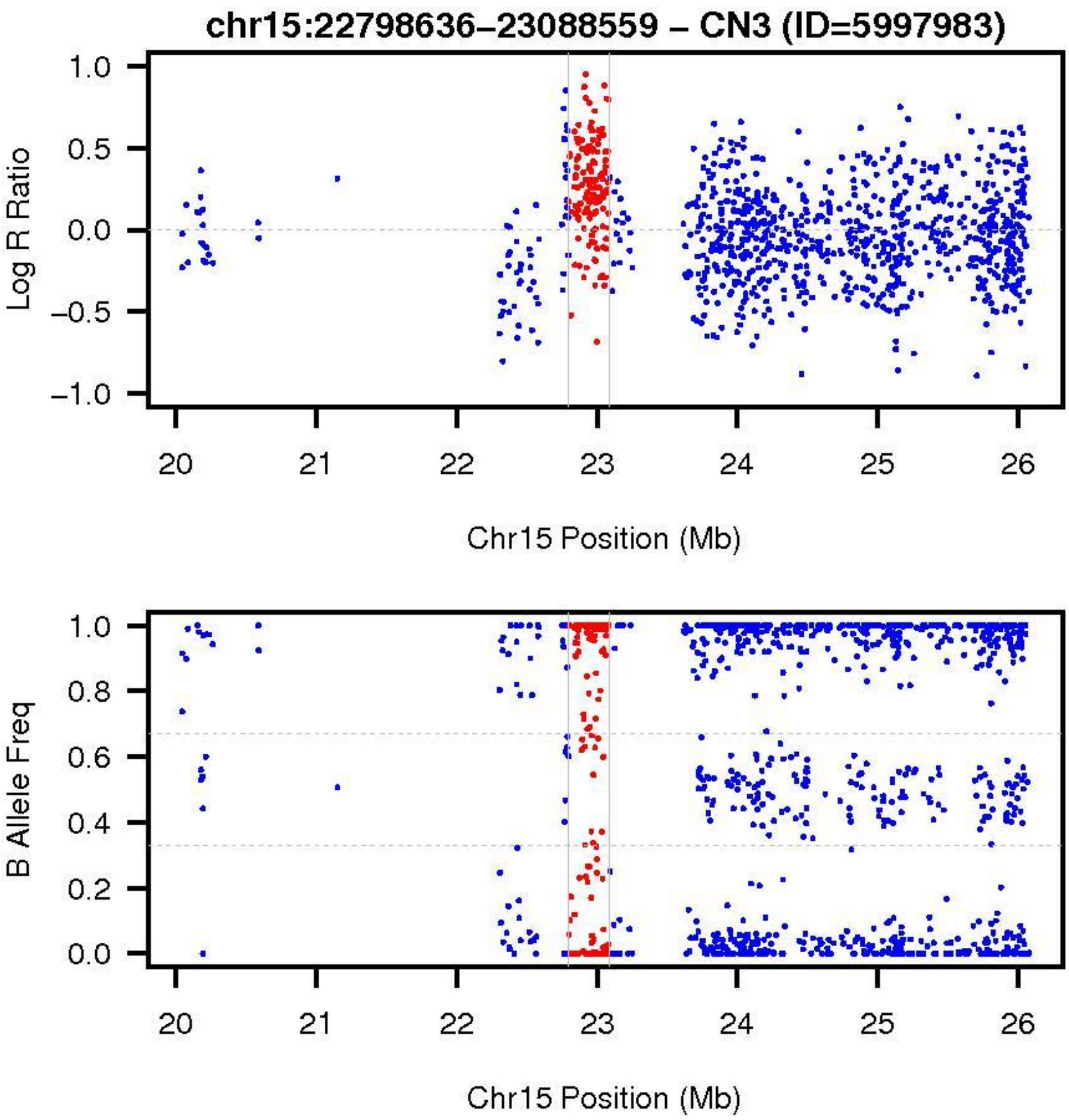

chr15:24818907–28426405 – CN3 (ID=1739561)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
22
24
26
28
30
32
Chr15 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
22
24
26
28
30
32
Chr15 Position (Mb)

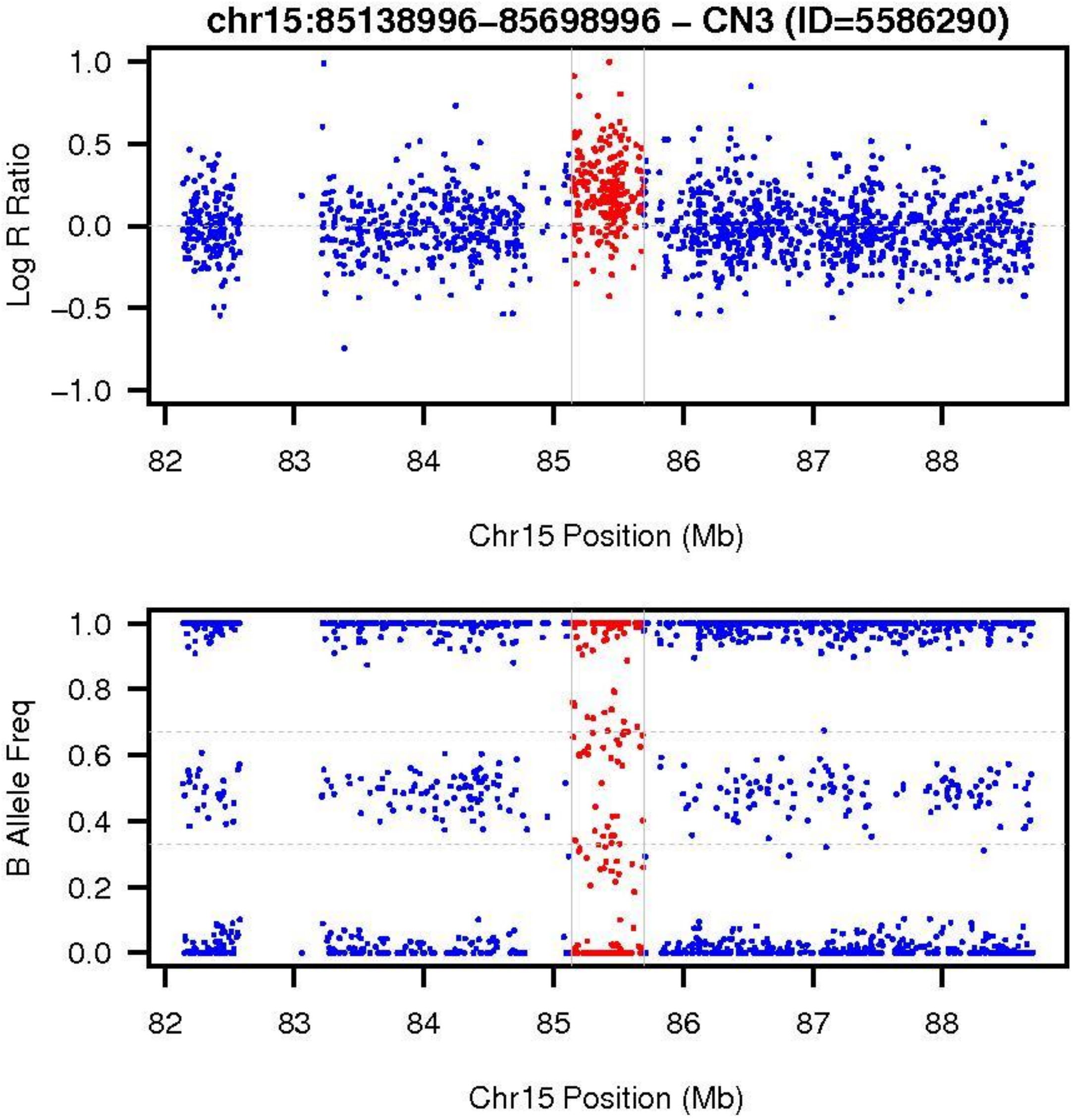

chr15:31132708–32482708 – CN3 (ID=5144054)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
28
30
32
34
Chr15 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
28
30
32
34
Chr15 Position (Mb)

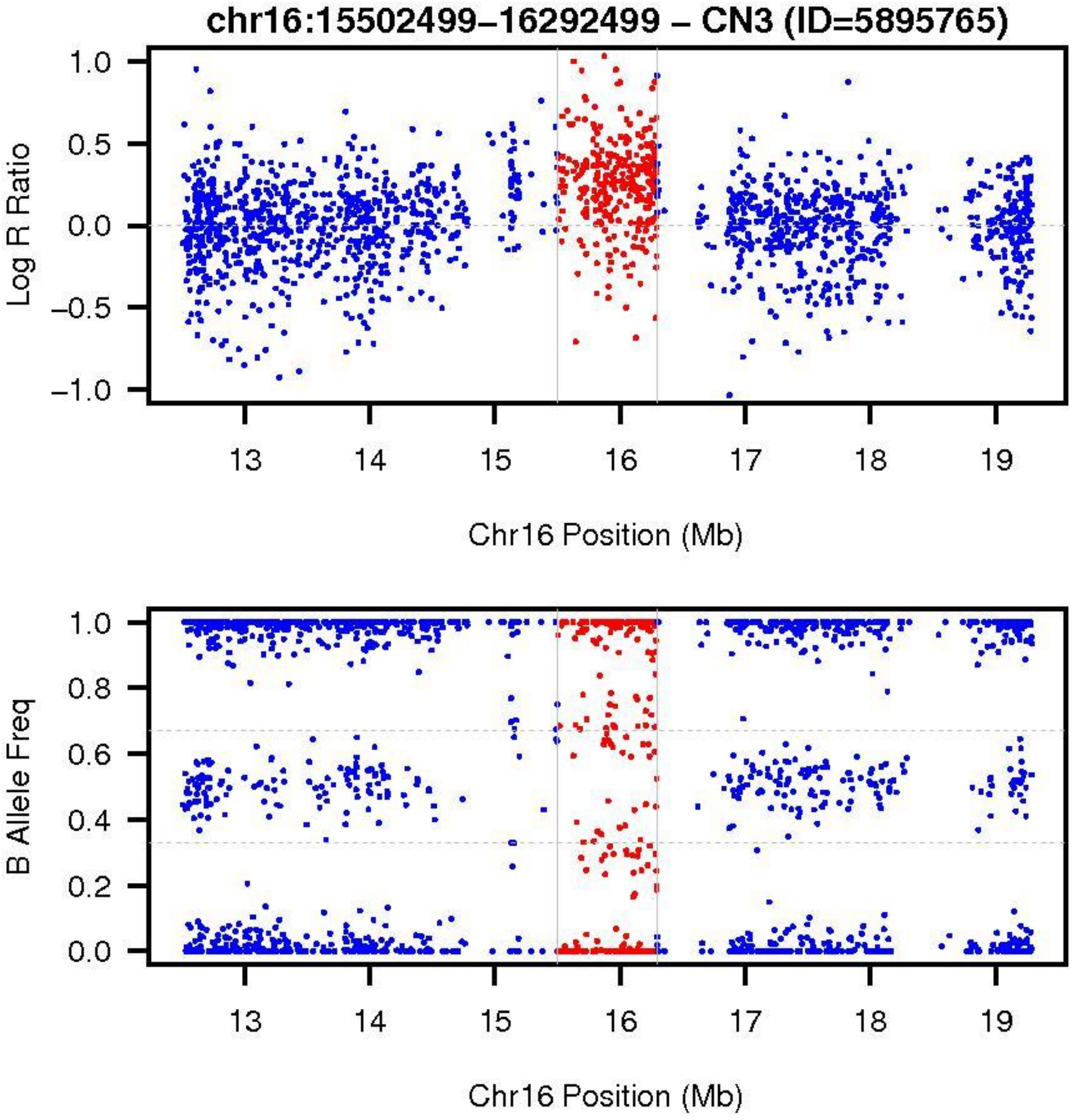

chr15:31132708–32482708 – CN1 (ID=5231063)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
28
30
32
34
Chr15 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
28
30
32
34
Chr15 Position (Mb)

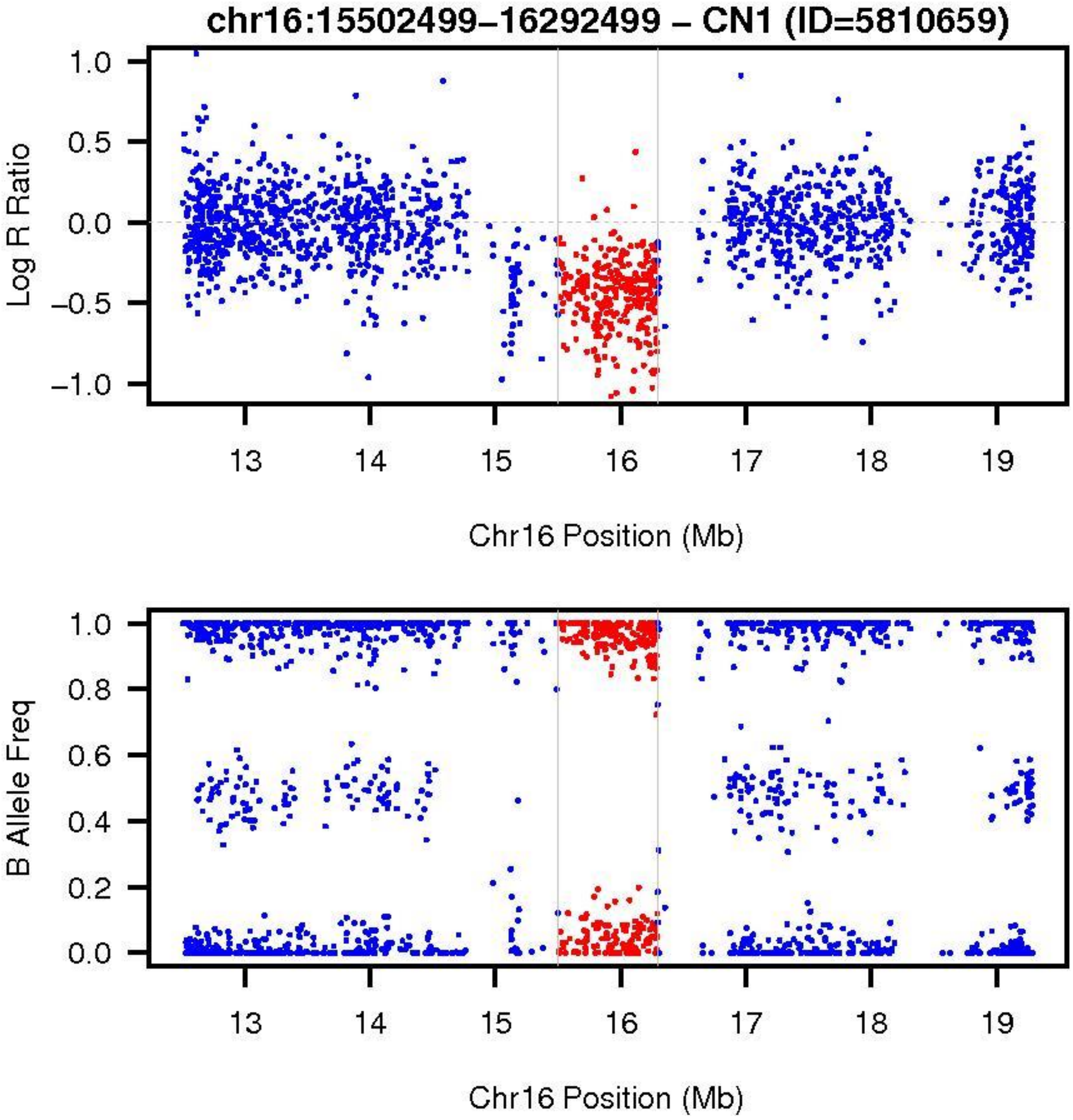

chr15:83182945–84738996 – CN3 (ID=2816413)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
80
82
84
86
88
Chr15 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
80
82
84
86
88
Chr15 Position (Mb)

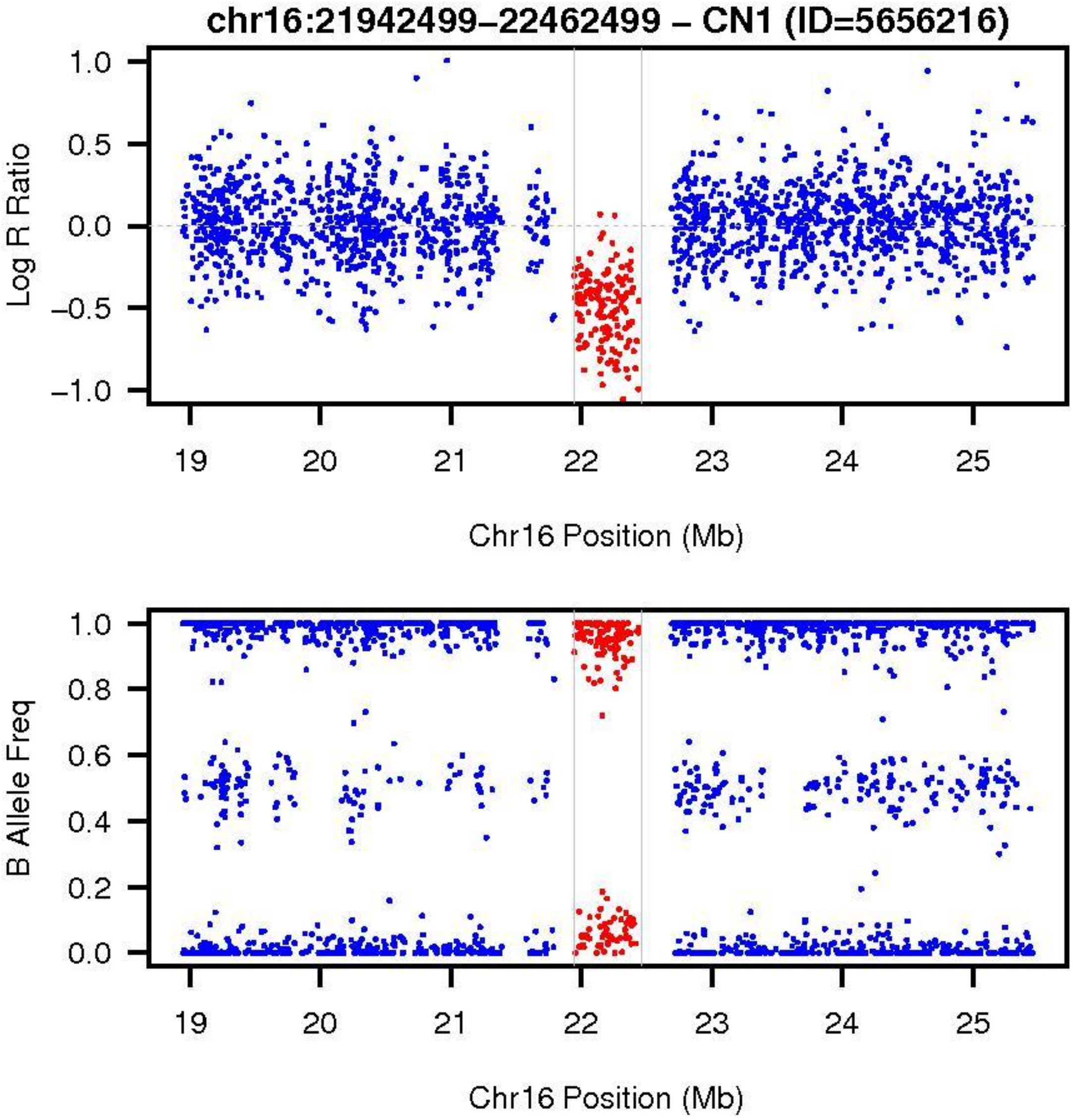

chr15:85138996–85698996 – CN1 (ID=5527390)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
82
83
84
85
86
87
88
Chr15 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
82
83
84
85
86
87
88
Chr15 Position (Mb)

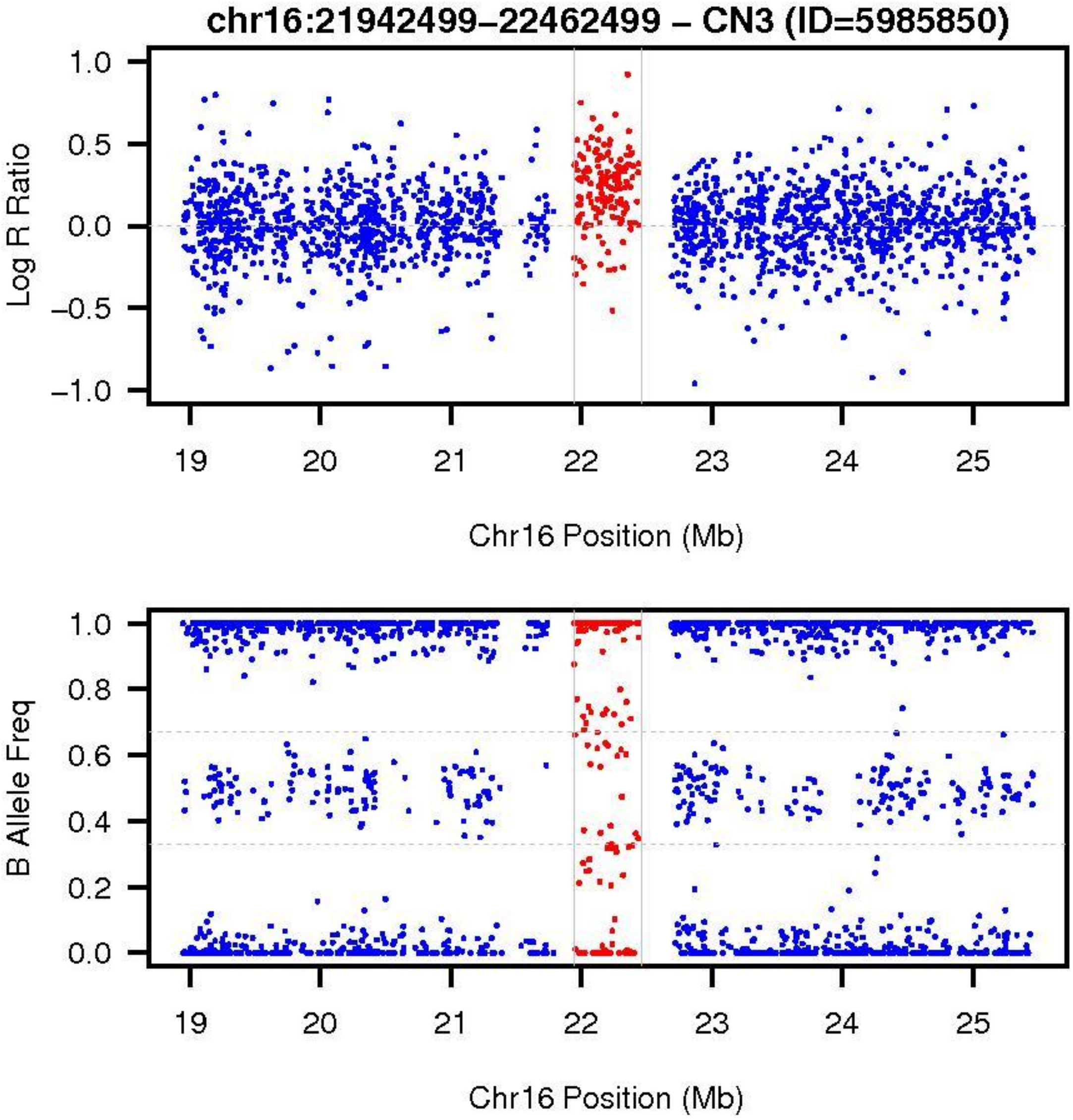

chr15:22798636–23088559 – CN3 (ID=5997983)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
20
21
22
23
24
25
26
Chr15 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
20
21
22
23
24
25
26
Chr15 Position (Mb)

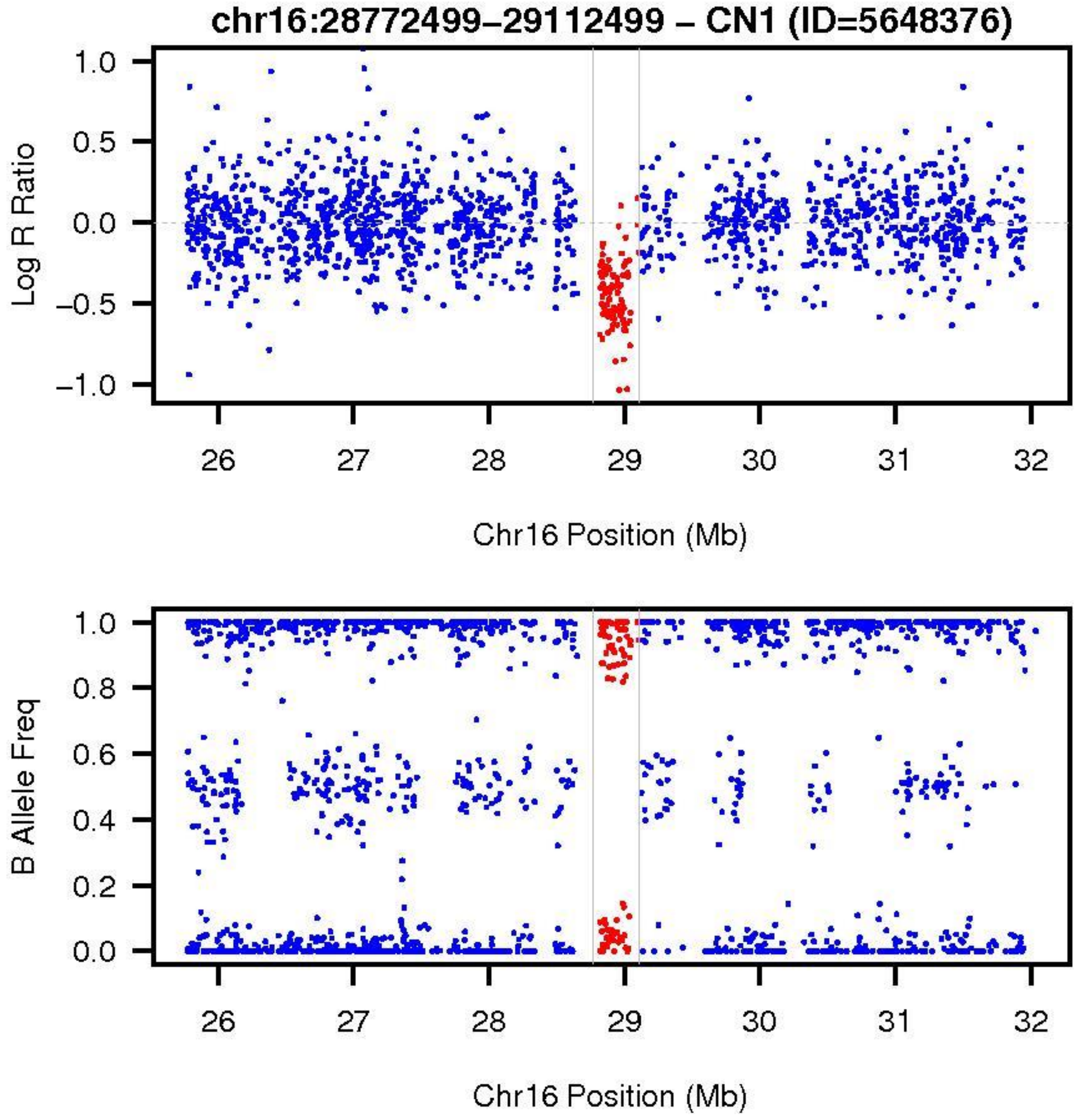

chr15:85138996–85698996 – CN3 (ID=5586290)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
82
83
84
85
86
87
88
Chr15 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
82
83
84
85
86
87
88
Chr15 Position (Mb)

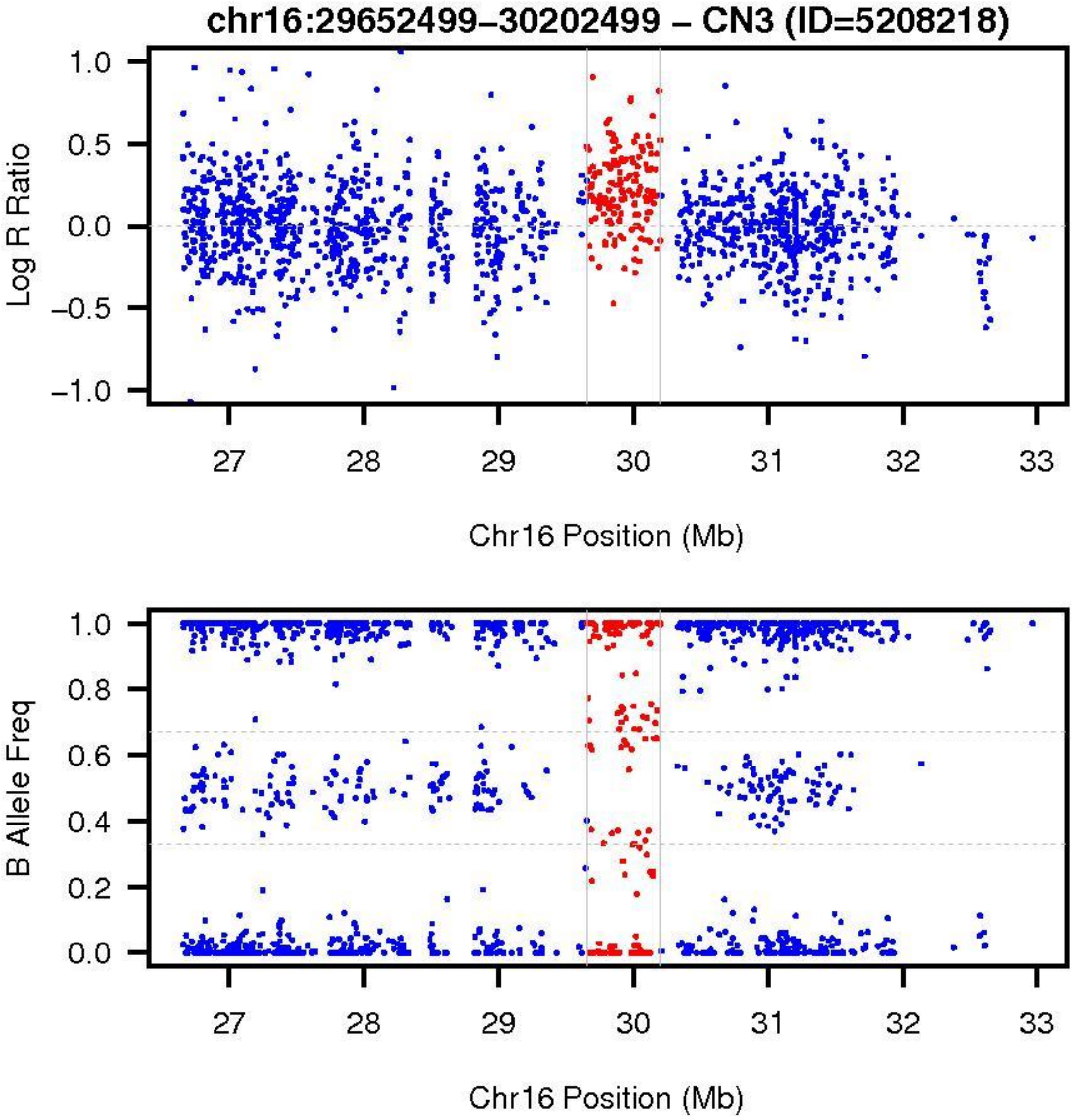

chr16:15502499–16292499 – CN3 (ID=5895765)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
13
14
15
16
17
18
19
Chr16 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
13
14
15
16
17
18
19
Chr16 Position (Mb)

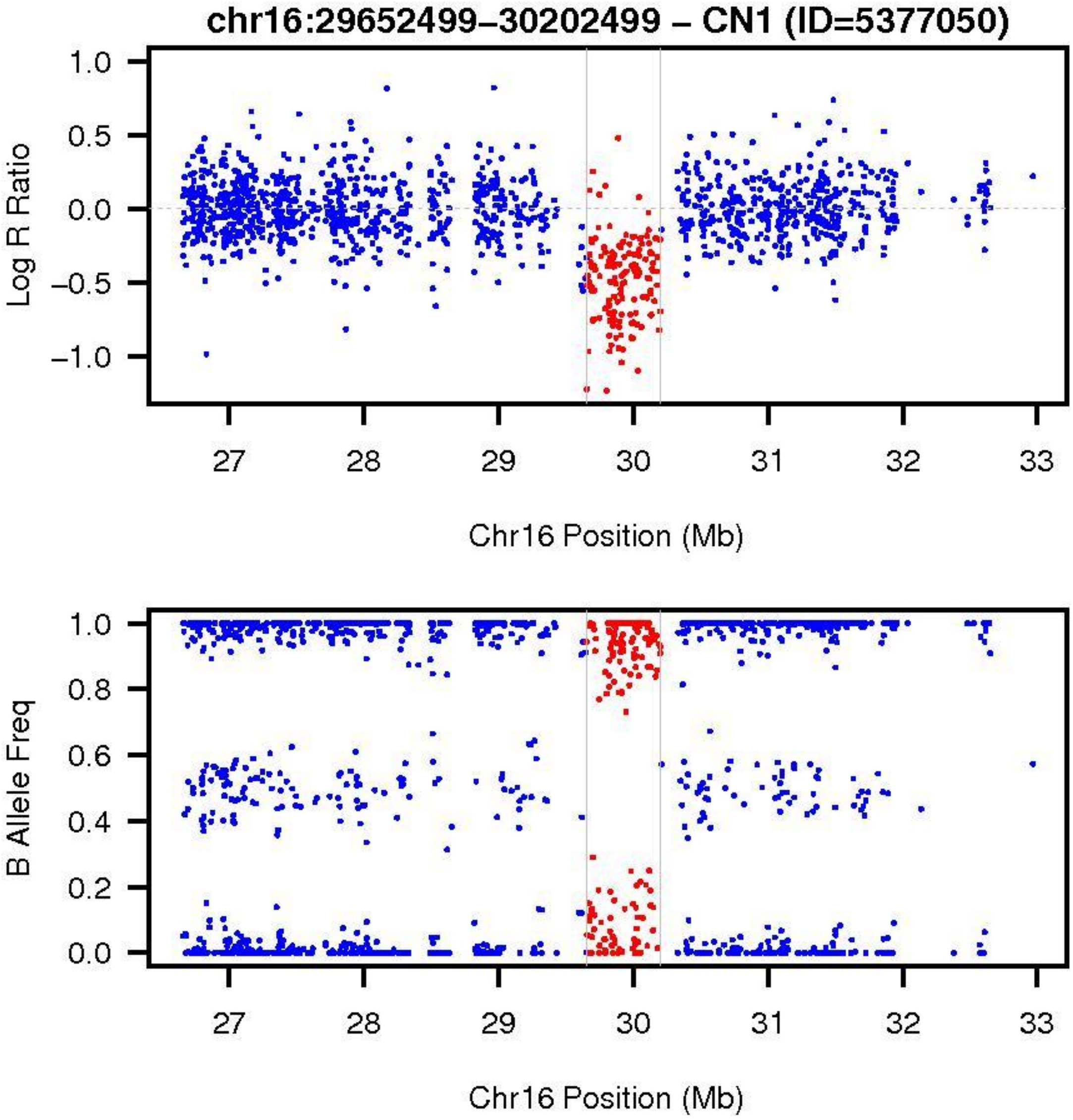

chr16:15502499–16292499 – CN1 (ID=5810659)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
13
14
15
16
17
18
19
Chr16 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
13
14
15
16
17
18
19
Chr16 Position (Mb)

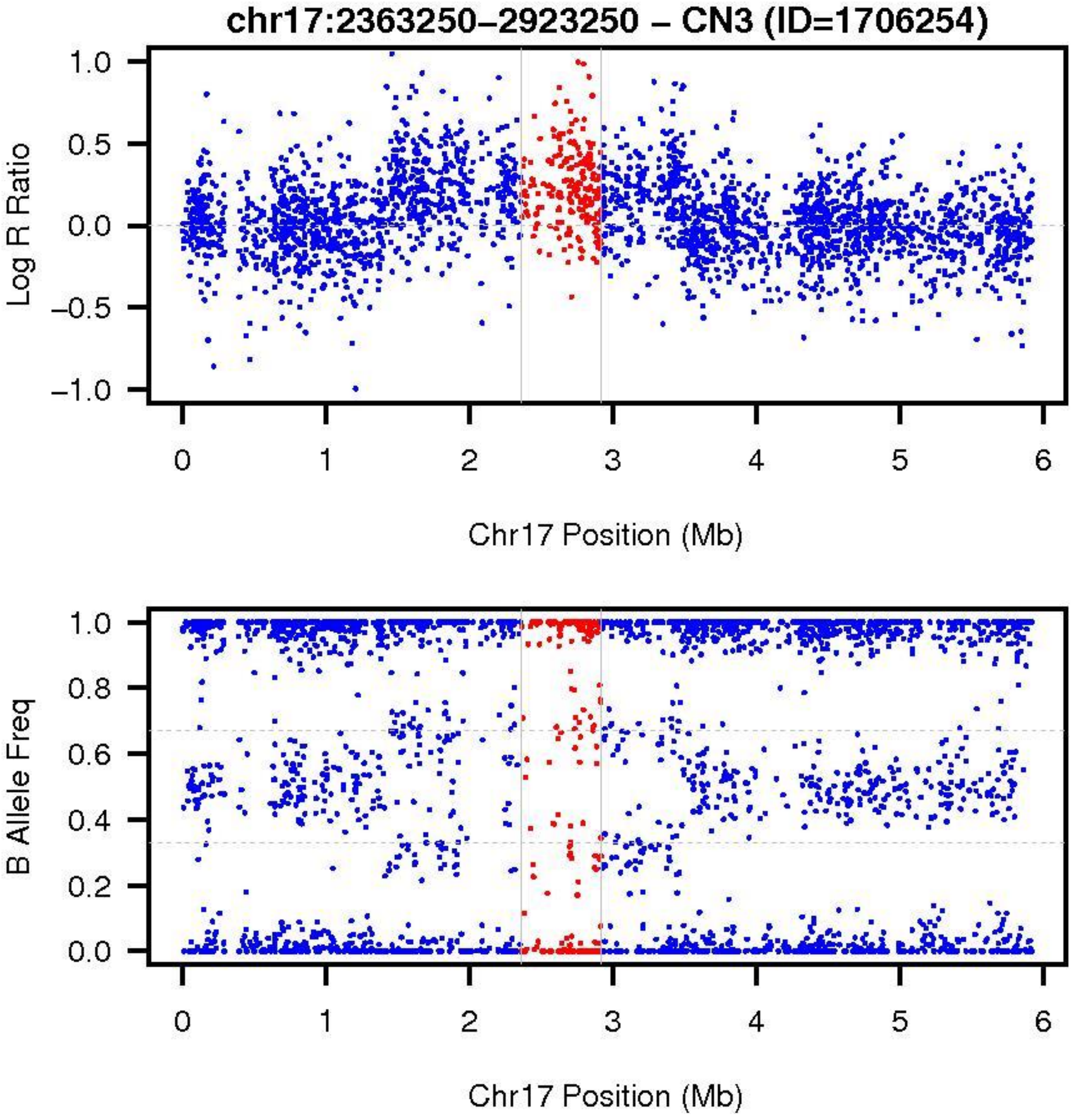

chr16:21942499–22462499 – CN1 (ID=5656216)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
19
20
21
22
23
24
25
Chr16 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
19
20
21
22
23
24
25
Chr16 Position (Mb)

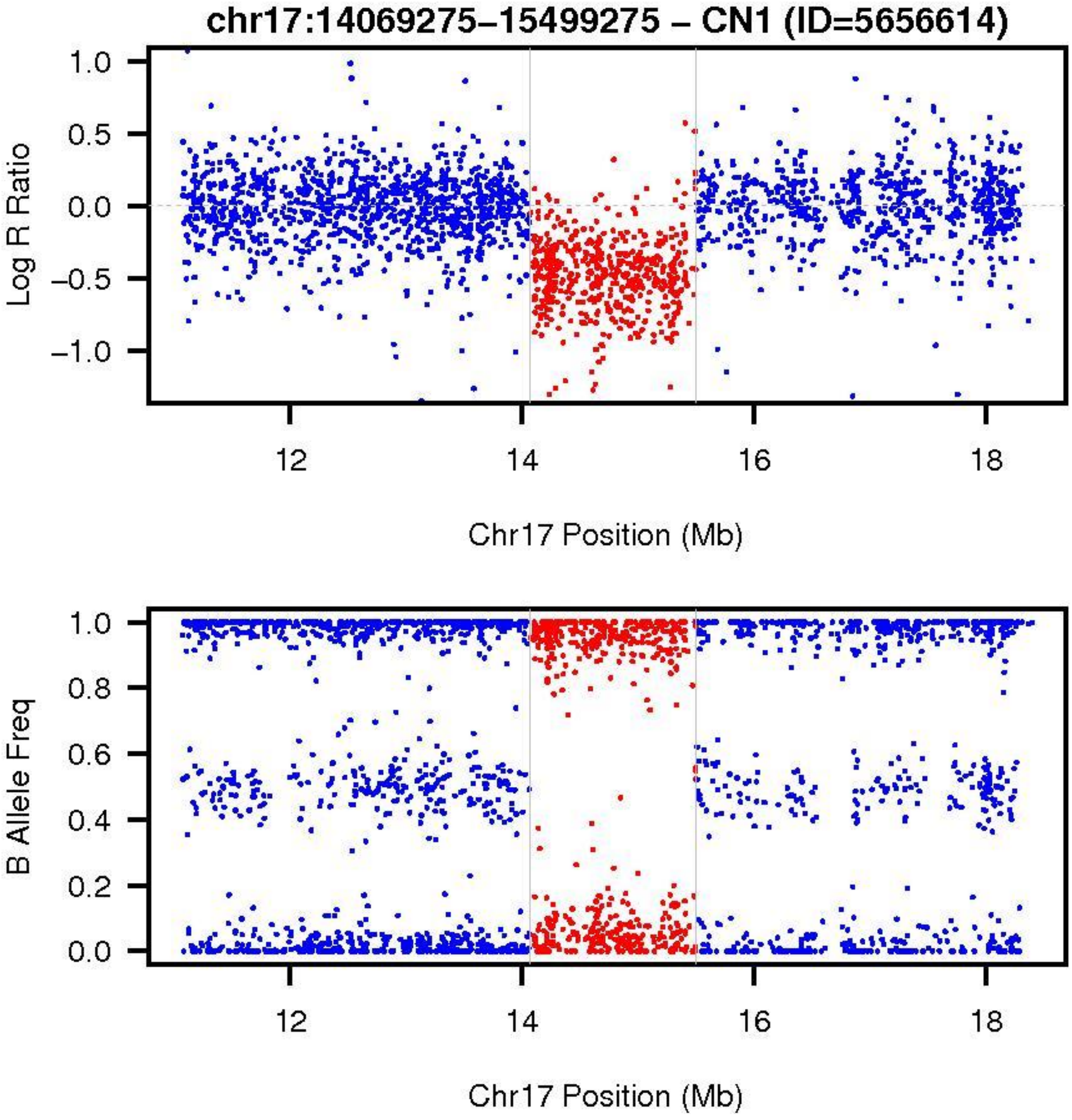

chr16:21942499–22462499 – CN3 (ID=5985850)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
19
20
21
22
23
24
25
Chr16 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
19
20
21
22
23
24
25
Chr16 Position (Mb)

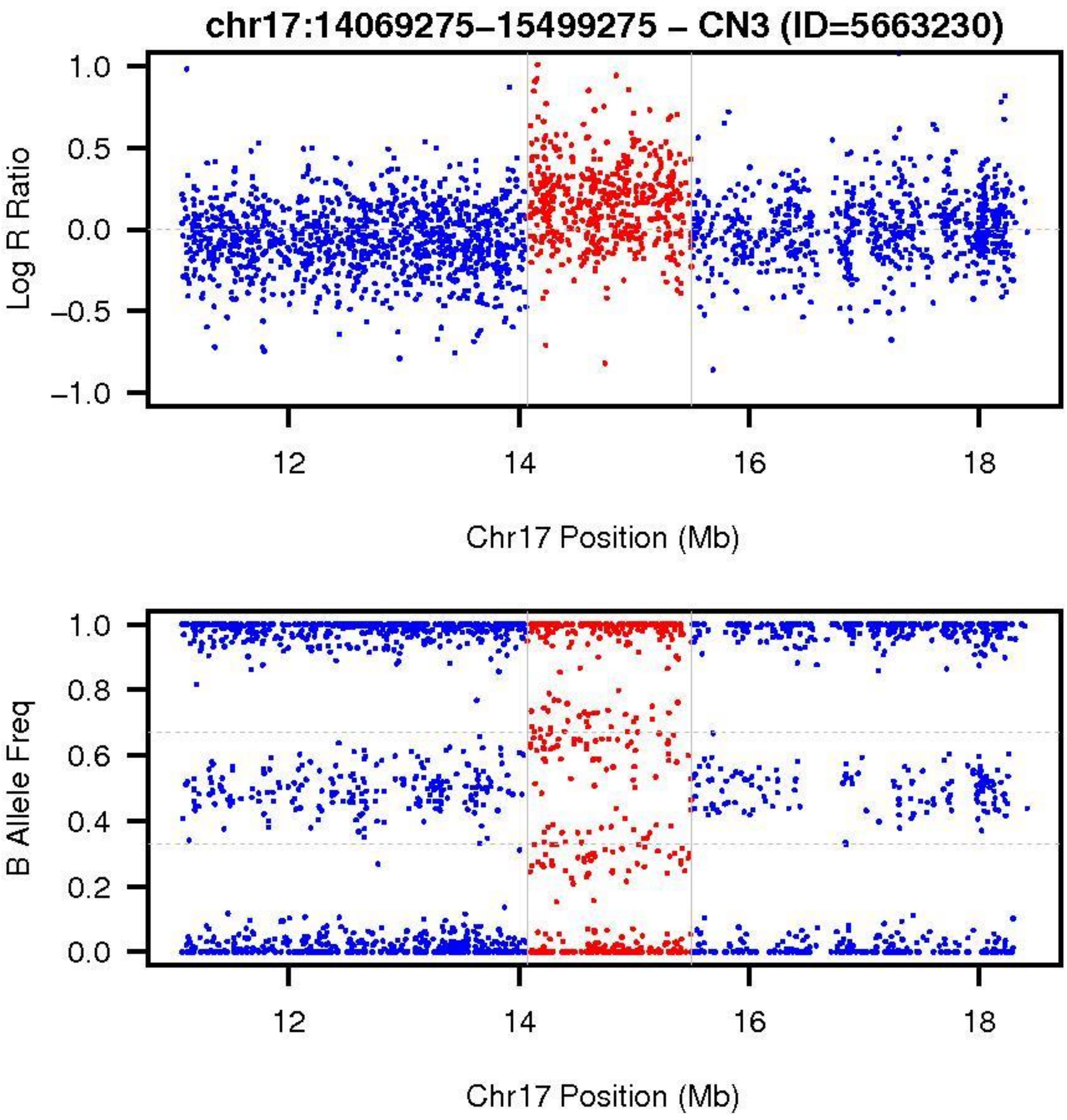

chr16:28772499–29112499 – CN1 (ID=5648376)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
26
27
28
29
30
31
32
Chr16 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
26
27
28
29
30
31
32
Chr16 Position (Mb)

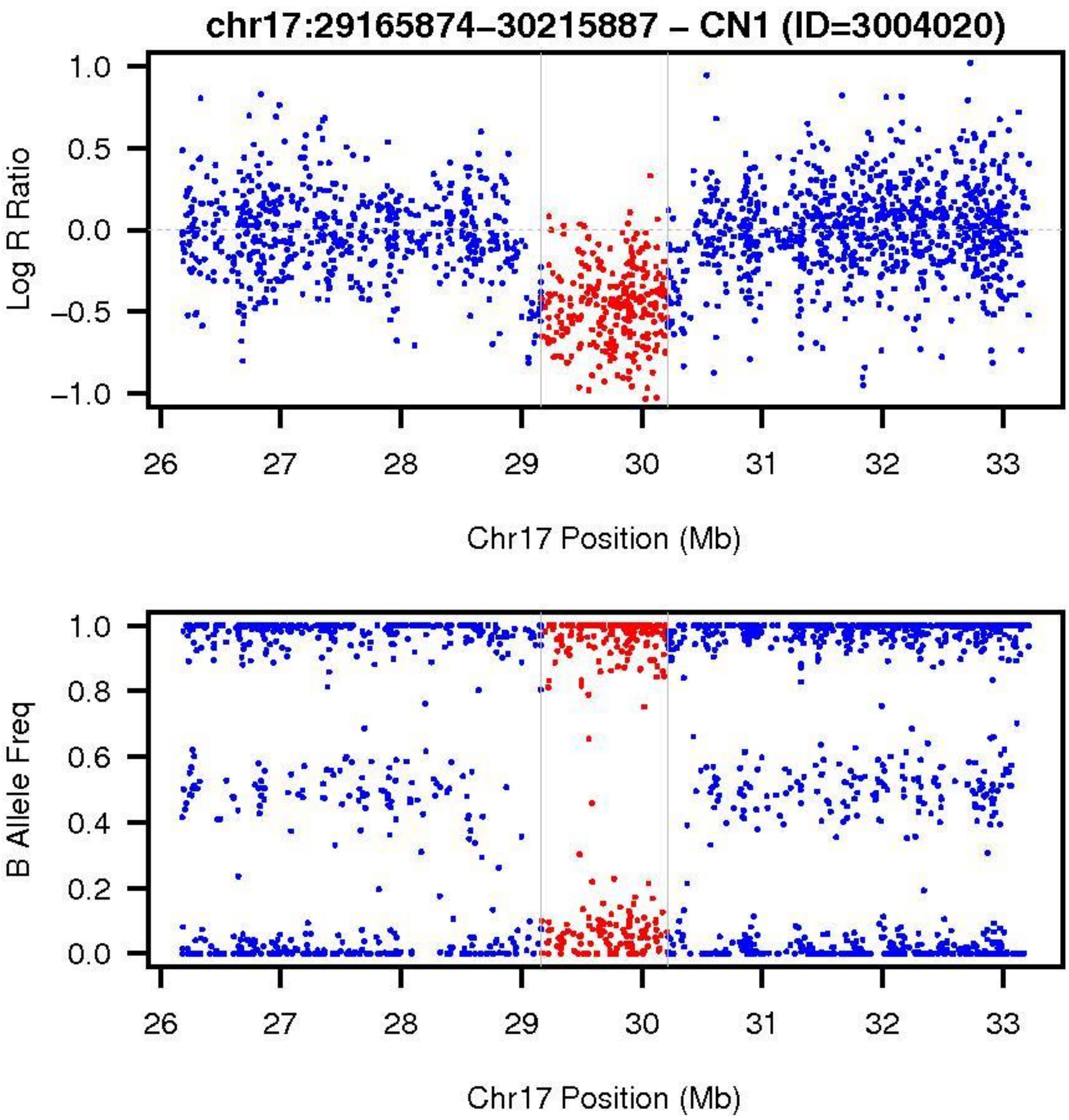

chr16:29652499–30202499 – CN3 (ID=5208218)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
27
28
29
30
31
32
33
Chr16 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
27
28
29
30
31
32
33
Chr16 Position (Mb)

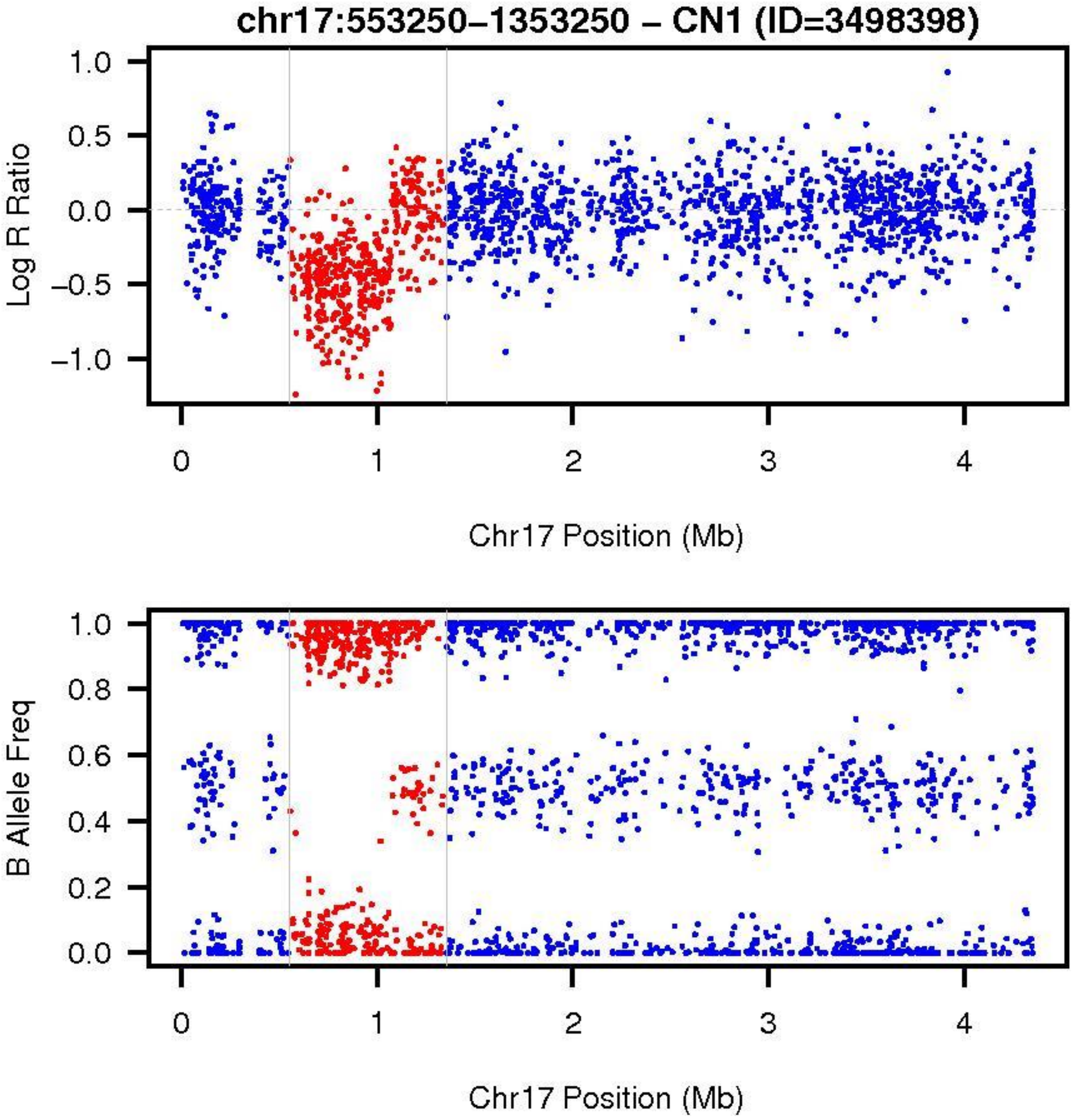

chr16:29652499–30202499 – CN1 (ID=5377050)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
27
28
29
30
31
32
33
Chr16 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
27
28
29
30
31
32
33
Chr16 Position (Mb)

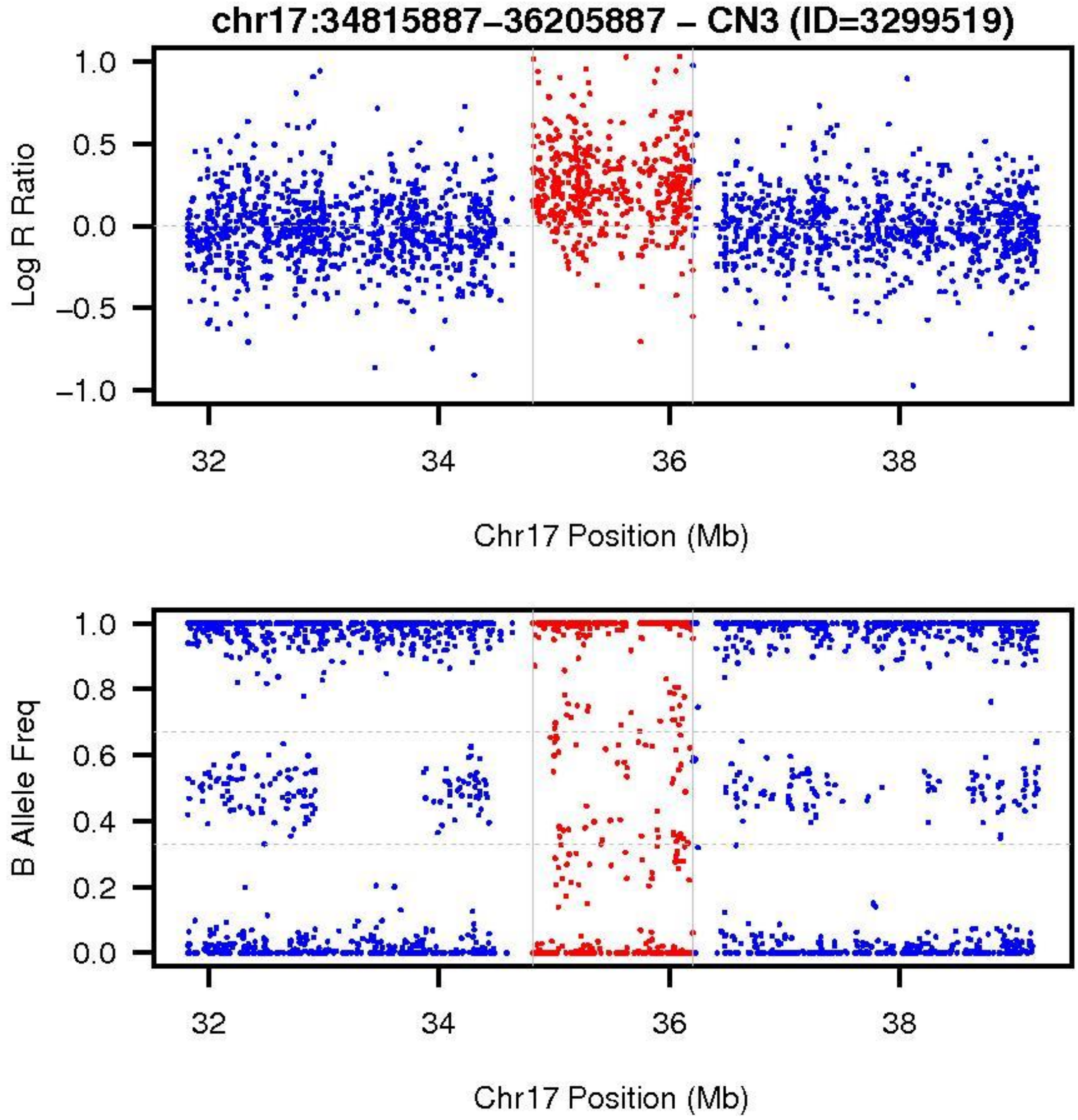

chr17:2363250–2923250 – CN3 (ID=1706254)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
0
1
2
3
4
5
6
Chr17 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
Chr17 Position (Mb)

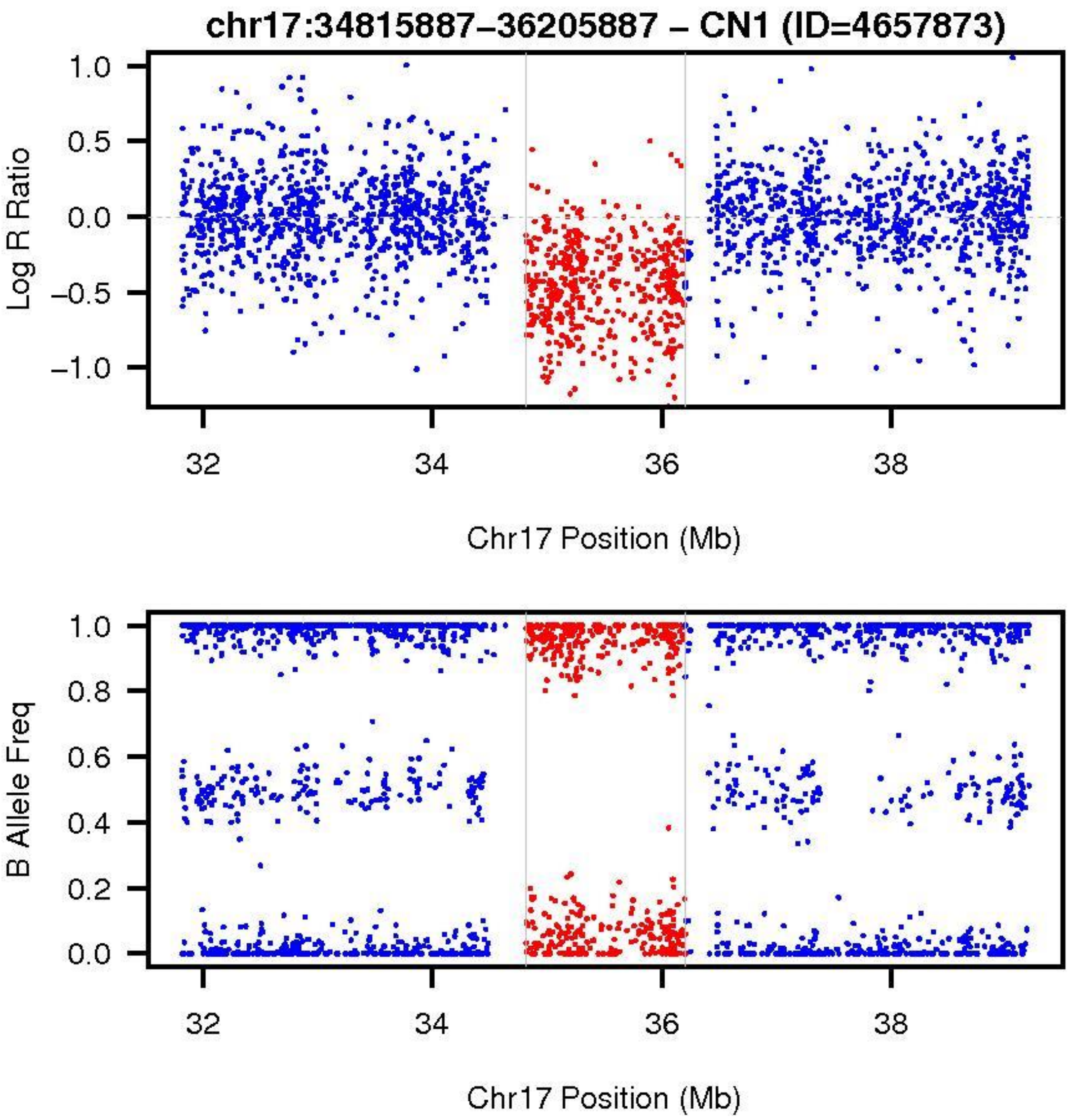

chr17:14069275–15499275 – CN1 (ID=5656614)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
12
14
16
18
Chr17 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
12
14
16
18
Chr17 Position (Mb)

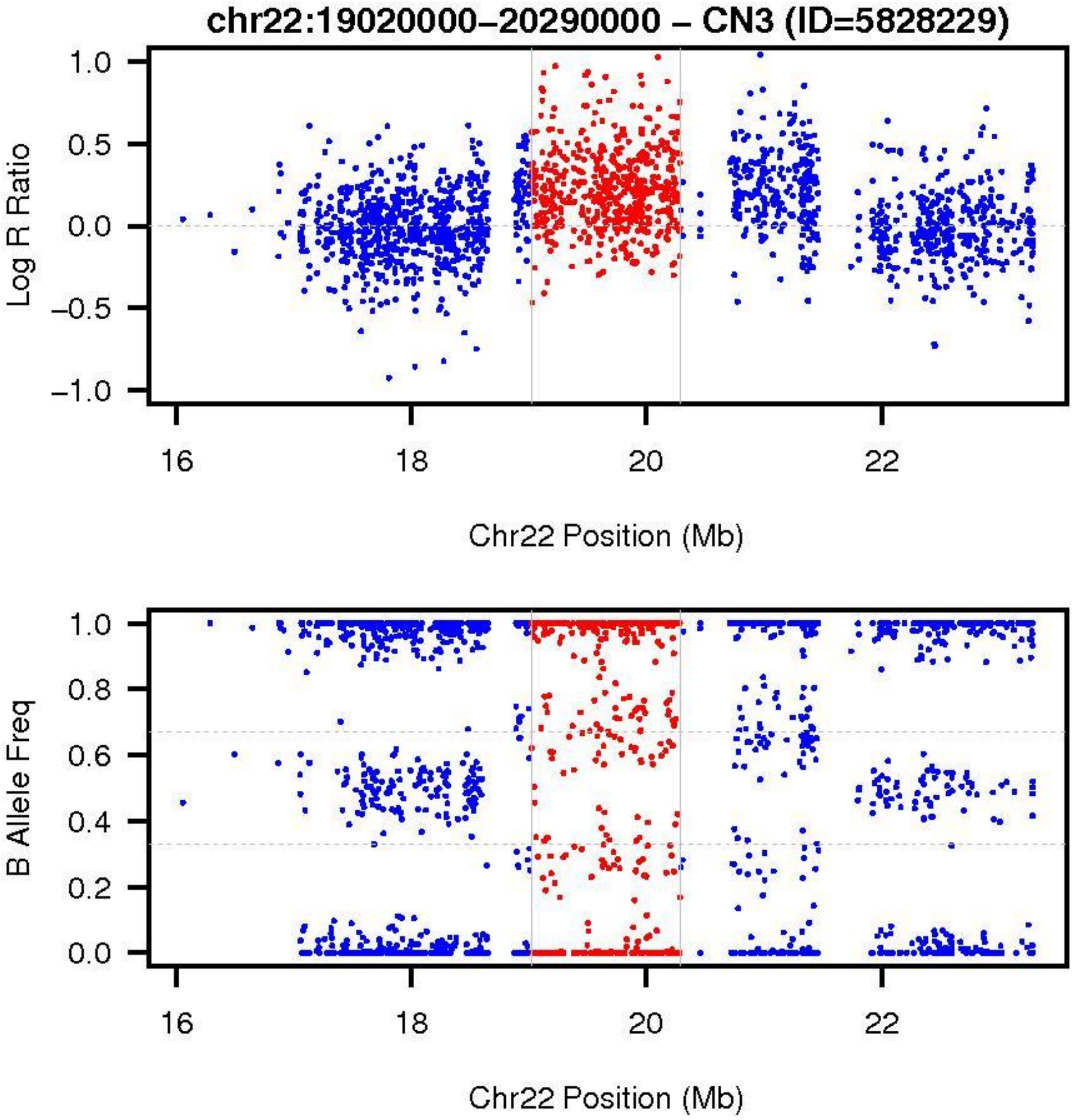

chr17:14069275–15499275 – CN3 (ID=5663230)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
12
14
16
18
Chr17 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
12
14
16
18
Chr17 Position (Mb)

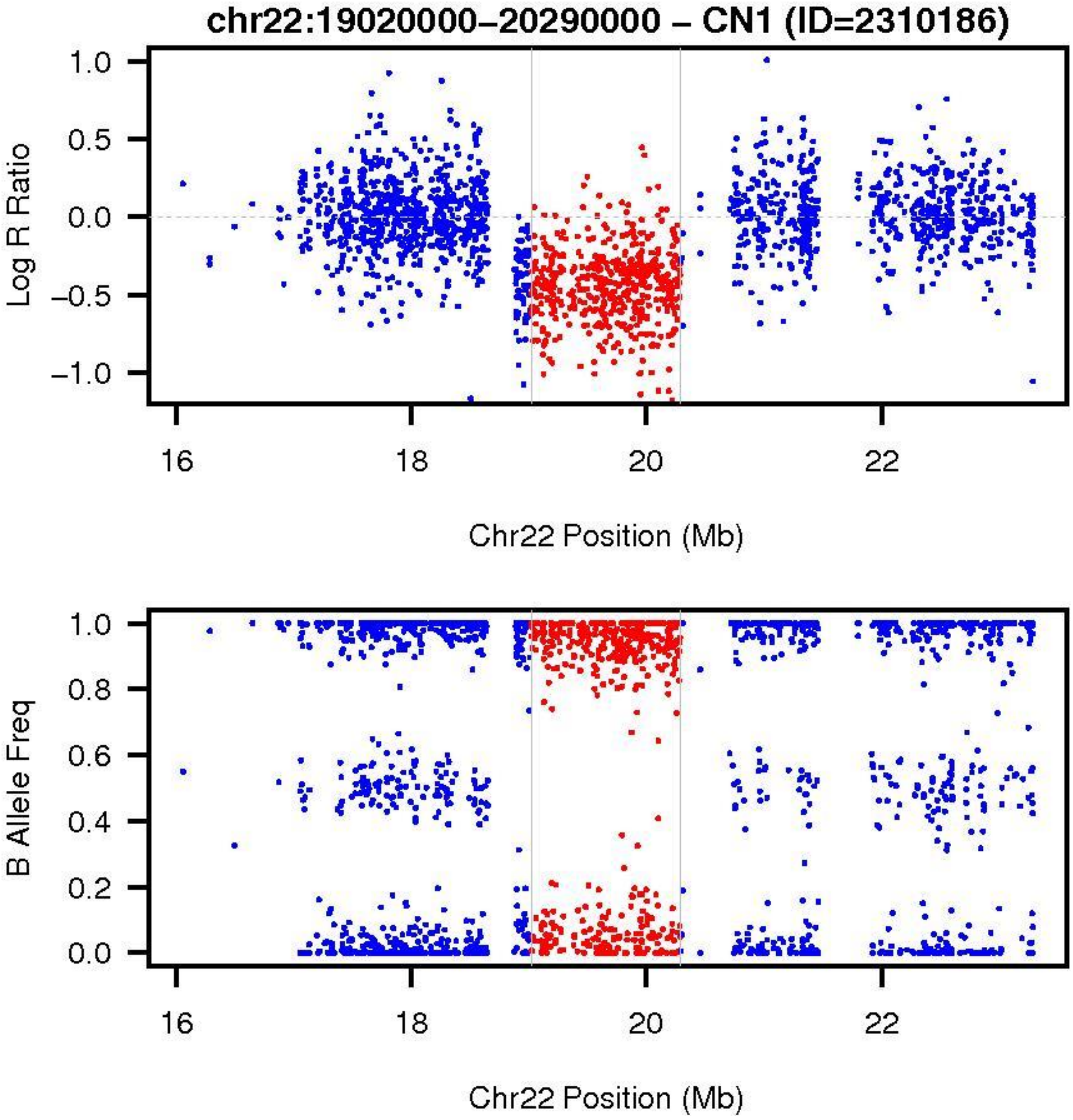

chr17:29165874–30215887 – CN1 (ID=3004020)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
26
27
28
29
30
31
32
33
Chr17 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
26
27
28
29
30
31
32
33
Chr17 Position (Mb)

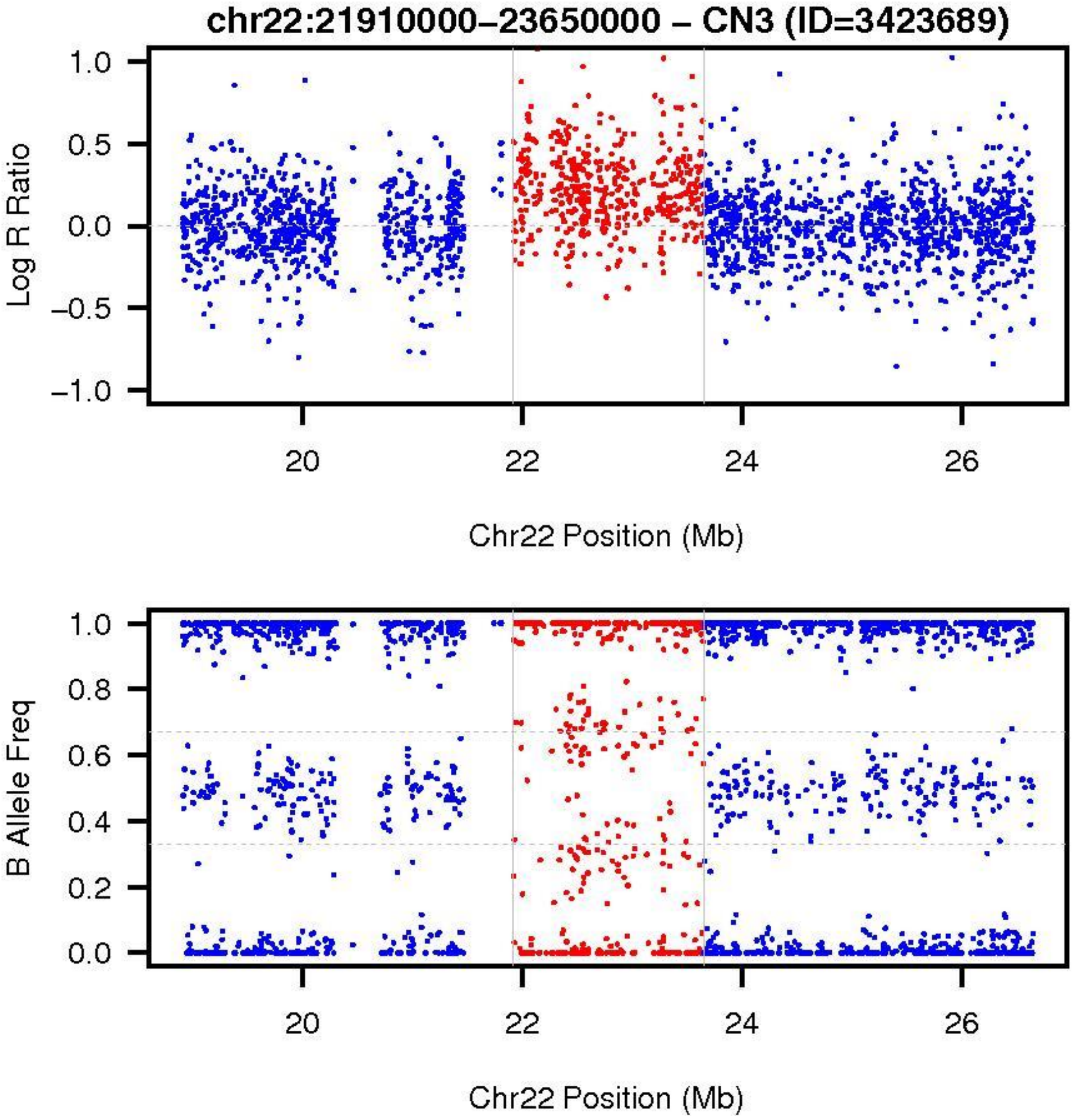

chr17:553250–1353250 – CN1 (ID=3498398)
Log R Ratio
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
Chr17 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
Chr17 Position (Mb)

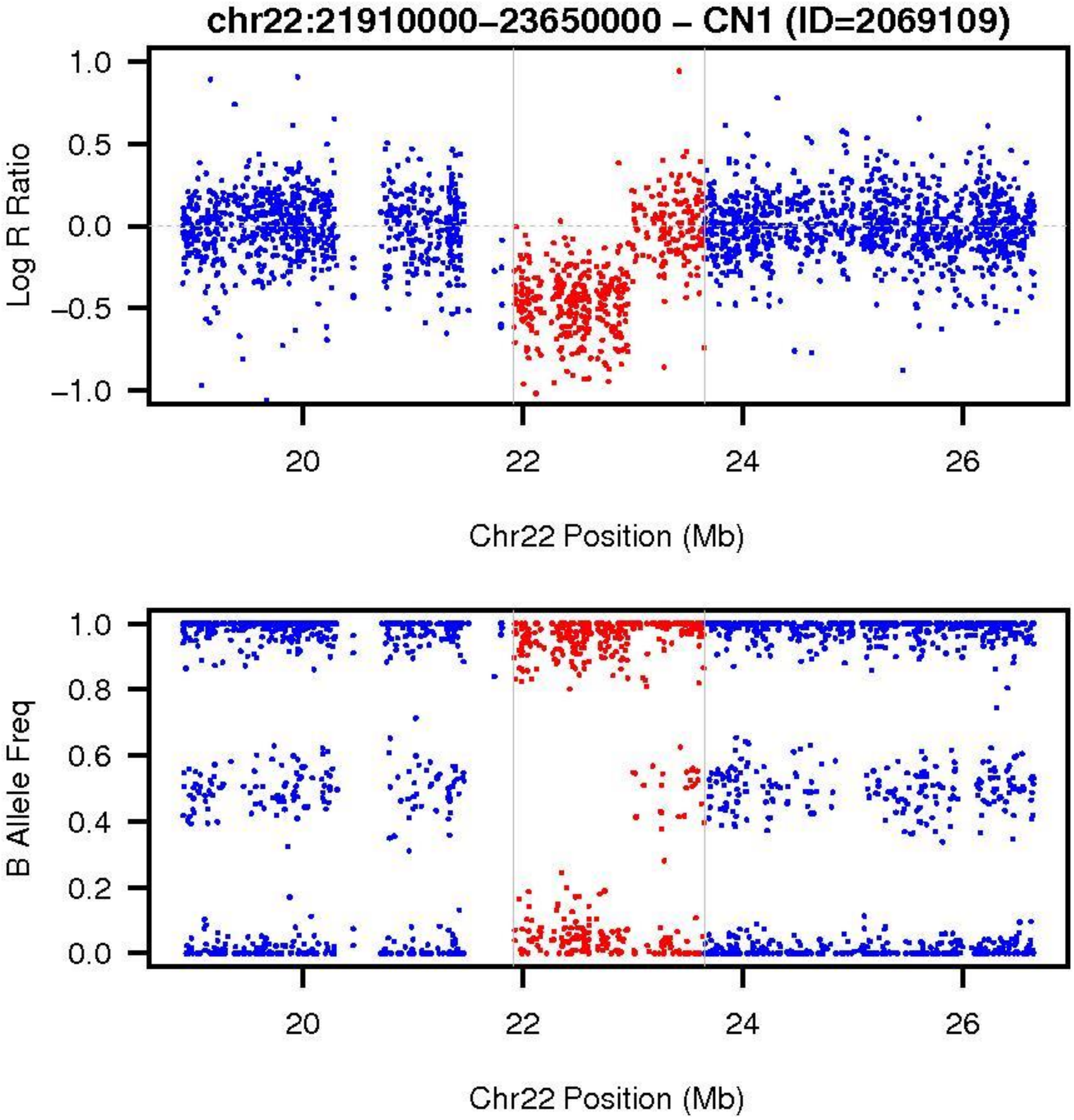

chr17:34815887–36205887 – CN3 (ID=3299519)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
32
34
36
38
Chr17 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
32
34
36
38
Chr17 Position (Mb)


chr17:34815887–36205887 – CN1 (ID=4657873)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
32
34
36
38
Chr17 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
32
34
36
38
Chr17 Position (Mb)


chr22:19020000–20290000 – CN3 (ID=5828229)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
16
18
20
22
Chr22 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
16
18
20
22
Chr22 Position (Mb)


chr22:19020000–20290000 – CN1 (ID=2310186)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
16
18
20
22
Chr22 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
16
18
20
22
Chr22 Position (Mb)


chr22:21910000–23650000 – CN3 (ID=3423689)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
20
22
24
26
Chr22 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
20
22
24
26
Chr22 Position (Mb)


chr22:21910000–23650000 – CN1 (ID=2069109)
Log R Ratio
1.0
0.5
0.0
–0.5
–1.0
20
22
24
26
Chr22 Position (Mb)
B Allele Freq
1.0
0.8
0.6
0.4
0.2
0.0
20
22
24
26
Chr22 Position (Mb)

APPENDIX III

List of publications

1. Yaghootkar, H., L. A. Lotta, J. Tyrrell, R. A. Smit, S. E. Jones, L. Donnelly, R. Beaumont, A. Campbell, **M. A. Tuke**, C. Hayward, K. S. Ruth, S. Padmanabhan, J. W. Jukema, C. C. Palmer, A. Hattersley, R. M. Freathy, C. Langenberg, N. J. Wareham, A. R. Wood, A. Murray, M. N. Weedon, N. Sattar, E. Pearson, R. A. Scott and T. M. Frayling (2016). "**Genetic evidence for a link between favorable adiposity and lower risk of type 2 diabetes, hypertension and heart disease.**" Diabetes.

Recent genetic studies have identified some alleles associated with higher BMI but lower risk of type 2 diabetes, hypertension and heart disease. These "favorable adiposity" alleles are collectively associated with lower insulin levels and higher subcutaneous-to-visceral adipose tissue ratio and may protect from disease through higher adipose storage capacity. We aimed to use data from 164,609 individuals from the UK Biobank and five other studies to replicate associations between a genetic score of 11 favorable adiposity variants and adiposity and risk of disease, test for interactions between BMI and favorable adiposity genetics and test effects separately in men and women.In the UK Biobank the 50% of individuals carrying the most favorable adiposity alleles had higher BMIs (0.120 Kg/m2 [0.066,0.174]; p=1E-5) and higher body fat percentage (0.301 % [0.230,0.372]; p=1E-16) compared to the 50% of individuals carrying the fewest alleles. For a given BMI, the 50% of individuals carrying the most favourable adiposity alleles were at: 0.837 OR [0.784,0.894] lower risk of type 2 diabetes (p=1E-7), -0.859 mmHg [-1.099,-0.618] lower systolic (p=3E-12) and -0.394 mmHg [-0.534,-0.254] lower diastolic blood pressure (p=4E-8), 0.935 OR [0.911,0.958] lower risk of hypertension (p=1E-7) and 0.921 OR [0.872,0.973] lower risk of heart disease (p=3E-3). In women, these associations could be explained by the observation that the alleles associated with higher BMI but lower risk of disease were also associated with a favourable body fat distribution, with a lower waist-hip ratio (-0.004 [-0.005,-0.003] 50% vs 50%; p=3E-14) but in men, the favourable adiposity alleles were associated with higher waist circumference (0.454 cm [0.267,0.641] 50% vs 50%; p=2E-6) and higher waist-hip ratio (0.0013 [0.0003,0.0024] 50% vs 50%; p=0.01). Results were strengthened when meta-analysing with five additional studies. There was no evidence of interaction between a genetic score consisting of known BMI variants and the favorable adiposity genetic score.In conclusion, different molecular mechanisms that lead to higher body fat percentage (with greater subcutaneous storage capacity) can have different impacts on cardiometabolic disease risk. While higher BMI is associated with higher risk of diseases, better fat storage capacity could reduce the risk.

2. Wood, A. R., J. Tyrrell, R. Beaumont, S. E. Jones, **M. A. Tuke**, K. S. Ruth, H. Yaghootkar, R. M. Freathy, A. Murray, T. M. Frayling and M. N. Weedon (2016). "**Variants in the FTO and CDKAL1 loci have recessive effects on risk of obesity and type 2 diabetes, respectively.**" Diabetologia 59(6): 1214-1221.

AIMS/HYPOTHESIS: Genome-wide association (GWA) studies have identified hundreds of common genetic variants associated with obesity and type 2 diabetes. These studies have usually focused on additive association tests. Identifying deviations from additivity may provide new biological insights and explain some of the missing heritability for these diseases. METHODS: We performed a GWA study using a dominance deviation model for BMI, obesity (29,925 cases) and type 2 diabetes (4,040 cases) in 120,286 individuals of British ancestry from the UK Biobank study. We also investigated whether single nucleotide polymorphisms previously shown to be associated with these traits showed any enrichment for departures from additivity. RESULTS: Known obesity-associated variants in FTO showed strong evidence of deviation from additivity (p DOMDEV = 3 x 10(-5)) through a recessive effect of the allele associated with higher BMI. The average BMI of individuals carrying zero, one or two BMI-raising alleles was 27.27 (95% CI 27.22, 27.31) kg/m(2), 27.54 (95% CI 27.50, 27.58) kg/m(2) and 28.07 (95% CI 28.00, 28.14) kg/m(2), respectively. A similar effect was observed in 105,643 individuals from the GIANT Consortium (p DOMDEV = 0.003; meta-analysis p DOMDEV = 1 x 10(-7)). For type 2 diabetes, we detected a recessive effect (p DOMDEV = 5 x 10(-4)) at CDKAL1. Relative to homozygous non-risk allele carriers, homozygous risk allele carriers had an OR of 1.48 (95% CI 1.32, 1.65), while the heterozygous group had an OR of 1.06 (95% CI 0.99, 1.14), a result consistent with that of a previous study. We did not identify any novel associations at genome-wide significance. CONCLUSIONS/INTERPRETATION: Although we found no evidence of widespread non-additive genetic effects contributing to obesity and type 2 diabetes risk, we did find robust examples of recessive effects at the FTO and CDKAL1 loci. ACCESS TO RESEARCH MATERIALS: Summary statistics are available at www.t2diabetesgenes.org and by request (a.r.wood@exeter.ac.uk). All underlying data are available on application from the UK Biobank.

3. Pilling, L. C., J. L. Atkins, K. Bowman, S. E. Jones, J. Tyrrell, R. N. Beaumont, K. S. Ruth, **M. A. Tuke**, H. Yaghootkar, A. R. Wood, R. M. Freathy, A. Murray, M. N. Weedon, L. Xue, K. Lunetta, J. M. Murabito, L. W. Harries, J. M. Robine, C. Brayne, G. A. Kuchel, L. Ferrucci, T. M. Frayling and D. Melzer (2016). "**Human longevity is influenced by many genetic variants: evidence from 75,000 UK Biobank participants.**" Aging (Albany NY) 8(3): 547-563.

Variation in human lifespan is 20 to 30% heritable in twins but few genetic variants have been identified. We undertook a Genome Wide Association Study (GWAS) using age at death of parents of middle-aged UK Biobank participants of European decent (n=75,244 with father's and/or mother's data, excluding early deaths). Genetic risk scores for 19 phenotypes (n=777 proven variants) were also tested. In GWAS, a nicotine receptor locus(CHRNA3, previously associated with increased smoking and lung cancer) was associated with fathers' survival. Less common variants requiring further confirmation were also identified. Offspring of longer lived parents had more protective alleles for coronary artery disease, systolic blood pressure, body mass index, cholesterol and triglyceride levels, type-1 diabetes, inflammatory bowel disease and Alzheimer's disease. In candidate analyses, variants in the TOMM40/APOE locus were associated with longevity, but FOXO variants were not. Associations between extreme longevity (mother >=98 years, fathers >=95 years, n=1,339) and disease alleles were similar, with an additional association with HDL cholesterol (p=5.7x10-3). These results support a multiple protective factors model influencing lifespan and longevity (top 1% survival) in humans, with prominent roles for cardiovascular-related pathways. Several of these genetically influenced risks, including blood pressure and tobacco exposure, are potentially modifiable.

4. Ruth, K. S., R. N. Beaumont, J. Tyrrell, S. E. Jones, **M. A. Tuke**, H. Yaghootkar, A. R. Wood, R. M. Freathy, M. N. Weedon, T. M. Frayling and A. Murray (2016). "**Genetic evidence that lower circulating FSH levels lengthen menstrual cycle, increase age at menopause and impact female reproductive health.**" Hum Reprod 31(2): 473-481.

STUDY QUESTION: How does a genetic variant in the FSHB promoter, known to alter FSH levels, impact female reproductive health? SUMMARY ANSWER: The T allele of the FSHB promoter polymorphism (rs10835638; c.-211G>T) results in longer menstrual cycles and later menopause and, while having detrimental effects on fertility, is protective against endometriosis. WHAT IS KNOWN ALREADY: The FSHB promoter polymorphism (rs10835638; c.-211G>T) affects levels of FSHB transcription and, as a result, circulating levels of FSH. FSH is required for normal fertility and genetic variants at the FSHB locus are associated with age at menopause and polycystic ovary syndrome (PCOS). STUDY DESIGN, SIZE, DURATION: We used cross-sectional data from the UK Biobank to look at associations between the FSHB promoter polymorphism and reproductive traits, and performed a genome-wide association study (GWAS) for length of menstrual cycle. PARTICIPANTS/MATERIALS, SETTING, METHODS: We included white British individuals aged 40-69 years in 2006-2010, in the May 2015 release of genetic data from UK Biobank. We tested the FSH-lowering T allele of the FSHB promoter polymorphism (rs10835638; c.-211G>T) for associations with 29, mainly female, reproductive phenotypes in up to 63 350 women and 56 608 men. We conducted a GWAS in 9534 individuals to identify genetic variants associated with length of menstrual cycle. MAIN RESULTS AND THE ROLE OF CHANCE: The FSH-lowering T allele of the FSHB promoter polymorphism (rs10835638; MAF 0.16) was associated with longer menstrual cycles [0.16 SD (c. 1 day) per minor allele; 95% confidence interval (CI) 0.12-0.20; P = 6 x 10(-16)], later age at menopause (0.13 years per minor allele; 95% CI 0.04-0.22; P = 5.7 x 10(-3)), greater female nulliparity [odds ratio (OR) = 1.06; 95% CI 1.02-1.11; P = 4.8 x 10(-3)] and lower risk of endometriosis (OR = 0.79; 95% CI 0.69-0.90; P = 4.1 x 10(-4)). The FSH-lowering T allele was not associated with other female reproductive illnesses or conditions in our study and we did not replicate associations with male infertility or PCOS. In the GWAS for menstrual cycle length, only variants near the FSHB gene reached genome-wide significance (P < 5 x 10(-9)). LIMITATIONS, REASONS FOR CAUTION: The data included might be affected by recall bias. Cycle length was not available for 25% of women still cycling (1% did not answer, 6% did not know and for 18% cycle length was recorded as 'irregular'). Women with a cycle length recorded were aged over 40 and were approaching menopause; however, we did not find evidence that this affected the results. Many of the groups with illnesses had relatively small sample sizes and so the study may have been under-powered to detect an effect. WIDER IMPLICATIONS OF THE FINDINGS: We found a strong novel association between a genetic variant that lowers FSH levels and longer menstrual cycles, at a locus previously robustly associated with age at menopause. The variant was also associated with nulliparity and endometriosis risk. These findings should now be verified in a second independent group of patients. We conclude that lifetime differences in circulating levels of FSH between individuals can influence menstrual cycle length and a range of reproductive outcomes, including menopause timing, infertility, endometriosis and PCOS. STUDY

FUNDING/COMPETING INTERESTS: None. TRIAL REGISTRATION NUMBER: Not applicable.

5. Tyrrell, J., S. E. Jones, R. Beaumont, C. M. Astley, R. Lovell, H. Yaghootkar, **M. Tuke**, K. S. Ruth, R. M. Freathy, J. N. Hirschhorn, A. R. Wood, A. Murray, M. N. Weedon and T. M. Frayling (2016). "**Height, body mass index, and socioeconomic status: mendelian randomisation study in UK Biobank.**" BMJ 352: i582.

OBJECTIVE: To determine whether height and body mass index (BMI) have a causal role in five measures of socioeconomic status. DESIGN: Mendelian randomisation study to test for causal effects of differences in stature and BMI on five measures of socioeconomic status. Mendelian randomisation exploits the fact that genotypes are randomly assigned at conception and thus not confounded by non-genetic factors. SETTING: UK Biobank. PARTICIPANTS: 119 669 men and women of British ancestry, aged between 37 and 73 years. MAIN OUTCOME MEASURES: Age completed full time education, degree level education, job class, annual household income, and Townsend deprivation index. RESULTS: In the UK Biobank study, shorter stature and higher BMI were observationally associated with several measures of lower socioeconomic status. The associations between shorter stature and lower socioeconomic status tended to be stronger in men, and the associations between higher BMI and lower socioeconomic status tended to be stronger in women. For example, a 1 standard deviation (SD) higher BMI was associated with a pound210 (euro276; $300; 95% confidence interval pound84 to pound420; P=6x10(-3)) lower annual household income in men and a pound1890 (pound1680 to pound2100; P=6x10(-15)) lower annual household income in women. Genetic analysis provided evidence that these associations were partly causal. A genetically determined 1 SD (6.3 cm) taller stature caused a 0.06 (0.02 to 0.09) year older age of completing full time education (P=0.01), a 1.12 (1.07 to 1.18) times higher odds of working in a skilled profession (P=6x10(-7)), and a pound1130 (pound680 to pound1580) higher annual household income (P=4x10(-8)). Associations were stronger in men. A genetically determined 1 SD higher BMI (4.6 kg/m(2)) caused a pound2940 (pound1680 to pound4200; P=1x10(-5)) lower annual household income and a 0.10 (0.04 to 0.16) SD (P=0.001) higher level of deprivation in women only. CONCLUSIONS: These data support evidence that height and BMI play an important partial role in determining several aspects of a person's socioeconomic status, especially women's BMI for income and deprivation and men's height for education, income, and job class. These findings have important social and health implications, supporting evidence that overweight people, especially women, are at a disadvantage and that taller people, especially men, are at an advantage.

6. Usher, C. L., R. E. Handsaker, T. Esko, **M. A. Tuke**, M. N. Weedon, A. R. Hastie, H. Cao, J. E. Moon, S. Kashin, C. Fuchsberger, A. Metspalu, C. N. Pato, M. T. Pato, M. I. McCarthy, M. Boehnke, D. M. Altshuler, T. M. Frayling, J. N. Hirschhorn and S. A. McCarroll (2015). "**Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity.**" Nat Genet 47(8): 921-925.

Hundreds of genes reside in structurally complex, poorly understood regions of the human genome. One such region contains the three amylase genes (AMY2B, AMY2A and AMY1) responsible for digesting starch into sugar. Copy number of AMY1 is reported to be the largest genomic influence on obesity, although genome-wide association studies for obesity have found this locus unremarkable. Using whole-genome sequence analysis, droplet digital PCR and genome mapping, we identified eight common structural haplotypes of the amylase locus that suggest its mutational history. We found that the AMY1 copy number in an individual's genome is generally even (rather than odd) and partially correlates with nearby SNPs, which do not associate with body mass index (BMI). We measured amylase gene copy number in 1,000 obese or lean Estonians and in 2 other cohorts totaling approximately 3,500 individuals. We had 99% power to detect the lower bound of the reported effects on BMI, yet found no association.

7. Wood, A. R.*, **M. A. Tuke***, M. Nalls, D. Hernandez, J. R. Gibbs, H. Lin, C. S. Xu, Q. Li, J. Shen, G. Jun, M. Almeida, T. Tanaka, J. R. Perry, K. Gaulton, M. Rivas, R. Pearson, J. E. Curran, M. P. Johnson, H. H. Goring, R. Duggirala, J. Blangero, M. I. McCarthy, S. Bandinelli, A. Murray, M. N. Weedon, A. Singleton, D. Melzer, L. Ferrucci and T. M. Frayling (2015). "Whole-genome sequencing to understand the genetic architecture of common gene expression and biomarker phenotypes." Hum Mol Genet 24(5): 1504-1512.

Initial results from sequencing studies suggest that there are relatively few low-frequency (<5%) variants associated with large effects on common phenotypes. We performed low-pass whole-genome sequencing in 680 individuals from the InCHIANTI study to test two primary hypotheses: (i) that sequencing would detect single low-frequency-large effect variants that explained similar amounts of phenotypic variance as single common variants, and (ii) that some common variant associations could be explained by low-frequency variants. We tested two sets of disease-related common phenotypes for which we had statistical power to detect large numbers of common variant-common phenotype associations-11 132 cis-gene expression traits in 450 individuals and 93 circulating biomarkers in all 680 individuals. From a total of 11 657 229 high-quality variants of which 6 129 221 and 5 528 008 were common and low frequency (<5%), respectively, low frequency-large effect associations comprised 7% of detectable cis-gene expression traits [89 of 1314 cis-eQTLs at P < 1 x 10(-06) (false discovery rate approximately 5%)] and one of eight biomarker associations at P < 8 x 10(-10). Very few (30 of 1232; 2%) common variant associations were fully explained by low-frequency variants. Our data show that whole-genome sequencing can identify low-frequency variants undetected by genotyping based approaches when sample sizes are sufficiently large to detect substantial numbers of common variant associations, and that common variant associations are rarely explained by single low-frequency variants of large effect.

8. Wood, A. R., **M. A. Tuke**, M. A. Nalls, D. G. Hernandez, S. Bandinelli, A. B. Singleton, D. Melzer, L. Ferrucci, T. M. Frayling and M. N. Weedon (2014). "**Another explanation for apparent epistasis.**" Nature 514(7520): E3-5.

Epistasis occurs when the effect of a genetic variant on a trait is dependent on genotypes of other variants elsewhere in the genome. Hemani et al. recently reported the detection and replication of many instances of epistasis between pairs of variants influencing gene expression levels in humans1. Using whole-genome sequencing data from 450 individuals we strongly replicated many of the reported interactions but, in each case, a single third variant captured by our sequencing data could explain all of the apparent epistasis. Our results provide an alternative explanation for the apparent epistasis observed for gene expression in humans.

9. Morrison, F. S., J. M. Locke, A. R. Wood, **M. Tuke**, D. Pasko, A. Murray, T. Frayling and L. W. Harries (2013). "**The splice site variant rs11078928 may be associated with a genotype-dependent alteration in expression of GSDMB transcripts.**" BMC Genomics 14: 627.

BACKGROUND: Many genetic variants have been associated with susceptibility to complex traits by genome wide association studies (GWAS), but for most, causal genes and mechanisms of action have yet to be elucidated. Using bioinformatics, we identified index and proxy variants associated with autoimmune disease susceptibility, with the potential to affect splicing of candidate genes. PCR and sequence analysis of whole blood RNA samples from population controls was then carried out for the 8 most promising variants to determine the effect of genetic variation on splicing of target genes. RESULTS: We identified 31 splice site SNPs with the potential to affect splicing, and prioritised 8 to determine the effect of genotype on candidate gene splicing. We identified that variants rs11078928 and rs2014886 were associated with altered splicing of the GSDMB and TSFM genes respectively. rs11078928, present in the asthma and autoimmune disease susceptibility locus on chromosome 17q12-21, was associated with the production of a novel Delta exon5-8 transcript of the GSDMB gene, and a separate decrease in the percentage of transcripts with inclusion of exon 6, whereas the multiple sclerosis susceptibility variant rs2014886, was associated with an alternative TFSM transcript encompassing a short cryptic exon within intron 2. CONCLUSIONS: Our findings demonstrate the utility of a bioinformatic approach in identification and prioritisation of genetic variants effecting splicing of their host genes, and suggest that rs11078928 and rs2014886 may affect the splicing of the GSDMB and TSFM genes respectively.