Title
Copy Number Variation in Neuropsychiatric Disorders

Author
Huang, Alden Yen-Wen

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Copy Number Variation in Neuropsychiatric Disorders

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Bioinformatics

by

Alden Yen-Wen Huang

2018

ABSTRACT OF THE DISSERTATION

Copy Number Variation in Neuropsychiatric Disorders

by

Alden Yen-Wen Huang

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2018

Professor Giovanni Coppola, Chair

In this thesis, I characterize the contribution of rare copy number variation (CNV) to the genetic etiology of Tourette syndrome (TS) and bipolar disorder (BP). I accomplish this using several different study designs and various methods for CNV detection. As array data was widely available for the majority of samples evaluated, I make extensive use of this technology throughout this project and first provide an overview of the technical challenges involved and describe the analytical pipeline I developed to produce reliable CNV calls from such data.

Then, in the largest TS CNV study conducted to date, I report the discovery of the first two genome-wide significant CNVs associated with the disorder, and demonstrate an increased global burden of large, singleton events and CNVs at known, pathogenic loci. Conditioned on this latter observation, I perform an exploratory analysis aimed at gene discovery through the identification of *de novo* copy number variants from whole-exome sequencing in a sample of affected proband, unaffected parent trios in TS.

I then describe a CNV study of 26 large, multigenerational families with a high incidence of BP from two population isolates, using both microarray and whole-genome sequencing data. While thorough examination of these extended BP pedigrees revealed no segregating variants of large effect, I observe a significant increase in CNV burden across a subset of BP-related genes. Finally, I explore this notion further in a larger North American sample of unrelated individuals ascertained for BP. I demonstrate that although BP cases show no observable differences in the rate or size of rare CNVs, case CNVs affect more highly constrained, brain-expressed genes, and I provide evidence for an increased female CNV burden for BP.

The dissertation of Alden Yen-Wen Huang is approved.

Jeremiah Scharf

Bogdan Pasaniuc

Nelson B. Freimer

Eleazar Eskin

Giovanni Coppola, Committee Chair

University of California, Los Angeles

2018

To my family.

Table of Contents

List of Figures and Tables

Acknowledgements

I am grateful for all the individuals who contributed to my post-graduate work. I am particularly indebted to my mentor, Dr. Giovanni Coppola, for not only first giving me the opportunity to conduct research, but also ensuring that I always had the resources to do so properly. From his guidance, support, and patience, I have grown tremendously both intellectually and as a person. I would also like to thank Dr. Jeremiah Scharf, Dr. Jae Hoon Sul, Dr. Carol Mathews, Dongmei Yu, and the rest of the TSAICG collaborators. The bulk of what I have learned about genetics and academic life in general has been garnered through the numerous interactions between us.

I would also like to acknowledge the other members of my committee, Drs. Nelson Freimer, Eleazar Eskin, and Bogdan Pasaniuc; the body of your collective works has been a source of inspiration. I would also like to thank Dr. Christopher J. Evans, for the quality of his character and stimulating discussions on opioids.

My time as a graduate student has certainly been more enjoyable in the company of the other members of the Coppola lab, many of whom have been with me for most of my doctoral studies. I will miss seeing all of you, day in and day out.

Finally, I am most grateful for family. I wish to thank my parents, sister, and wife, for their love and support.

In Chapter 1, I adapted a figure from a co-authored published paper entitled "Transcriptome Profiling of Peripheral Blood in 22q11.2 Deletion Syndrome Reveals Functional Pathways Related to Psychosis and Autism Spectrum Disorder", by Maria Jalbrzikowski et al. Carrie E. Bearden is the corresponding author for this study. I designed the custom aCGH assay, processed the data, and generated Figure 1, which was incorporated into this thesis.

The work described in Chapter 3 is adapted from a published paper entitled “Rare Copy Number Variants in *NRXN1* and *CNTN6* Increase Risk for Tourette Syndrome”, by Alden Y. Huang et al. Peristera Pachou, Carol A. Mathews, Jeremiah M. Scharf, and Giovanni Coppola are senior authors of this study. We would like to thank Michael C. O’Donovan and Michael Owen for graciously providing us with additional control samples.

Chapter 5 is adapted from a manuscript in preparation, tentatively entitled “Contribution of common and rare variants to bipolar disorder susceptibility in extended pedigrees from population isolates”, by Jae Hoon Sul, et al. Nelson B. Freimer is the senior author of this study. Jae Hoon Sul processed the WGS data for SNVs/indels, performed the linkage analysis, and devised the segregation analysis. Susan K. Service processed the SNP data, performed the linkage analysis, performed the PRS analysis, and contributed to statistical aspects of the manuscript. I processed both SNP array and WGS data specifically for CNV calling, performed the burden analysis of rare CNVs, and assisted with the segregation analysis of rare CNVs.

Chapter 6 represents unpublished work. We thank James Knowles, Carlos M. Pato, Michele T. Pato, and Jordan Smoller of the GPC for providing us with the intensity data.

VITA

Education

2008	B.S. Mathematics/Applied Science, UCLA

Employment

2016	Teaching Assistant Department of Molecular Cellular and Developmental Biology, UCLA
2009 – 2013	Staff Research Associate Department of Neurology, UCLA
2007 – 2010	Staff Scientist Proteinlinks, Inc., Pasadena, California
2004 – 2006	Undergraduate Researcher Department of Pediatrics/Endocrinology, UCLA

Publications

Peer Reviewed

Chen JA, Fears SC, Jasinska AJ, **Huang A**, Al-Sharif NB, Scheibel KE, Dyer TD, Fagan AM, Blangero J, Woods R, Jorgensen MJ, Kaplan JR, Freimer NB & Coppola G (2018) Neurodegenerative disease biomarkers Aβ1-40, Aβ1-42, tau, and p-tau181 in the vervet monkey cerebrospinal fluid: Relation to normal aging, genetic influences, and cerebral amyloid angiopathy. *Brain Behav.* **8:** e00903

Fontes MM, Guvenek A, Kawaguchi R, Zheng D, **Huang A**, Ho VM, Chen PB, Liu X, O'Dell TJ, Coppola G, Tian B & Martin KC (2017) Activity-Dependent Regulation of Alternative Cleavage and Polyadenylation During Hippocampal Long-Term Potentiation. *Sci. Rep.* **7**: 17377

Huang AY, Yu D, Davis LK, Sul JH, Tsetsos F, Ramensky V, Zelaya I, Ramos EM, Osiecki L, Chen JA, McGrath LM, Illmann C, Sandor P, Barr CL, Grados M, Singer HS, Nöthen MM, Hebebrand J, King RA, Dion Y, et al (2017) Rare Copy Number Variants in NRXN1 and CNTN6 Increase Risk for Tourette Syndrome. *Neuron* **94:** 1101–1111.e7

Willsey AJ, Fernandez TV, Yu D, King RA, Dietrich A, Xing J, Sanders SJ, Mandell JD, **Huang AY**, Richer P, Smith L, Dong S, Samocha KE, Tourette International Collaborative Genetics (TIC Genetics), Tourette Syndrome Association International Consortium for Genetics (TSAICG), Neale BM, Coppola G, Mathews CA, Tischfield JA, Scharf JM, et al (2017) De Novo Coding Variants Are Strongly Associated with Tourette Disorder. *Neuron* **94:** 486–499.e9

Jalbrzikowski M, Lazaro MT, Gao F, **Huang A**, Chow C, Geschwind DH, Coppola G & Bearden CE (2015) Transcriptome Profiling of Peripheral Blood in 22q11.2 Deletion Syndrome Reveals Functional Pathways Related to Psychosis and Autism Spectrum Disorder. *PLoS One* **10:** e0132542

Chen JA, Wang Q, Davis-Turak J, Li Y, Karydas AM, Hsu SC, Sears RL, Chatzopoulou D, **Huang AY**, Wojta KJ, Klein E, Lee J, Beekly DL, Boxer A, Faber KM, Haase CM, Miller J, Poon WW, Rosen A, Rosen H, et al (2015) A multiancestral genome-wide exome array study of Alzheimer disease, frontotemporal dementia, and progressive supranuclear palsy. *JAMA Neurol.* **72:** 414–422

Hsu SC, Sears RL, Lemos RR, Quintáns B, **Huang A**, Spiteri E, Nevarez L, Mamah C, Zatz M, Pierce KD, Fullerton JM, Adair JC, Berner JE, Bower M, Brodaty H, Carmona O, Dobricić V, Fogel BL, García-Estevez D, Goldman J, et al (2013) Mutations in SLC20A2 are a major cause of familial idiopathic basal ganglia calcification. *Neurogenetics* **14:** 11–22

Coppola G, Chinnathambi S, Lee JJ, Dombroski BA, Baker MC, Soto-Ortolaza AI, Lee SE, Klein E, **Huang AY**, Sears R, Lane JR, Karydas AM, Kenet RO, Biernat J, Wang L-S, Cotman CW, Decarli CS, Levey AI, Ringman JM, Mendez MF, et al (2012) Evidence for a role of the rare p.A152T variant in MAPT in increasing the risk for FTD-spectrum and Alzheimer's diseases. *Hum. Mol. Genet.* **21:** 3500–3512

Fogel BL, Lee JY, Lane J, Wahnich A, Chan S, **Huang A**, Osborn GE, Klein E, Mamah C, Perlman S, Geschwind DH & Coppola G (2012) Mutations in rare ataxia genes are uncommon causes of sporadic cerebellar ataxia. *Mov. Disord.* **27:** 442–446

Russell JF, Steckley JL, Coppola G, Hahn AFG, Howard MA, Kornberg Z, **Huang A**, Mirsattari SM, Merriman B, Klein E, Choi M, Lee H-Y, Kirk A, Nelson-Williams C, Gibson G, Baraban SC, Lifton RP, Geschwind DH, Fu Y-H & Ptáček LJ (2012) Familial cortical myoclonus with a mutation in NOL3. *Ann. Neurol.* **72:** 175–183

Miyako K, Cobb LJ, Francis M, **Huang A**, Peng B, Pintar JE, Ariga H & Cohen P (2009) PAPA-1 Is a nuclear binding partner of IGFBP-2 and modulates its growth-promoting actions. *Mol. Endocrinol.* **23:** 169–175

Submitted

Huang AY, Taylor AMW, Ghogha A, Pribadi M, Wang Q, Kim TSJ, Coppola G, Evans CJ (2017) Genetic and functional analysis of a Pacific Hagfish opioid system.

Pribadi M, Yang Z, Kim TS, Swartz EW, **Huang AY**, Chen JA, Dokuru D, Baek J, Gao F, Fua AT, Wojta K, Wang Q, Lezcano E, Ng S, Chehab FF, Karydas A, Fong J, Vinters HV, Miller BL & Coppola G (2016) CRISPR-Cas9 targeted deletion of the C9orf72 repeat expansion mutation corrects cellular phenotypes in patient-derived iPS cells. *bioRxiv*: 051193

In Preparation

Sul JH, Service SK, **Huang AY**, Coppola G et al (2018) Contribution of common and rare variants to bipolar disorder susceptibility in extended pedigrees from population isolates.

Chapter 1:

Introduction

1.1 The landscape of human genetic variation

Following the completion of the human reference genome, there has been, in my view, no single scientific outcome of greater consequence to the field of human genetics than the resultant effort of the International HapMap Project (Abelson et al., 2005; International HapMap, 2003). From this and other earlier surveys of human variation, it became apparent that the individual genomes of any two given individuals differ, on average, at a rate of about 1 in every 300 bases at specific positions throughout the genome. The second phase of this collaborative effort characterized over 3.1 million single nucleotide polymorphisms (SNPs) across individuals from diverse geographic populations, about a third of the 10 million common SNPs that had been identified at the time (International HapMap et al., 2007). Most importantly, the knowledge gained helped fuel the rapid development of accessible, low-cost, high-throughput commercial assays to genotype these common variants across a large numbers of individuals enabling genome-wide association studies (GWAS) to be performed at scale.

In part due to sheer numbers, SNPs have long been considered as the primary source of human phenotypic variation. However, in a landmark study published in 2004, Sebat et al. utilized the hybridization intensity information from SNP arrays in 20 individuals to *quantify*, as opposed to simply type, the individual alleles at each polymorphic site throughout the genome (Sebat et al., 2004). Their findings closely mirrored those of study published in the same year by Iafrate et al., and revealed that variations in diploid copy number (copy number variants, or CNVs) were much more widespread among normal human individuals than previously anticipated (Iafrate et al., 2004). We now know that, although present in far fewer numbers, the amount of human genetic variation attributable to CNVs is roughly equal that of SNPs when measured in terms of base pairs. However, with the advancement of technology, particularly next-generation sequencing (NGS), a more complete picture of the overall landscape of human genetic variation continues to develop.

1.2 Overview of methods for genome-wide detection of copy number variation

Karyotyping was the first technique that allowed for the detection of gross chromosomal abnormalities genome-wide. The process for producing the “classical” karyotype image involves the staining of metaphase-arrested chromosomes from a cell culture sample and fixing them to slides for direct observation under a microscope (Barcia, 2007). It is not high throughput and severely limited in resolution. However, since abnormalities are visualized on the level of individual cells, karyotyping can unequivocally identify instances of mosaicism, and remains particularly useful for the evaluation of cell lines (Figure 1-1A). To improve upon the resolution limitations of traditional karyotyping, comparative genomic hybridization (CGH) was developed. In CGH, the sample of interest and a control sample are first differentially labeled with fluorescence and mixed in equal proportions before hybridization to a normal set of metaphase chromosomes fixed to a slide (Kallioniemi et al., 1992). As the name implies, CNVs are determined through comparison of the relative intensities of the two differently-labeled DNA samples.

Array-based comparative genomic hybridization (aCGH) was the first technology to specifically enable the detection of submicroscopic CNVs on a genome-wide scale in a high-throughput fashion. In contrast to traditional CGH, labeled DNA samples are instead hybridized to an array spotted with clones of known DNA fragments. Early versions of the technology utilized spots composed of bacterial artificial chromosome or cDNA clones. This practice has largely been supplanted by the refinement of the human reference genome alongside technical improvements in the production of long synthesized oligonucleotides. Modern aCGH platforms that utilize synthesized oligonucleotide probes allow for fully customized designs with the placement of millions of features, and offer a dramatic increase in resolution with accurate detection rates down to the kilobase-range (Conrad et al., 2009; Greshock et al., 2007). I have effectively used this technology to distinguish between different genetic forms of the 22q11.2

deletion that causes DiGeorge syndrome (Jalbrzikowski et al., 2015) (Figure 1-1B). aCGH is often considered the “gold standard” of high-throughput, genome-wide CNV detection, and remains the most prevalent means for CNV detection in the clinic (Keren, 2014).

Although not originally designed for the purpose of CNV detection, the majority of large-scale genetic studies of CNVs are conducted using SNP genotyping arrays. SNP arrays offer higher throughput, lower cost per sample, and lower DNA input requirements compared to aCGH. Their widespread use stems mainly from the combined information that SNP arrays provide—samples originally collected and assayed for GWAS using common SNP variation can subsequently be assessed for CNVs without the generation of new data. We describe a detailed protocol for CNV detection from SNP arrays in Chapter 2.

The advent of NGS has made a tremendous impact in both research and medical genetics, since whole genome sequencing (WGS) has, in theory, the ability to catalog all types of variation. However, in contrast to single nucleotide variants (SNVs) and small insertion/deletions (indels), there remains, at present, no established consensus on how best to detect structural variants. Current methods for CNV detection from standard paired-end, Illumina WGS broadly make use of two types of information to detect CNVs: breakpoint detection and read depth. Breakpoints are defined by either mapping “split” reads that directly overlap CNV breakpoints, or by identifying regions were the intervening space between paired alignments differ substantially from expectation. Some popular programs that use split or paired-end read mapping exclusively are Pindel, BreakDancer, and Lumpy (Trost et al., 2018). Other detection algorithms, such as CNVnator, cn.MOPS, and Canvas, use only read-depth information to detect CNVs throughout the genome (Trost et al., 2018). In Chapter 5, we utilize GenomeSTRiP (Handsaker et al., 2015), an algorithm developed by the 1000 Genomes Project that incorporates both breakpoint detection and read-depth information across all samples for both detection and genotyping of CNVs (Figure 1-1C). Methods for CNV detection from whole-

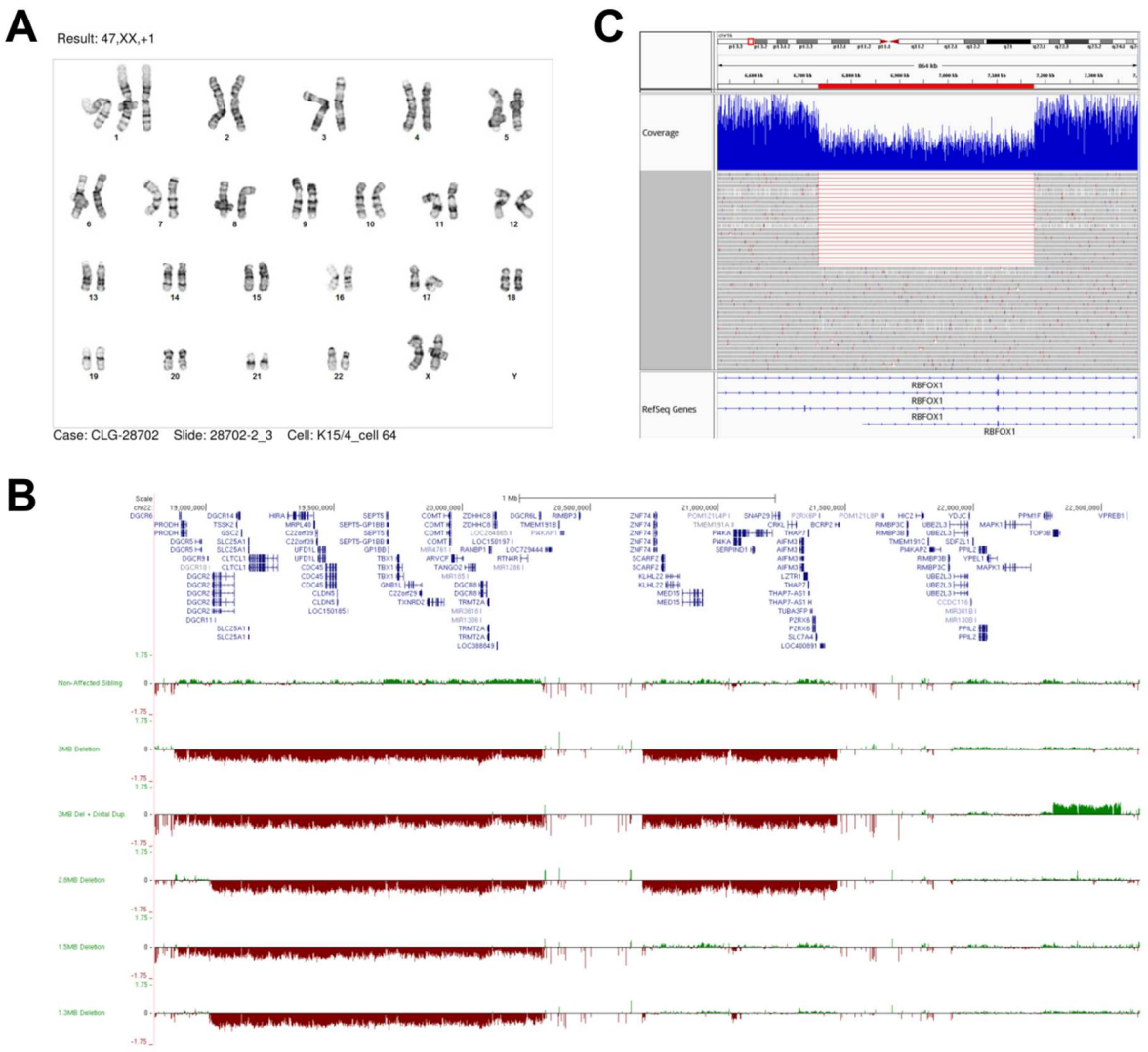


Figure 1-1. Methods for CNV detection.

(A) Analysis of a human induced pluripotent stem cell line an abnormal karyotype (47,XX +1), indicating an extra copy of chromosome 1. This abnormality was mosaic in the cell population. (B) Custom aCGH profiling of an unaffected individual and five 22q11.2 deletion patients distinguishes unique microscopic CNV characteristics in patients that are not visible using traditional cytogenic analysis techniques. (C) Detection of a large heterozygous deletion (red bar) in RBFOX1 using whole-genome sequencing indicated by a drop in read depth coverage and discordant read-pair spacing.

exome sequencing (WES) have also been developed, all of which exclusively make use of read depth information. However, since the capture process used in WES severely alters the uniformity of sequencing coverage, these methods all rely on some initial form of dimensionality reduction to remove systematic noise, limiting sensitivity.

1.3 Copy number variation and human disease

The notion that rare CNVs contribute to disease phenotypes far preceded our ability to detect and appreciate their general prevalence in the population. The first CNV causing a human disease, chromosome 21 aneuploidy as the molecular basis for Down's syndrome, was described in 1959 (Lejeune et al., 1959), only 3 years after the number of chromosomes in humans had been correctly identified (Tjio and Levan, 2010). Large chromosomal abnormalities among individuals with developmental disorders would first be observed some three decades later. Following these early observations, numerous large-scale studies utilizing first microscopic observation, then high-throughput technologies, have been conducted in order to understand both the heterogeneity and overall extent to which submicroscopic CNVs contribute to underlying genetic etiology of pervasive disorders of developmental delay (DD). In one of the largest of these, Cooper at al. characterized CNVs in a sample of 15,767 children with intellectual disability (ID) and congenital malformations compared to 8,329 controls, and estimated that rare, large CNVs accounted for approximately 14% of disease cases (Cooper et al., 2011). They observed two trends also seen in many other large, phenotypically diverse CNV surveys: that increasing phenotypic severity generally correlated with increased CNV size (Girirajan et al., 2011), and that many, if not all of the large, recurrent CNVs observed were associated with a diverse clinical presentation among carriers.

In 1935, Haldane first proposed that for a disease with a stable prevalence, the extent to which contributing mutations occur in a population must balance the removal of such by selection (Haldane, 1935). This principle extends naturally to the study of developmental and neuropsychiatric disorders, which are, for the most part, accompanied by a marked reduction in overall fecundity (Power et al., 2013). Sebat et al. were the first to examine this principal formally in autism spectrum disorder (ASD), by directly comparing the rate of de novo CNVs identified between 118 ASD and 196 control trios (Sebat et al., 2007). They observed that the

rate of de novo CNVs was significantly higher in sporadic cases (10%) compared to both cases with an affected first-degree relative (3%) and controls (1%). Success in this regard undoubtedly influenced the formation of the Simons Simplex Collection (SSC), a large cohort of over 2,600 simplex families of idiopathic ASD with extensive phenotypic characterization and a wealth of genetic data (Fischbach and Lord, 2010). Studies based on the SSC have firmly establish a role for de novo CNVs in ASD (Levy et al., 2011; Sanders et al., 2011), and by combining this data with *de novo* SNVs determined from WES, a total of 71 risk loci with strong statistical support have now been identified (Sanders et al., 2015).

In contrast to ASD, our current understanding of the role for CNVs in the genetic etiology of schizophrenia (SCZ) have been informed, for the most part, by large-scale case-control surveys, although *de novo* CNVs have also been shown to confer risk. In one of the earliest of such surveys, Stefansson et al. examined the transmission of CNVs genome wide among 2,160 trios and 5,558 parent-offspring pairs in *normal* individuals to discover recurrent de novo events (Stefansson et al., 2008). They then screened for these CNVs in a large sample of patients with SCZ and related psychoses along with controls, identifying three novel significant associations on 1q21.1, 15q11.2, and 15q13.3. In a separate study from the International Schizophrenia Consortium, published in same issue of the journal *Nature*, Stone et al. performed a genome-wide screen for rare CNVs in a total sample of 6,572 ethnically matched cases and controls (International Schizophrenia, 2008). In additional to also reporting significant associations on 1q21.1 and 15q13.3, they conclusively demonstrated an increased burden of rare CNVs in SCZ, particularly among singletons, large deletions, and those affecting genes. The largest CNV study in SCZ performed to date, which includes more than 40,000 samples, has confirmed these early findings, and defined at least eight distinct genome-wide significant CNVs that confer substantial risk for the disease (Marshall et al., 2016).

1.4 Rationale

The work in this dissertation is primarily aimed at characterizing the contribution of rare CNVs for two neuropsychiatric disorders for which a role for such variants has not yet been well established, Tourette syndrome (TS) and bipolar disorder (BP). Although ultimately, the goal is to identify *bona fide* risk factors and specific genes for both TS and BP, CNV studies conducted over the last decade in ID/DD, ASD, and SCZ provide us with a sobering fact on such an endeavor: none of the definitive CNV risk loci identified in any of these disorders accounts for more than one percent of affected individuals (Kirov, 2015). Therefore, I will also focus on discerning more distributed differences between cases and controls through careful burden analysis and by restricting analyses to certain types of CNVs with a high likelihood of effect, such as known pathogenic events, *de novo* variants, or CNVs that affect specific subsets of genes. Such analyses are critical for identifying disease-relevant pathways and can inform whether larger future gene discovery efforts will be warranted.

Chapter 2:

A Workflow for Robust Detection of CNVs Using SNP Arrays

2.1 Abstract

Multiple software packages implementing a number of different algorithms have been developed in order to solve the specific problem of identifying CNVs from SNP array intensity data. However, in contrast to SNPs, where it is routine to expect >99% accuracy and specificity for genotyping calls “out of the box”, the process for producing reliable CNV calls suitable for downstream analysis is comparably much more labor intensive, requiring numerous stages of quality control, manual inspection, and post-call cleaning. In this chapter, I provide a brief background on the data generated from SNP arrays used for CNV calling, followed by a description of our generalized CNV calling pipeline, highlighting key aspects I have determined to be critical for producing robust, reliable results. Finally, I describe a computational framework that I developed to help facilitate analyses.

2.2 Background

For each SNP assayed, Illumina arrays generate two intensity measures, corresponding to each allele. CNV detection is derived from two parameters calculated from these signal intensities: *R*, the sum of intensity values for each allele, and θ, a polar-transformed ratio of the allele intensities which measures the relative contribution of each allele to the overall signal intensity. From these two parameters, *R* and θ, two values are calculated which serve as the basis for CNV detection from SNP arrays:

$$LRR = \log_2(\frac{R}{R_e})$$

$$BAF = \begin{cases} 0 & \text{if } \theta < \theta_{AA} \\ 0.5(\frac{\theta-\theta_{AA}}{\theta_{AB}-\theta_{AA}}) & \text{if } \theta_{AA} \leq \theta < \theta_{AB} \\ 0.5 + 0.5(\frac{\theta-\theta_{AB}}{\theta_{BB}-\theta_{AB}}) & \text{if } \theta_{AB} \leq \theta < \theta_{BB} \\ 1 & \text{if } \theta > \theta_{BB} \end{cases}$$

The calculation of these values for a single SNP is summarized in Figure 2-1A. The log-R ratio (LRR) is a normalized signal intensity at a particular SNP and is indicative of the total amount of DNA present. SNPs that have a normal, diploid copy number of 2 will have an *R* value close to the expected value (R_e) and hence, an LRR close to 0. The B-allele frequency (BAF) value provides an additional layer of information from which to infer copy number state at a given SNP. For normal, diploid copy number values of 2, homozygous SNPs will have BAF values close to either 0 or 1 (representing an intensity measure composed entirely of a signal from either AA alleles or BB alleles), and heterozygous SNPs will have a BAF value close to 0.5 (an equal contribution of the A and B alleles). For a heterozygous duplication (Figure 2-1B), we observe an overall increase in LRR values coupled with a deviation in the BAF "banding" pattern of heterozygous SNPs away from the expected value of 0.5. Heterozygous deletions (Figure 2-

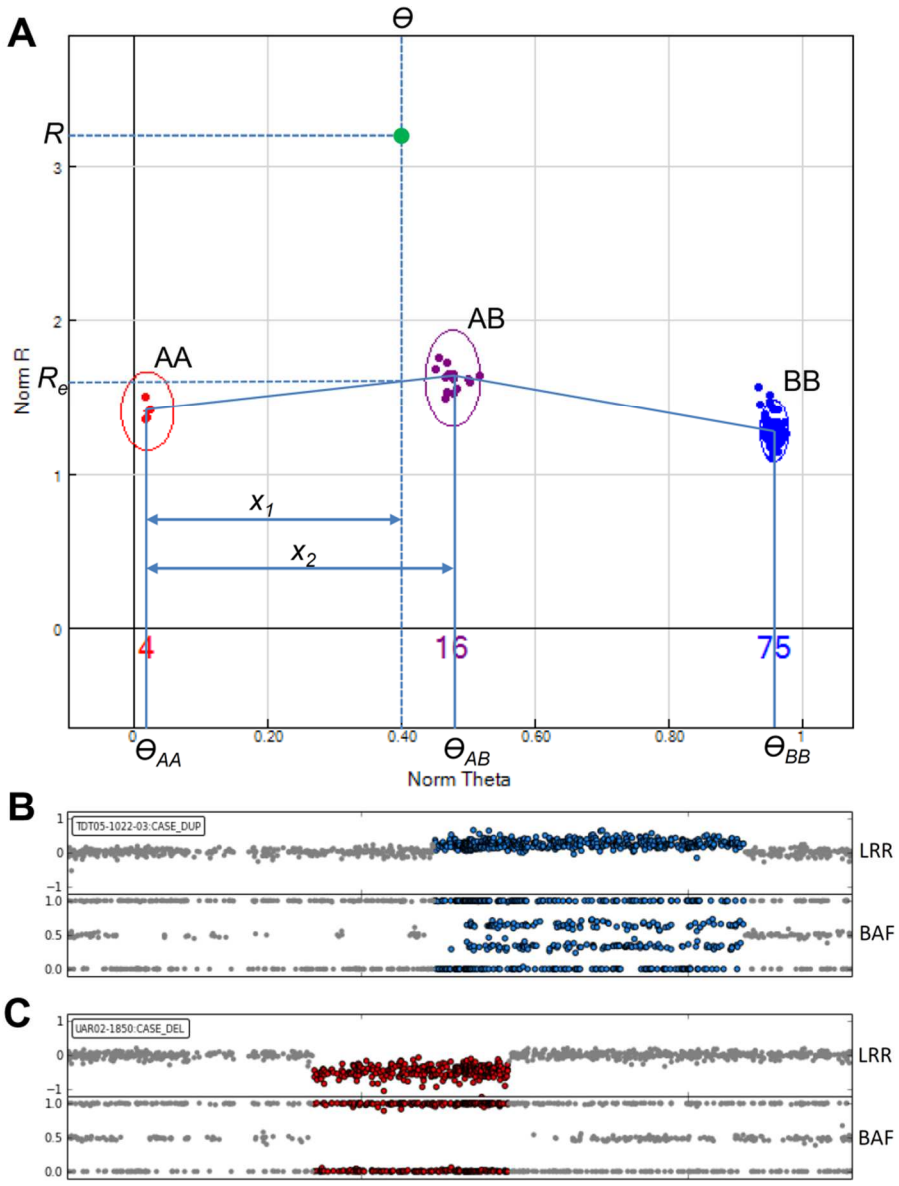


Figure 2-1. Parameters used for CNV detection from SNP arrays.

(A) For the highlighted SNP (green dot), the sum of allelic intensities (LRR) is calculated by subtracting the observed value (R) from the expected value (R_e), which are determined based on the position of cluster centers (colored circles). Allelic intensity (BAF) is calculated in this instance by 0.5 (x_1 / x_2). Prototypical probe-level plots are depicted for a heterozygous duplication (B) and deletion (C).

1C) exhibit a drop in LRR alongside an absence of heterozygosity, indicated by a BAF banding pattern along the values of 0 and 1, exclusively.

2.3 Overview of the CNV calling pipeline

An outline of the pipeline we use throughout this thesis for calling CNVs from SNP array data is provided in Figure 2-2. The pipeline can roughly be divided into four main parts: data preprocessing, CNV detection, quality control (QC), and post-call filtering.

Preprocessing of intensity data

An often ignored, but absolutely critical factor for the accurate detection of CNVs is the appropriate determination of expected values R and θ for each respective genotype cluster. These values are typically provided by Illumina in the form of a canonical cluster file (*.egt) based on empirical data generated from diverse set of samples drawn from the International HapMap Consortium. Although the use of a canonical cluster file typically demonstrates adequate performance for SNP calling, generating a custom cluster file specifically within each individual genotyping batch is essential for producing CNV data that is comparable between batches. The effect of this is illustrated in Figure 2-3. For SNP kgp12087965, the canonical cluster file produces accurate and complete genotype calls in all samples (indicated by the clear separation of the differently colored clusters). However, the position of the canonical cluster centers (colored ellipses) are clearly mis-specified (Figure 2-3A), resulting in a substantial underestimation of LRR across all genotypes. Simple reclustering of the raw intensity data appropriately defines cluster centers based on the empirical data (Figure 2-3B). The overall effect of SNP reclustering on downstream CNV calling can be quite substantial. The number of resultant CNV calls per sample is shown for three separate genotyping batches of samples from the same study all genotyped on the Omni2.5 array in a single lab, produced from intensity measures derived from canonical clusters (Figure 2-3C) and after custom cluster generation within each genotyping batch (Figure 2-3D). Additionally, we have observed that cluster

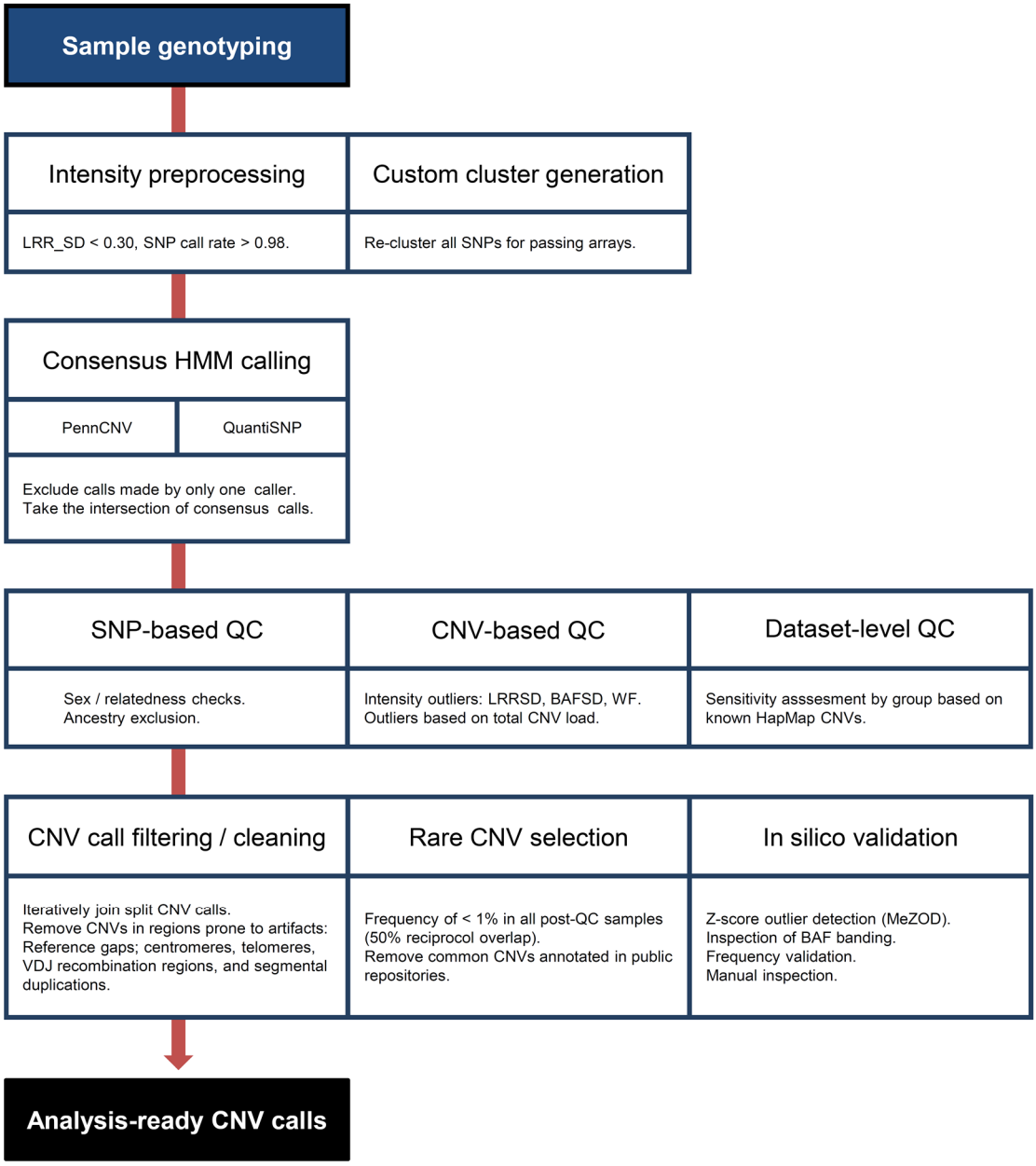


Figure 2-2. Schematic overview of the CNV processing pipeline.

definition can be affected by the presence of poorly performing samples. Therefore, we first process all samples with canonical cluster file using Beeline (Illumina), which provides rapid calculation of the LRR standard deviation (LRR_SD) and call rate. We exclude samples with an LRR_SD > 0.30 and call rate < 0.98, and perform re-clustering on the remaining, high-quality samples only.

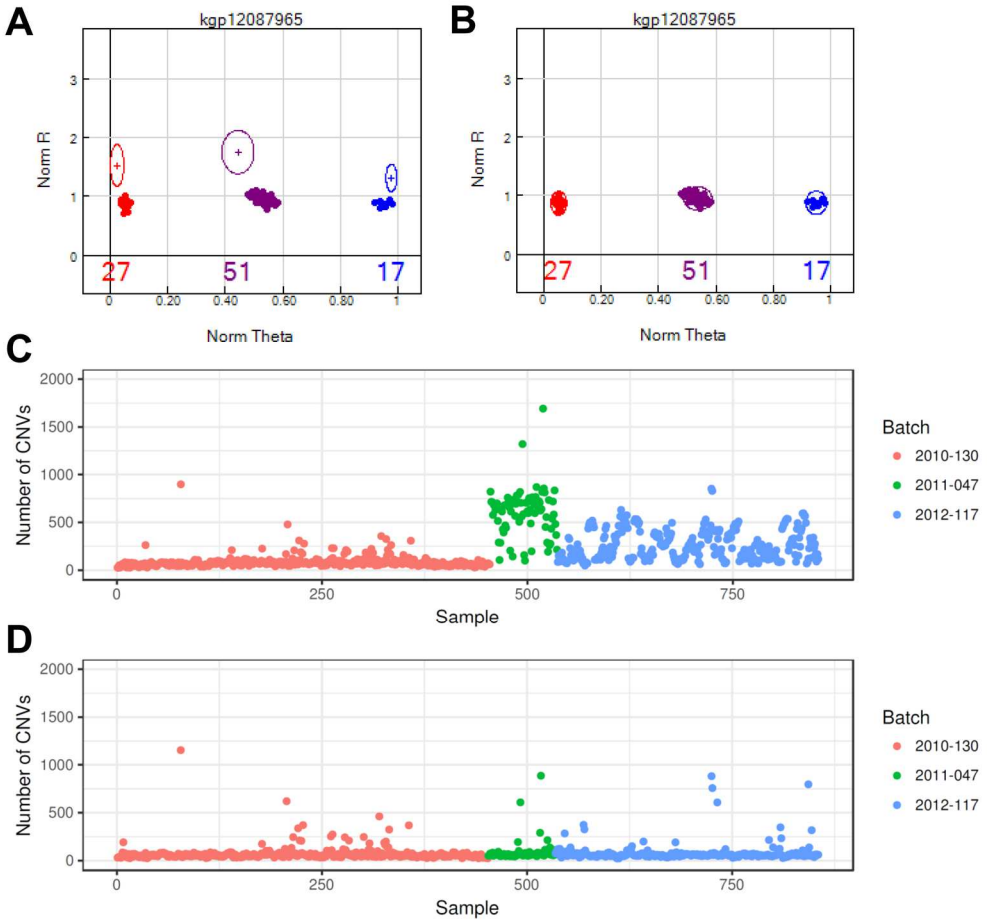


Figure 2-3. Effect of custom cluster definition on downstream CNV calling.
(A) Genotype calls for kgp12087965 based on canonical cluster files derived from 270 HapMap samples produces reliable genotype calls (differently colored dots); however cluster centers are clearly mis-specified (colored elipses). (B) Reclustering this SNP accurately sets cluster centers based on empirical data. The number of CNVs called per individual across the same three genotype batches is shown based on intensity measures derived from using the canonical cluster file (C) and with generating a custom cluster file within in batch separately (D). Custom cluster generation within each batch greatly reduces systematic differences and improves comparability across CNV calls.

Consensus CNV detection

Numerous studies have evaluated the performance of different CNV calling algorithms, and have noted the low reproducibility, high false-positive rates, and general discordance between methods (Pinto et al., 2011). To ameliorate this and its potential influence on

subsequent association testing, we adopt a consensus approach using two Hidden markov model (HMM)-based algorithms for CNV detection, PennCNV (Wang et al., 2007) and QuantiSNP (Colella et al., 2007). With the aim of increasing specificity, we take a conservative consensus approach by taking the innermost boundaries of CNVs of the same relative type (gain or loss) called by both algorithms, therefore excluding discordant CNVs and those called by only one method.

Quality control for CNV detection

Sample-level QC: Although our re-clustering procedure alongside standard QC measures based on SNP genotype data tends to remove most gross assay failures, oftentimes assays that produce data of sufficient quality for genotypes will produce poor CNV calls. QC for CNV calling is composed of removing outliers with regard to overall intensity metrics and total CNV load. As CNV intensity metrics are strongly correlated, we focus mainly on three intensity measures:

1. LRR standard deviation (LRR_SD): The standard deviation of LRR values and the primary metric for assay quality and overall signal to noise ratio.
2. BAF standard deviation (BAF_SD): The standard deviation of heterozygous BAF values within the range of [0.25-0.75].
3. Waviness Factor (WF): The waviness factor is calculated as defined in Diskin, et al., and essentially measures the dispersion or “waviness” of the LRR signal after accounting for local GC content (Diskin et al., 2008).

Though there has been some general consensus literature as to appropriate inclusion thresholds for these metrics (Pinto et al., 2011), in most cases, given a sufficiently large number of samples, it is preferable to determine these values empirically by observing the distribution of these metrics against measures of overall CNV load, including the total number of CNVs and

the aggregate CNV length, as intensity metrics vary considerably between datasets (Figure 2-4A and B). We then further remove samples that exhibit excessive CNV load in terms of number of raw CNV calls and/or total length of CNVs. A chromosome-level plot of overall CNV load is useful to distinguish possible aneuploidies from poorly-performing assays (Figure 2-4C and D).

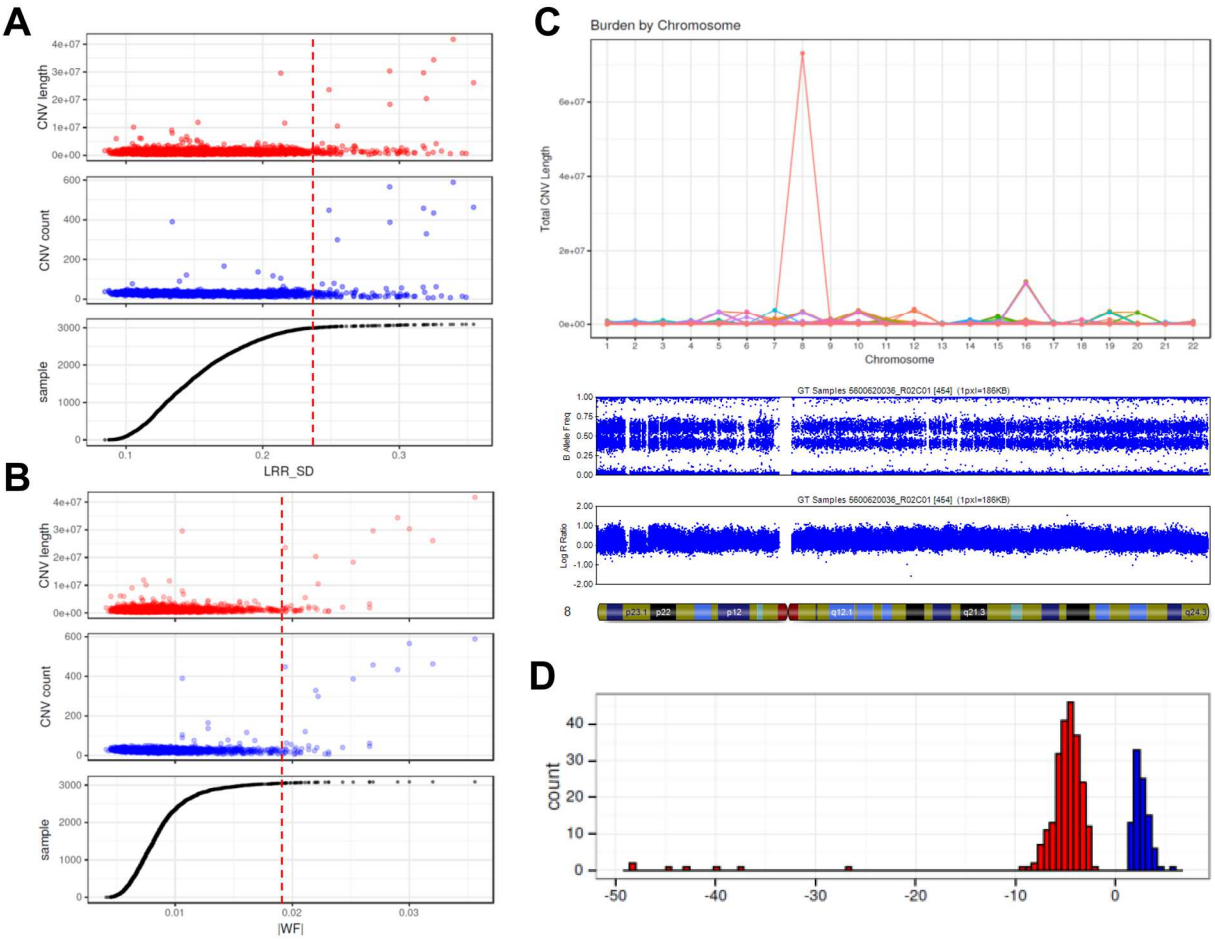


Figure 2-4. Quality control measures for CNV calling.

Individual CNV load, in terms of both the total length of CNVs and the number of CNV calls, is plotted against the intensity metrics LRR_SD (A) and WF (B). Empirical inspection of these distributions is invaluable for determining reasonable thresholds for inclusion (red dotted lines) and for balancing the tradeoff between assay quality and sample attrition. Chromosome level plots of CNV load per individual (C) are useful for revealing instances of aneuploidy; in this case there is a clear evidence for a chromosome 8 trisomy. (D) Median-summarized Z-scores for LRR from Illumina arrays across rare, validated CNVs from 270 HapMap samples. Rare CNVs show very discreet Z-score values for both deletions (red) and duplications (blue).

Dataset-level QC: The HMM-based methods we use for CNV detection work within each sample individually, without incorporating intensity information across all samples. Specifically, this means that both rare CNVs and known, common CNVs are determined within an individual sample in the exact same manner. Therefore, the overall sensitivity for detection at known, common CNV loci serves as an appropriate proxy for the overall sensitivity of CNV detection genome-wide, including rare CNVs. To calculate detection sensitivity, we first probabilistically assign genotype calls for all samples across a set of known, common CNVs with a locus-specific genotyping method that clusters median Z-score summarized intensity values for a particular CNV across all samples using a Gaussian mixture model (GMM). We then calculate the proportion of concordance with CNV calls determined from genome-wide detection using HMMs. This sensitivity metric is particularly useful in determining if there is any differential bias in detection between different datasets or phenotypic groups, and can even be extended to compare detection sensitivities between different assay types (e.g. SNP array vs aCGH). This methodology was used for the TS CNV study and is described in detail in Chapter 3.11.

Post-call CNV cleaning, call filtering, and validation

CNV cleaning and call filtering: Large CNVs are frequently segmented into a number of smaller, disjoint calls by CNV detection algorithms. For example, the whole-chromosome duplication in Figure 2-4C was split into 862 distinct segments. Excessive segmentation of a single, large CNV into smaller segments may bias downstream burden analyses based on CNV counts and/or when stratified by CNV size. Therefore, we merge CNV calls at the sample level when the number of SNPs in the intervening space between adjacent CNVs of the same type is no more than 20% of the total of both segments combined, repeating this merging process iteratively until no more joining occurs. After CNV merging, we remove calls overlapping by more than 50% of their length with regions known to produce artifacts, including genomic gaps, centromeric, telomeric, and VDJ recombination regions (Pinto et al., 2011; Wang et al., 2007).

In silico validation: CNV studies often cite verification of putative calls by manual inspection as a final step prior to analysis (Leppa et al., 2016; Pinto et al., 2010). However, as CNV studies continue grow in size and arrays continue to improve in detection resolution, the number of CNVs that require manual inspection becomes prohibitively large. Therefore, we calculate a series of *in silico* metrics to qualify rare CNV calls. The goal of these scoring metrics is not to impose any hard filtering, but rather to flag possibly problematic events that require further inspection.

For each putative rare CNV, we generate two different metrics, based on LRR intensity and BAF banding. For qualifying CNVs based on intensity, we standardize LRR values within each individual sample to Z-scores, and adopt a simple scoring scheme similar to the median Z-score outlier detection (MeZOD) method (McCarthy et al., 2009; Vacic et al., 2011). Each rare CNV is scored by taking the median of LRR intensity Z-scores (LRR-Z) for all probes within the region. For Illumina arrays, based on empirical data from a large number of qualified rare CNVs from publically available HapMap samples (Illumina), we observed that for rare CNVs, standardized intensity measures typically range from < -20 for homozygous deletions, [-6,-2.3] for heterozygous deletions, and > 1.3 for duplications. To qualify CNVs based on BAF banding, we calculate the proportion of probes within the CNV region that show evidence of a duplication event (BAF of [0.25-04] or [0.6-0.8]), denoting this measure "BAF-D." We typically flag deletions that had a LRR-Z > -2 or BAF-D > 0.02, and duplications with a LRR-Z < 1 or BAF-D < 0.1. *In silico* validation is particularly useful for identifying problematic CNV calls likely due to individual mosaicism. An example of this is shown in Chapter 3 (Figures 3-S3A and B). Furthermore, misspecification of rare events will add to the burden of multiple tests. To ensure the rarity of each CNV call, we also evaluate the proportion of samples whose LRR-Z metric falls outside of [-2.3, 1.3]. Putative rare CNV loci that show substantial evidence for extensive polymorphism

are subsequently scored for frequency using the GMM genotyping method described above and excluded if necessary.

2.4 A computational framework for the analysis of large-scale CNV studies

In practice, the proper analysis of CNV data is not trivial. In contrast to the genotype data, one cannot simply perform imputation to "homogenize" different datasets for downstream analysis; meticulous quality control and manual inspection are essential but also extremely labor intensive. For these tasks, access to the raw intensity data is required and is typically provided as a single file per sample. This quickly becomes untenable for large samples sizes and is further complicated by the fact that even versions of the same array may have slightly different annotations and SNP content.

To address this issue, I developed a computation framework that provides a means to manage and track a large number of samples, and perform several of the specific tasks outlined in this chapter, including genotyping known, common CNVs, performing *in silico* validation of rare CNVs, and rapidly visualizing CNV calls/loci of interest across one or many samples simultaneously. An overview of this framework is provided in Figure 2-5. At the center is the creation of an HDF5 data file that consolidates raw intensity data, various SNP annotations, sample metadata and QC metrics into a single, on-disk database that can be rapidly accessed through a collection scripts via an interactive Python session. The data is chunked on disk for rapid access using hd5py and pandas and easily scales to hundreds of thousands of samples across millions of SNPs on modest hardware. The Python interface makes extensive use of Numpy masking arrays to handle missing SNP calls, exclude specific samples, and, if desired, for evaluating a set of CNV calls on samples assayed on different platforms. The framework is designed to accept CNV calls supplied in standard PLINK rare CNV format, or specific intervals in a simple tab-delimited table (for CNV genotyping).

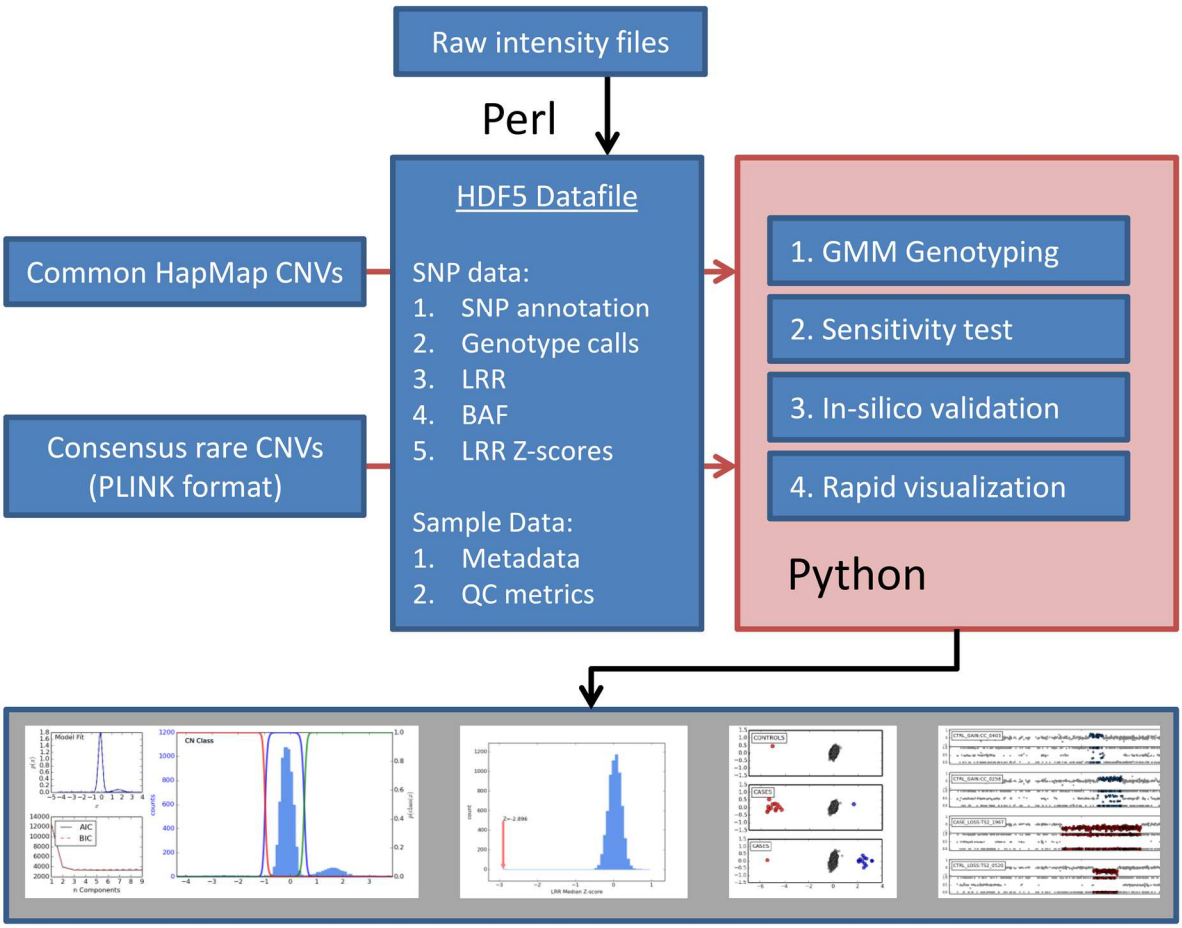


Figure 2-5. A framework for data handling, QC, and rapid visualization for large CNV studies.

Chapter 3:

Rare Copy Number Variation Increases Risk for Tourette Syndrome

3.1 Abstract

TS is a model neuropsychiatric disorder thought to arise from abnormal development and/or maintenance of cortico-striato-thalamo-cortical circuits. TS is highly heritable, but its underlying genetic causes are still elusive, and no genome-wide significant loci have been discovered to date. We analyzed a European sample of 2,434 TS cases and 4,093 ancestry-matched controls for rare (<1% frequency) CNVs using SNP microarray data. We observed an enrichment of global CNV burden that was prominent for large (>1 Mb), singleton events (OR=2.28, 95%CI [1.39-3.79], $P=1.2x10^{-3}$) and known, pathogenic CNVs (OR=3.03 [1.85-5.07], $P=1.5x10^{-5}$). We also identified two individual, genome-wide significant loci, each conferring a substantial increase in TS risk (NRXN1 deletions, OR=20.3, 95%CI [2.6-156.2]; CNTN6 duplications, OR=10.1, 95% CI [2.3-45.4]). Approximately 1% of TS cases carry one of these CNVs, indicating that rare structural variation contributes significantly to the genetic architecture of TS.

3.2 Background

TS is a complex developmental neuropsychiatric disorder, characterized by multiple involuntary motor and vocal tics, with an estimated population prevalence of 0.3-0.9 percent (Scharf et al., 2015). Tics typically emerge during childhood and peak early within the second decade of life; in most cases, there is a marked reduction in symptoms during late adolescence, supporting the notion that the aetiology of TS is neurodevelopmental in origin (Paschou, 2013). The majority of TS patients (>85%) present with one or more additional neuropsychiatric co-morbidities, the most prevalent of which include two other childhood-onset disorders: attention deficit hyperactivity disorder (ADHD), and obsessive-compulsive disorder (OCD) (Hirschtritt et al., 2015). An elevated risk for mood, anxiety, major depressive, and ASD is also common (Burd et al., 2009; Hirschtritt et al., 2015). Consequently, TS is frequently considered a model neuropsychiatric disorder such that identification of the underlying molecular, cellular, and neurophysiologic etiology of TS may be broadly applicable to a wide range of childhood and adolescent psychiatric disorders.

Neuroimaging (Marsh et al., 2009; Worbe et al., 2015), neurophysiology (Draper et al., 2014; Gilbert et al., 2004) and neuropathological (Kataoka et al., 2010) studies in TS patients suggest that the disorder arises from dysregulated development and/or maintenance of cortico-striatal-thalamo-cortical (CSTC) circuits involved in initiation, selection, learning and reinforcement of intended movements and motor sequences (Graybiel, 2008; Jahanshahi et al., 2015; Mink, 2006). Decades of work in rodents and primates have demonstrated the critical role of CSTC circuitry in procedural learning of complex movements and context-dependent habit formation (Jog et al., 1999; Knowlton et al., 1996); in this framework, tics can be viewed as an over-exuberant, pathological version of this fundamental behavior (Graybiel, 2008). Partially segregated motor CSTC loops retain a somatotopic specificity which parallels that found in primary motor cortex (M1) and supplementary motor area (SMA), with M1 loops located more

laterally in the putamen and conveying information on simple movements, while SMA loops project more medially and convey information on complex movements (Flaherty and Graybiel, 1991; Nambu, 2011; Romanelli et al., 2005). This segregation by body location and complexity is reflected clinically in the subgroups of tics seen in TS patients (Hirschtritt et al., 2016). CSTC circuits have also been demonstrated to subserve parallel cognitive, behavioral and affective regulatory functions via networks connecting dorsolateral, medial frontal and orbitofrontal cortex to corresponding regions of the striatum, globus pallidus and thalamus (Alexander et al., 1986; Haber and Knutson, 2010). Abnormal development of these cognitive and limbic CSTC circuits is thought to underlie the frequent co-occurrence of ADHD and OCD in TS patients (Jahanshahi et al., 2015), as supported by corresponding structural and functional neuroimaging studies (Delorme et al., 2016; Wang et al., 2011; Worbe et al., 2015).

Though a variety of environmental factors are thought to contribute to tic severity, and perhaps also to development or emergence of tics (Lin et al., 2007; Mell et al., 2005; Motlagh et al., 2010), it is widely acknowledged that TS is primarily a genetic disorder. Depending on the exact diagnostic criteria used, twin studies have shown a monozygotic concordance rate of 58-77% compared to only 8% for dizygotic twins (Pauls et al., 2014). Recent family studies of TS and chronic tic disorder (CT), a closely related condition with the same characteristics and natural history as TS (comprising motor or vocal tics, but not both), have demonstrated that siblings of TS/CT probands exhibit a 15-fold increased risk of TS/CT compared to that of the general population, while children of TS/CT-affected parents have a 60-fold higher disease risk (Browne et al., 2015). Similarly, a recent population-based family study estimates the heritability (h^2) of TS to be 0.77 (Mataix-Cols et al., 2015), placing it among the most highly heritable complex neuropsychiatric disorders.

Despite this strong genetic component, the identification of *bona-fide* TS susceptibility genes has proven challenging. Although linkage analysis conducted over the last several

decades has identified a number of candidate disease regions, there has been little consensus across these studies, suggesting that, as with other neuropsychiatric disorders, the genetics of TS is both heterogeneous and complex (Pauls et al., 2014). An initial GWAS did not yield any loci that met criteria for genome-wide significance, most likely due to small sample size (Scharf et al., 2013). However, analyses of TS genetic architecture using aggregated GWAS data have demonstrated that TS is highly polygenic, with the majority of inherited TS genetic risk distributed widely throughout the genome including both common (5-50% population frequency) and rare (<5% population frequency) variants (Davis et al., 2013).

Studies examining rare structural variation in individuals with TS have implicated several neurodevelopmental genes involved in neurite outgrowth and axonal migration. Rare chromosomal abnormalities affecting *CNTNAP2* (Verkerk et al., 2003) and *SLITRK1* (Abelson et al., 2005) have been identified in isolated TS families, and multiple exonic copy-number variants (CNVs) in *NRXN1* have been observed in small genome-wide CNV studies (Nag et al., 2013; Sundaram et al., 2010), though none of these loci have surpassed the threshold for genome-wide significance. Because of the evidence suggesting that rare CNVs may have a role in TS etiology (McGrath et al., 2014), and since CNV studies have repeatedly demonstrated that such variants contribute significantly to susceptibility for other heritable neurodevelopmental disorders, such as ID, ASD, SCZ, and epilepsy (Cooper et al., 2011; Helbig et al., 2009; McCarthy et al., 2009; Pinto et al., 2010), we assessed the impact of rare CNVs on TS disease risk in a large sample of 6,527 unrelated individuals of European ancestry. Our results demonstrate a global increase in the burden of large, rare CNVs in TS cases compared to controls that is driven primarily by large, singleton events, in particular large (>1Mb) deletions, consistent with marked genetic heterogeneity. In addition, we report the first two TS susceptibility loci that meet genome-wide significance--heterozygous exonic deletions in *NRXN1*

and exonic duplications in *CNTN6*. Each confers a substantial increase in disease risk and collectively, is present in approximately 1% of TS cases.

3.3 Sample collection and genome-wide detection of rare CNVs

All TS cases and controls were recruited either through the Tourette Syndrome Association International Consortium for Genetics (TSAICG) or through European TS clinics within the Gilles de la Tourette Syndrome GWAS Replication Initiative (GGRI), with additional controls selected from several external studies (See Chapter 3.11 for sample details). All DNA samples were genotyped on Illumina OmniExpress SNP array platforms (Table A-1.A). Prior to CNV calling, we unified our dataset by restricting to SNP assays common to all versions of the OmniExpress array to avoid any potential systematic bias due to minor variances in probe density. We conducted extensive quality control analyses including both SNP-based and CNV-based exclusion of outliers (Table A-1.B), as well as genotype-based determination of ancestry (Figure A-1). The final dataset consisted of 6,527 unrelated European ancestry samples, including 2,434 individuals diagnosed with TS based on the Diagnostic and Statistical Manual of Mental Disorders-Fourth edition, Text Revision (DSM-IV-TR) criteria and 4,093 unselected controls. CNV calls were generated with two widely-used HMM-based segmentation algorithms, PennCNV (Wang et al., 2007) and QuantiSNP (Colella et al., 2007). To improve specificity of our CNV detection, we only retained the consensus of CNVs called by both algorithms.

To validate our intensity pre-processing and CNV calling pipeline prior to performing further analyses, we first assessed for any differential sensitivity in CNV detection rates between cases and controls by comparing the sensitivity of CNV call rates at a subset of known, common CNV loci across phenotypic groups. Since HMM-based CNV detection is performed within each sample individually, it is agnostic to population CNV frequency; therefore, sensitivity measures assessed across common CNVs serve as a reasonable proxy for rare CNV detection rates

genome-wide. We genotyped our samples across 11 common HapMap3 CNV loci that each produced distinct CNV clusters using a sensitive, locus-specific intensity GMM-based method, generating a total of 4,758 non-reference CNV calls across all case-control samples (Figure A-2, Table A-2). We subsequently used the proportion of concordant HMM-based calls as a sensitivity metric for CNV detection and confirmed the absence of any differential bias in CNV detection sensitivity between phenotypic groups, whether assessed for individual loci (Table A-3.A), across all loci (P=0.53, Fisher's exact test, Supplementary Table A-3.B), or as the average sensitivity of individuals within groups (P=0.15, Welch's *t*-test, Supplementary Table A-3.C). We performed post-call cleaning by joining large CNVs that appeared to be fragmented, and assigned each CNV based on a 50% reciprocal overlap with other variants. We then filtered our call set for rare CNVs (as defined by maximum frequency of 65 across all samples, corresponding approximately to a minor allele frequency, or MAF < 1%) of at least 30kb in length and spanning a minimum of 10 probes. Finally, using a heuristically derived series of *in silico* validation metrics, we removed several instances of aberrant CNV calls due to mosaicism and misclassified rare events (Supplementary Figure A-3). Rare CNVs were annotated for gene content according to RefSeq annotation. Conservatively, we only considered a CNV as "genic" if it overlapped any exon of a known protein-coding transcript. In total, we resolved 9,375 rare CNV calls.

3.4 Global burden analysis of rare CNVs in TS

Each CNV was assigned a specific frequency and annotated with the number of protein-coding genes affected by it, and an increase in CNV burden was assessed genome-wide using three different metrics stratified by CNV type and aggregated across individual samples: the total number of CNVs (CNV count), the aggregate length of all CNVs (total size), and the number of genes affected (gene count). We used a logistic regression framework to assess for an increased CNV burden on TS status, adjusting for sources of potential bias including sex,

CNV assay quality, and subject ancestry; an odds ratio (OR) greater than one indicates an increased risk for TS. For genic CNVs (n=4,604), we observed we observed modest but significant increases in burden across all metrics (Figure 3-1.A): CNV count (OR 1.05 [1.01-1.10], P=0.027), CNV gene count (OR 1.09 [1.01-1.17], P=0.019), and CNV length (OR 1.15 [1.07-1.24], $P=1.9x10^{-4}$). By contrast, no enrichment of CNV burden was observed when a comparable number of non-genic events (n=4,771) were considered (Figure 3-1.A). To further ensure that the enrichment of CNV size was not due to unmeasured confounding between cases and one or more subsets of control samples, we stratified this burden analysis by control study source, comparing TS cases separately to controls collected as part of this study and also to the controls from each external source individually. We observed a consistent enrichment of CNV size across all control sets (Supplementary Figure A-4), noting that the magnitude of the effect size was most pronounced in the stratum using control samples ascertained and genotyped concurrently with the TS cases. To explore the CNV length burden further, we partitioned the data across a range of CNV size and frequency bins and observed the enrichment was mainly attributable to large (>1Mb; OR 1.26 [1.08-1.49], $P=5.3x10^{-3}$) (Figure 3-1.B) and/or singleton CNVs (OR 1.13 [1.04-1.24], $P=2.9x10^{-3}$) (Figure 3-1.C) distributed across different loci throughout the genome.

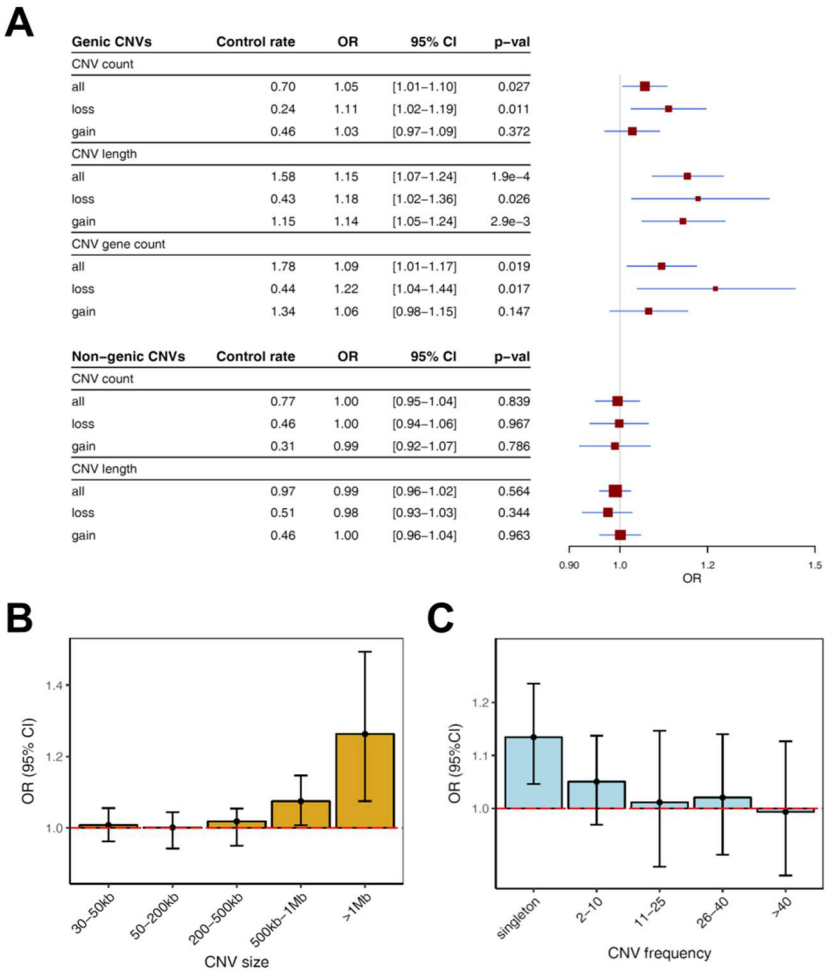


Figure 3-1. Rare CNV burden in 2,434 TS cases and 4,093 controls.

(A) The global burden of all rare (<1% frequency) CNVs > 30kb is shown for genic (top) and non-genic (bottom) CNVs and stratified by CNV type (all, loss (deletions), gain (duplications)). Global CNV burden is compared using three different metrics: CNV count, total number of CNVs per subject; CNV length, aggregate length of all CNVs (in Mb); and CNV gene count, number of genes spanned by CNVs. Control rate, averaged baseline burden metric per control subject. Red boxes, odds ratios (box size is proportional to standard error); Blue lines indicate 95% confidence intervals. Genic CNVs are defined as those that overlap any exon of a known protein-coding gene. (B) Analyses in (A) were assessed further by partitioning CNV length burden of all CNVs (deletions + duplications) into different CNV size categories. Whiskers represent 95% confidence intervals. (C) The analysis in (B) was repeated for CNVs binned by frequency. Odds ratios (OR) were calculated from logistic regression adjusted for covariates using standardized burden metrics. ORs >1 indicate an increased TS risk.

3.5 Enrichment of large, singleton events and CNVs affecting conserved genes

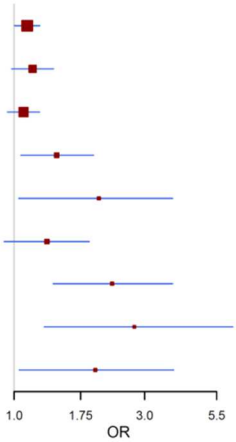
Having observed that the elevated burden in TS was largely confined to large and/or very rare events, we reevaluated burden by CNV count restricted to singleton events and stratified by CNV size. We observed a significant enrichment of singletons >500kb (OR 1.43 [1.06-1.95], P=0.020) that was further increased in the largest size category (>1Mb, OR 2.28 [1.39-3.79], $P=1.2x10^{-3}$). The enrichment of >1Mb genic CNVs was greater for deletions (OR 2.75 [1.28-5.23], P=0.012) than duplications (OR 1.98 [1.04-3.83], P=0.038) (Figure 3-2.A). Recent large-scale studies of human genomic variation have demonstrated marked variability in the sensitivity of individual genes to deleterious, gene-disrupting variation (Lek et al., 2016; MacArthur et al., 2012; Petrovski et al., 2015; Ruderfer et al., 2016). As predicted by evolutionary theory, negative selection limits the ability of deleterious genetic variation to be passed on through multiple generations, such that individual gene-disrupting variants in loss of function-intolerant genes will be quite rare in the general population, and these genes will demonstrate a high degree of evolutionary conservation (Samocha et al., 2014). Two studies have defined metrics of evolutionary constraint, residual variation intolerance score and pLI (probability of loss of function intolerance), for base-pair level variation using exome sequencing data from the Exome Sequencing Project and Exome Aggregation Consortium (ExAC). Ruderfer and colleagues have recently applied a similar approach to define CNV intolerance per gene (Ruderfer et al., 2016) and demonstrated that the ExAC pLI score served as a better predictor of CNV deleteriousness than CNV intolerance score alone, given the tendency for most CNVs to span multiple genes.

Based on these findings, we hypothesized that the large (>1 Mb), singleton CNVs demonstrated above to be enriched in TS cases would be more likely to harbor highly conserved, LoF intolerant genes compared to those in controls. Considering both deletions and

A

Burden of Genic Singleton CNVs

Size	Type	Control rate	OR	95% CI	p−val
<500kb	all	0.251	1.12	[1.00−1.23]	0.041
	loss	0.091	1.17	[0.98−1.39]	0.082
	gain	0.161	1.09	[0.95−1.24]	0.201
>500kb	all	0.029	1.43	[1.06−1.95]	0.020
	loss	0.006	2.04	[1.04−3.79]	0.041
	gain	0.023	1.32	[0.92−1.88]	0.129
>1MB	all	0.008	2.28	[1.39−3.79]	1.2e−03
	loss	0.003	2.75	[1.28−6.29]	0.012
	gain	0.005	1.98	[1.04−3.83]	0.038



B

Burden by Clinical Relevance

ACMG class	Type	Control rate	OR	95% CI	p−val
BENIGN	all	0.101	1.07	[0.91−1.24]	0.417
	loss	0.037	1.02	[0.77−1.33]	0.907
	gain	0.069	1.09	[0.90−1.32]	0.357
UNCERTAIN	all	0.496	1.04	[0.96−1.12]	0.344
	loss	0.150	1.12	[0.97−1.28]	0.122
	gain	0.346	1.00	[0.91−1.10]	0.922
PATHOGENIC	all	0.008	3.03	[1.85−5.07]	1.5e−5
	loss	0.003	3.94	[1.83−8.95]	6.3e−4
	gain	0.005	2.48	[1.31−4.84]	5.9e−3

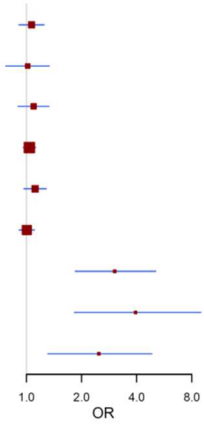


Figure 3-2. Large, singleton CNVs and known pathogenic variants are overrepresented in TS.
(A) CNV count burden restricted to genic singleton events, stratified by CNV size and type (deletion/duplication). (B) CNV burden of all rare CNVs, separated by clinical relevance (benign, uncertain, pathogenic) according to the ACMG guidelines. Red boxes, odds ratios; Blue lines, 95% CIs. ORs > 1 represent an increase in risk for TS per CNV.

duplications together, TS cases had a 2.7-fold increased rate of large (>1 Mb) singleton CNVs spanning genes under strong evolutionary constraint (pLI>0.9) (Relative risk=2.65 [1.40-5.00], $P=3.0x10^{-3}$; Poisson regression controlling for sex, assay quality and ancestry principal components). This increased burden was primarily driven by large, singleton deletions spanning

highly-constrained genes (RR=3.95 [1.33-11.68], P=0.01), though large singleton duplications appeared to contribute additional risk as well (RR=2.12 [0.96, 4.67], P=.06).

3.6 Enrichment of known pathogenic neurodevelopmental CNVs

It is well established that multiple regions of the human genome are particularly prone to large, rare, recurrent genomic deletions and/or duplications, each of which is typically associated with a broad range of neurodevelopmental disorder (NDD) phenotypes (Malhotra and Sebat, 2012). To characterize the extent to which known these pathogenic CNVs might also confer risk for TS, we annotated all CNV calls according to the American College of Medical Genetics (ACMG) guidelines (Kearney et al., 2011), based on curated databases (ClinVar and DECIPHER), automated annotation (Scripps Genome Annotator), and manual review of the literature. Known pathogenic CNVs were identified in 1.9% of TS cases vs. 0.8% of controls (OR 3.03 [1.85-5.07], $P=1.5x10^{-5}$) (Figure 3-2.B). Consistent with an increased pathogenicity of deletions compared to duplications, this enrichment was greater for deletions alone (OR per CNV 3.94 [1.83-8.95], $P=6.3x10^{-4}$). By contrast, no increase in burden was observed among CNVs classified as either benign or of unknown clinical significance.

3.7 Deletions in *NRXN1* and duplications in *CNTN6* confer substantial risk for TS

We next evaluated the dataset for enrichment of rare CNVs at individual genomic loci, conducting an unbiased, point-wise (segmental) test of association genome-wide, treating deletions and duplications independently. As non-overlapping CNVs that affect the same gene would be unaccounted for by segmental assessments of enrichment, we also performed a complementary gene-based test, conditioned on CNVs affecting exons. In contrast to genome-wide SNP association studies, there is no specific p-value threshold to indicate genome-wide significance for CNVs, since the number of rare CNV breakpoints per genome varies across individuals and detection platform. Therefore, for both tests, we established empirical locus-

specific p-values (P_{seg} and P_{gene} for segment and gene-based tests, respectively) and genome-wide corrected (P_{corr}) p-values through 1,000,000 label-swapping permutations, using the max(T) method to control for family-wise error rate (Westfall and Troendle, 2008). Both tests converged on the same two distinct loci, one for deletions and another for duplications, which were enriched among TS cases and survived genome-wide correction for multiple testing.

For deletions, the peak segmental association signal was located on chromosome 2p16 at rs13418185 ($P_{seg}=7.0x10^{-6}$; $P_{seg\text{-}corr}=1.0x10^{-3}$; Figure 3-3.A), corresponding to heterozygous losses across the first two exons of *NRXN1*, and found exclusively among TS cases (N=10, Figure 3-3.B). In the genome-wide gene-based test of exonic CNVs, heterozygous losses across *NRXN1* were also the most significant association for deletions ($P_{gene}=5.9x10^{-5}$; $P_{gene\text{-}corr}=8.5x10^{-4}$), representing 12 cases (0.49%) and a single control (0.02%), and correspond to a substantial increased risk for TS (OR=20.3 [2.6-156.2]). Consistent with pathogenic deletions previously identified within this gene in ASD, SCZ, and epilepsy patients, these exon-spanning CNVs clustered at the 5' end of *NRXN1* and predominantly affected the *NRXN1*-α isoform (Ching et al., 2010).

The segmental association test for CNV duplications identified one genome-wide significant locus at marker rs4085434 on chromosome 3p26 within *CNTN6* ($P_{seg}=5.4x10^{-5}$, $P_{seg\text{-}corr}=6.9x10^{-3}$) with a secondary peak at marker rs981235 located directly upstream of *CNTN6* ($P_{seg}=5.9x10^{-5}$, $P_{seg\text{-}corr}=6.9x10^{-3}$, Figure 3-3.A and C). Closer inspection of the locus revealed an enrichment of large duplications spanning this gene. The gene-based test identified the same locus, duplications overlapping *CNTN6*, with gains found in 12 cases (0.49%) and 2 controls (0.05%), corresponding to an OR=10.1 [2.3-45.4] which was also genome-wide significant ($P_{gene}=2.5x10^{-4}$, $P_{gene\text{-}corr}=8.3x10^{-3}$). Notably, the *CNTN6* duplications identified in our sample are

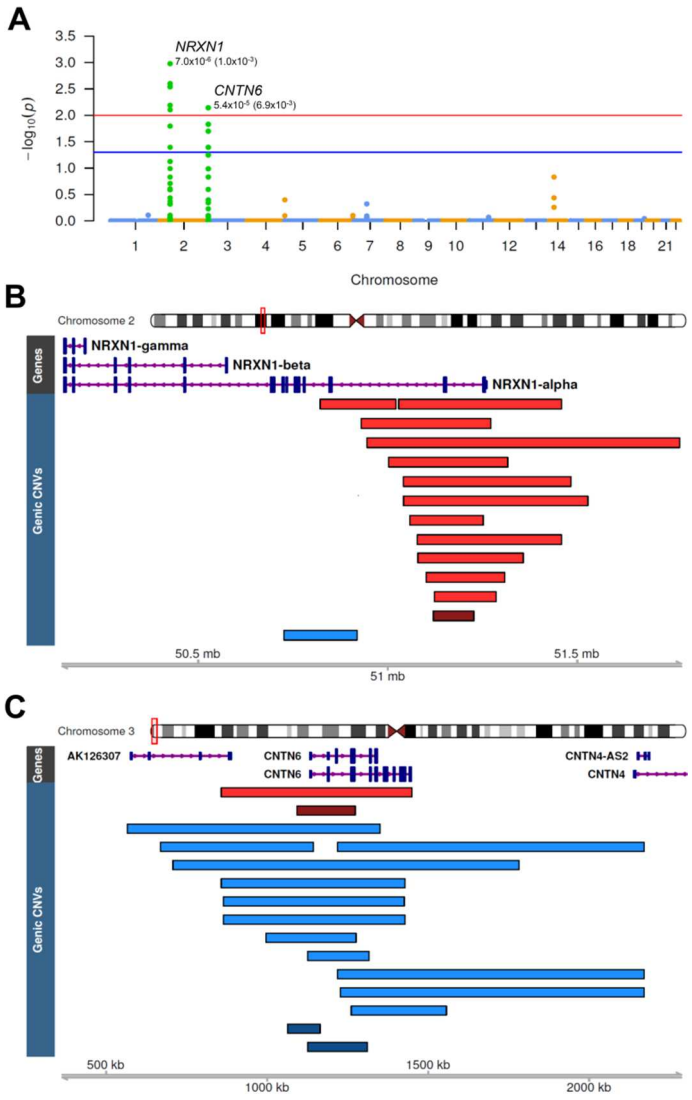


Figure 3-3. Significant results from segmental and gene-based tests converge on two distinct loci. (A) Manhattan plot of segmental association test results representing genome-wide corrected p-values calculated at each CNV breakpoint. The two genome-wide significant association peaks correspond to deletions at *NRXN1* (P_{locus}=7.0x10^{-6}, P_{corr}=1.0x10^{-3}) and duplications at *CNTN6* (P_{locus}=5.4x10^{-5}, P_{corr} =6.9x10^{-3}). Red and blue levels correspond to a genome-wide corrected α of 0.05 and 0.01, respectively. (B) Heterozygous exonic deletions in *NRXN1* found in 12 cases (0.49%) and 1 control (0.02%), corresponding to an OR=20.3, 95% CI (2.6-156.0). Exon-affecting CNVs cluster at the 5' end with deletions across exons 1-3 found in 10 cases and no controls. Red, deletions in TS cases; Dark Red, deletion in controls; Blue, case duplication. (C) Exon-spanning duplications over *CNTN6* found in 12 cases and 2 controls (OR=10.1, [2.3-45.4]) *CNTN6* duplications are considerably larger in cases compared to controls (641 vs. 143 kb, on average). Blue, case duplications; Dark blue, control duplications; Red, case deletion; Dark red, control deletion.

considerably larger in TS cases than those in controls (641 vs. 143 kbp, on average). 9 out of 12 TS carriers harbored a duplication exceeding 500 kbp in length, while both of the *CNTN6* duplications found in controls were less than 200kbp. All duplications detected across *CNTN6* were heterozygous and spanned exons.

We verified all genic CNV calls across *NRXN1* by manual inspection of probe-level intensity plots (Figures A-5 and A-6). Furthermore, we genotyped CNV calls at both *NRXN1* and *CNTN6* using alternative, intensity-only based outlier detection method and by plotting the median-summarized and standardized intensity measures of probes present in the CNV region against probes taken from flanking invariant regions with no CNV calls. We observed a complete concordance with calls generated by HMM-based methods, and the distinct separation between carriers and non-carriers indicated high sensitivity of variant detection across both regions (Figure A-7). No additional loci were significant after controlling for FWER, under either segmental or gene-based tests of association. Additionally, for both deletions and duplications, we obtained similar results after pair-matching each individual case with its closest ancestrally matched control, demonstrating that these results are robust and not the result of inter-European population stratification or case-control sample biases (Figure A-8).

3.8 Phenotypic characteristics of *NRXN1* and *CNTN6* CNV carriers

Analyses of TS age-of-onset, tic severity and rates of co-morbid OCD and/or ADHD revealed no significant differences between TS cases carrying *NRXN1* deletions or *CNTN6* duplications compared to non-carriers (Table 3-1). Four of 12 *NRXN1* deletion carriers were noted to have a broadly-defined NDD (2 ASD, 1 Developmental Delay, 1 Developmental Speech/Language Disorder unspecified); none of the *CNTN6* carriers were noted to have ASD/ID/DD. Although control individuals were not phenotypically characterized, it is worth noting that all control carries of genic deletions in *NRXN1* or duplications in *CNTN6* were

female. Among cases, 8/12 (75%) *NRXN1* carriers and 11/12 (92%) *CNTN6* carriers were males. The singe female TS patient with a *CNTN6* duplication also carried an additional clinically relevant 22q11.2 deletion (Bassett and Scherer). This is suggests that the general theory of a "female protective effect" among carriers of NDD-related CNVs also extends to TS and may, in part, help to explain the gender imbalance of the disorder.

Sample ID	Gene	Chr	Start	End	Type	Length (kb)	Variant Effect	OCD	ADHD	Atypical	Notes
TS1_0630	NRXN1	2	50821559	51021488	DEL	199.9	CODING	N	N	Y	Unspecified Developmental Delay (ICD-9: 315.9)
TS1_0180	NRXN1	2	50930181	51272375	DEL	342.2	CODING	N	N	Y	Asperger Syndrome
TS1_0446	NRXN1	2	50945471	51770480	DEL	825	CODING	Y	Y	N	
TS1_0105	NRXN1	2	51002606	51316822	DEL	314.2	CODING	N	Y	N	
TS2_1256	NRXN1	2	51028662	51458570	DEL	429.9	CODING	N	Y	Y	Other developmental speech or language disorder (ICD-9: 315.39)
TS2_0026	NRXN1	2	51041472	51483528	DEL	442.1	CODING	N	N	N	
TS2_0924	NRXN1	2	51041603	51528298	DEL	486.7	CODING	N	Y	N	
TS2_0750	NRXN1	2	51058745	51252137	DEL	193.4	CODING	Y	Y	Y	Asperger Syndrome
TS2_1238	NRXN1	2	51077569	51458570	DEL	381	CODING	Y	N	Y	Paranoid personality disorder
TS1_0573	NRXN1	2	51079482	51357902	DEL	278.4	CODING	NA	NA	NA	
TS1_0776	NRXN1	2	51101583	51308895	DEL	207.3	CODING	N	Y	N	Brother with Asperger Syndrome
TS1_0698	NRXN1	2	51123048	51286169	DEL	163.1	CODING	Y	Y	N	
TS2_1805	CNTN6	3	565961	1350458	DUP	784.5	CODING	Y	NA	N	
TS2_1405	CNTN6	3	668832	1143424	DUP	474.6	5' UTR	Y	Y	N	
TS2_1624	CNTN6	3	707257	1781739	DUP	1074	CODING	N	N	N	
TS2_1525	CNTN6	3	857325	1427769	DUP	570.4	CODING	Y	Y	N	
TS2_1568	CNTN6	3	864513	1425997	DUP	561.5	CODING	Y	Y	N	
TS2_1545	CNTN6	3	864513	1427769	DUP	563.3	CODING	N	Y	N	
TS2_1320	CNTN6	3	946290	1276092	DUP	329.8	CODING	Y	N	N	
TS1_0618	CNTN6	3	1125605	1315900	DUP	190.3	CODING	N	N	N	
TS2_1156	CNTN6	3	1218279	2170519	DUP	952.2	CODING	N	Y	N	
TS1_0558	CNTN6	3	1218279	2170519	DUP	952.2	CODING	N	Y	N	
TS2_0827	CNTN6	3	1226953	2170519	DUP	943.6	CODING	N	N	N	
TS2_0452	CNTN6	3	1260932	1556680	DUP	295.7	CODING	N	N	N	

Table 3-1. Clinical phenotypes of *NRXN1* and *CNTN6* CNV carriers.
Clinical phenotypes for all CNV carriers of the two significant TS loci detected in this study: deletions at *NRXN1* and duplications at *CNTN6*. Genomic location is given in hg19 coordinates. For each CNV carrier, the presence of common comorbid disorders for TS, ADHD and OCD is indicated, and atypical diagnoses are flagged and described (Notes). NA, No clinical information available.

3.9 Spatio-temporal brain expression patterns of *NRXN1* and *CNTN6*

We next examined patterns of normalized gene expression of the two genome-wide significant loci, *NRXN1* and *CNTN6,* across human brain development utilizing RNA-sequencing data from the Brainspan project. *NRXN1* is widely expressed across different regions of the prenatal human brain, with most prominent expression in mid-to-late gestational cerebral cortex

(17 post-conception weeks (pcw) to post-natal month 4), as well as in early-to-mid gestational thalamus (13-21 pcw). Postnatally, *NRXN1* is expressed at much lower levels compared to its expression in fetal brain, though is present across all cortical regions and cerebellum, with relatively little or no expression in samples from striatum, hippocampus, amygdala and thalamus (Figure 3-4.A). In contrast, *CNTN6* has more restricted gene expression in the prenatal human brain (Figure 3-4.B), with expression in striatum and cerebellum as early as 8-12 pcw, followed by high expression in thalamus from 13 to 21 pcw, and continued expression in cerebellum and thalamus through birth. *CNTN6* expression in fetal cortex appears to begin as early as 13 pcw in pre-frontal cortex (PFC), with highest expression at 21 pcw in medial PFC, orbital PFC as well as primary motor and sensory cortex. Postnatally, *CNTN6* is expressed widely across all samples except striatum, with most prominent expression in primary motor cortex, thalamus and cerebellum from age 13 to adulthood.

3.10 Discussion

The current study represents the largest survey of copy number variation in TS to date, and demonstrates a significant role for rare structural variation in the pathogenesis of this still poorly understood developmental neuropsychiatric disorder. We report the first two definitive TS risk variants surpassing the threshold for genome-wide significance, deletions in *NRXN1* and duplications in *CNTN6*, and show that CNVs in these two loci markedly increase TS risk in approximately 1% of cases. Furthermore, our CNV burden analysis provides convincing evidence that large CNVs lying within loci previously annotated to be clinically deleterious by ACMG criteria confer additional, high-impact TS genetic risk, as this increased burden remained significant even after removing all associated CNVs determined in this study (1.3% of cases vs. 0.7% of controls; OR per CNV of 2.0 [1.2-3.5], P=0.015).

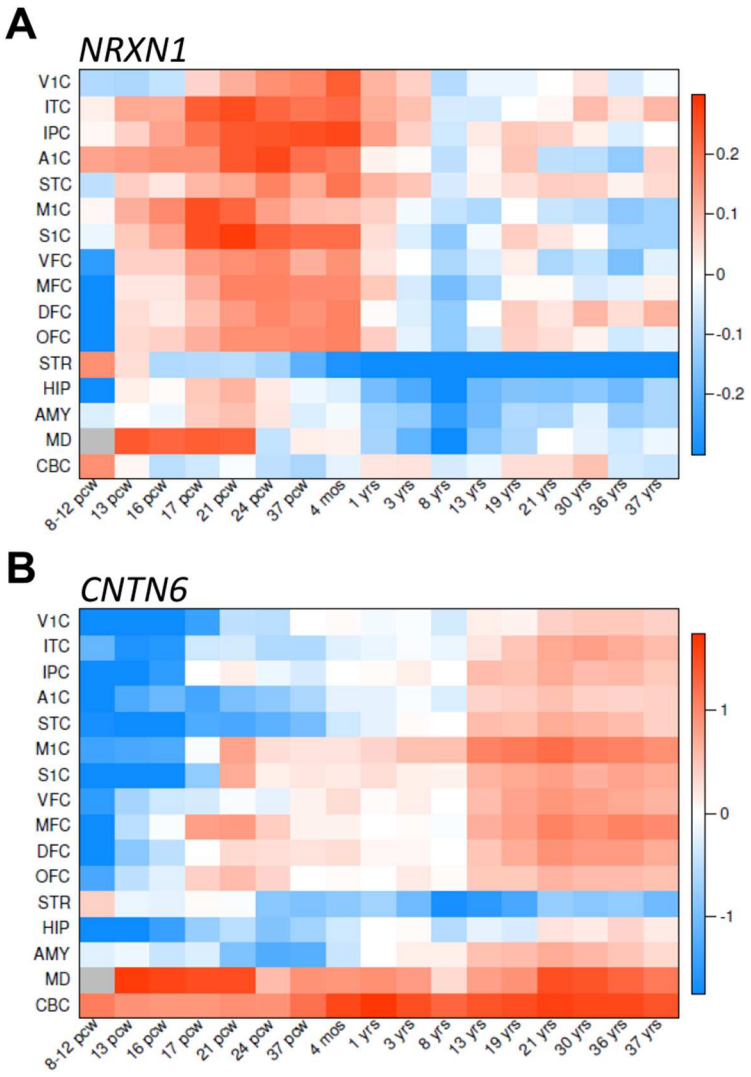


Figure 3-4. Spatio-temporal expression for TS-associated genes in the developing human brain. RNASeq expression data for (A) *NRXN1* and (B) *CNTN6* from Brainspan (www.brainspan.org). Color bar value indicates scaled expression level of mean-centered log-2 transformed Reads Per Kilobase of transcript per Million mapped reads (RPKM) values for each gene from all available samples across all tissues and developmental time points. Regions: primary visual cortex (V1C), inferolateral temporal cortex (ITC), posteroventral (inferior parietal) cortex (IPC), primary auditory cortex (A1C), posterior superior temporal cortex (STC), primary motor cortex (M1C), primary somatosensory cortex (S1C), ventrolateral prefrontal cortex (VFC), anterior cingulate (medial frontal) cortex (MFC), dorsolateral prefrontal cortex (DFC), orbital frontal cortex (OFC), striatum (STR), hippocampus (HIP), amygdaloid complex (AMY), mediodorsal nucleus of thalamus (MD), cerebellar cortex (CBC). Temporal abbreviations: post-conception weeks (pcw), months (mos), years (yrs).

Studies in other neurodevelopmental disorders demonstrating a global enrichment of CNV burden have shown that *de novo* variation plays a considerable role (Malhotra et al., 2011; Sanders et al., 2015). These highly-pathogenic variants are, on average, much larger in affected patients than in controls, and prior CNV study of 148 TS parent-proband trios has noted a similar phenomenon in TS, where the average size of *de novo* CNVs was approximately 20-fold greater in TS cases (Fernandez et al., 2012). In the current study, the observation that the excess TS CNV size burden is limited to large singleton CNVs is consistent with these variants being under strong negative selection, and suggests that these pathogenic CNVs have likely arisen either *de novo* or within a few prior generations.

Although previous studies have reported heterozygous exonic *NRXN1* deletions in a total of 6 TS patients (Nag et al., 2013; Sundaram et al., 2010), the small sample sizes in these prior studies precluded definitive association of this deletion with TS. We now demonstrate, in a sample sufficiently large to achieve genome-wide significance, that exonic deletions affecting *NRXN1*, particularly those spanning exons 1 and 2 of the *NRXN1α* isoform, confer a substantial increase in risk (approximately 20-fold) for the disorder. The association of heterozygous *NRXN1α* deletions with different neurodevelopmental disorders, including ID/DD, ASD, SCZ, epilepsy, and language delay, represents one of the most reliable findings regarding CNVs in neuropsychiatry (Ching et al., 2010; Glessner et al., 2009; Rujescu et al., 2009). Consistent with the view that CNVs in *NRXN1α* convey a broad risk for multiple neurodevelopmental disorders (NDDs), 4 of the 12 TS cases with exonic *NRXN1α* deletions in our sample had a previously diagnosed NDD, including two with Asperger syndrome, one with a speech/language disorder and one with unspecified developmental delay. The diverse clinical presentation exhibited by similar *NRXN1α* deletion carriers in the reported literature supports the hypothesis that these CNV deletions may interfere with a generalized neurodevelopmental process which, when combined with other disease-specific mutations and/or background polygenic risk, result in one

or more NDDs. Of note, our data suggest an approximately twofold higher prevalence of exonic *NRXN1α* deletions in TS compared to that of other neuropsychiatric disorders (Rujescu et al., 2009), though further studies in larger cohorts will be necessary to affirm this apparent comparative enrichment.

NRXN1 is a highly-studied, pre-synaptic cell-adhesion molecule involved in synaptogenesis and synaptic transmission at both glutamatergic and GABAergic synapses (Pak et al., 2015; Südhof, 2008). The *NRXN1* gene is transcribed from two alternative promoters, resulting in a full-length *NRXN1-α* isoform and a short C-terminal *NRXN1-β* isoform (Ushkaryov et al., 1992). *NRXN1-α* contains six alternative splice sites which, in combination, generate hundreds of unique transcripts that have been demonstrated to segregate within specific brain regions and cell-types (Fuccillo et al., 2015; Ullrich et al., 1995). These alpha neurexin isoforms also preferentially bind to various trans-synaptic partners, including neuroligins, cerebellins, neurexophilins and leucine-rich repeat transmembrane proteins, each of which subserve different synaptic functions (de Wit and Ghosh, 2016). Somewhat surprisingly, homozygous null *Nrxn1* mice form synapses and demonstrate only subtle impairments in synaptic transmission (Etherton et al., 2009). Similarly, heterozygous *NRXN1* LoF mutations in human iPSC-derived neurons *in vitro* have no effect on neuronal differentiation or synaptogenesis, but distinct from mouse models, display a consistent impairment of neurotransmitter release in a stimulus dependent pattern (Pak et al., 2015).

To date, *CNTN6* deletions, but not duplications, have been demonstrated to be significantly enriched in ASD cases (Mercati et al., 2016). *CNTN6* duplications have not been associated with other neuropsychiatric disorders. However, on the basis of structural variation in a small number of case reports, *CNTN6* has also been proposed to be a candidate gene for ID/DD, and both deletions and duplications have been observed (Hu et al., 2015; Kashevarova et al., 2014). Interestingly, of the patients within the ID/DD studies with detailed clinical

information, 7/9 (77%) presented with common TS comorbidities, including four with ADHD, two with OCD, and one with both. Thus, the significant enrichment of largely single-gene duplications in *CNTN6* presented here represents a novel NDD-associated CNV for TS.

Like *NRXN1*, *CNTN6* encodes a cell-adhesion molecule expressed primarily in the central nervous system (Ogawa et al., 1996). The contactin family is a member of the L1 immunoglobulin superfamily of proteins, and can either be membrane-bound via a glycosyl-phosphatidyl inositol anchor or secreted (de Wit and Ghosh, 2016). Mouse *Cntn6*, also known as NB-3, is expressed at low levels prenatally, with peak expression in the forebrain at P7 in cortical layers II/III, V and the subplate of the neocortex, as well as in the accessory olfactory bulb, piriform cortex, CA1 region of the hippocampus, amygdala, and several thalamic nuclei, with strongest expression in the anterior nucleus of the thalamus (Lee et al., 2000). *Cntn6* is also expressed in the Purkinje cell layer of the cerebellum and deep cerebellar nuclei, though its expression in cerebellum onsets post-natally and reaches maximal expression in adulthood. *Cntn6* has multiple functions in the developing mouse nervous system, including orientation of apical dendrites in cortical pyramidal neurons (Ye et al., 2008), regulation of Purkinje cell development and synaptogenesis, (Sakurai et al., 2009) and oligodendrocyte differentiation from neuroprogenitor cells (Cui et al., 2004). Mice with homozygous inactivation of *Cntn6* display motor dyscoordination and delayed corticospinal tract formation (Huang et al., 2012; Takeda et al., 2003).

How might the significant CNVs observed in this study inform our understanding of TS pathogenesis? As *NRXN1* is widely expressed across the developing human brain and engages in multiple functions in different cell types and brain regions, inference as to the specific brain regions, neurodevelopmental time points and underlying mechanisms through which *NRXN1* promotes TS pathogenesis can only be speculative at present. However, the prominent expression of *NRXN1* in primary sensorimotor cortex (Figure 3-4.A, M1C, S1C) and thalamus in

the mid-fetal period would fit with the prevailing theory of TS as arising from abnormal sensorimotor CSTC circuit development. In particular, Singh et al. have recently demonstrated a critical role for *NRXN1-α* trans-synaptic interactions in thalamo-cortical synaptogenesis and plasticity (Singh et al., 2016).

In contrast to *NRXN1*, *CNTN6* expression is more restricted in the prenatal human brain, with highest expression in thalamus, followed by primary sensorimotor cortex, orbitofrontal and medial frontal cortex, and cerebellum during the mid-fetal period (13-21 pcw) (Figure 3-4.B). Of note, the prominent *CNTN6* expression in mid-fetal thalamus matches the timing of high *NRXN1* expression in this region. The thalamus has long been implicated in TS pathophysiology based on multiple levels of evidence including thalamic lesions, human physiology, and by recent treatment successes using deep-brain stimulation (Gunduz et al., 2015; Hassler and Dieckmann, 1970). In addition, a recent meta-analysis of voxel-based morphometry studies in 103 pediatric TS patients and an equal number of age- and study-matched controls found the strongest association between TS and increased gray matter volume in the posterior thalamus (corrected P=0.001) (Greene et al., 2016). Postnatally, in additional to strong expression in primary motor cortex, thalamus and cerebellum, human *CNTN6* is much more broadly expressed throughout the adolescent and adult cortex, which may parallel the emerging cognitive control networks that mark maturation of the adolescent and adult brain, and which have been proposed to correspond with the psychological development from childhood and adolescence to adulthood (Casey et al., 2016; Draganski et al., 2010; Marsh et al., 2009). Furthermore, Peterson and colleagues have hypothesized that impaired development of top-down cognitive control circuitry may be a key contributor for persistence of severe TS in adulthood as opposed to the 2/3^{rds} of TS patients whose symptoms improve or resolve in late adolescence/early adulthood. Further work will be needed to dissect which brain region(s) and developmental period(s).

Several limitations of the current study can inform future inquiry. First, although our sample represents the largest survey of CNVs in TS to date, it is still underpowered to detect extremely rare CNVs and/or those of moderate effect. While we show strong evidence for the involvement of deletions across *NRXN1*, our data also does not support other similarly implicated loci, including deletions in *COL8A1* (Nag et al., 2013), which we in a single TS patient. This emphasizes the need for larger, independent samples for continued discovery and refinement of candidate TS loci. Second, while our TS cases were well characterized for OCD and ADHD, we did not formally assess ASD, ID, SCZ or epilepsy. Parents and/or adult subjects were queried about existing diagnoses of these NDDs as well as learning disorders/developmental delay, but cases with milder ASD/DD may not have been detected. Additional efforts should focus on characterizing the full scope of phenotypes associated with *NRXN1* and *CNTN6* CNVs. A comprehensive molecular analysis of these CNVs is also needed to fully understand how they increase risk for such phenotype(s). Finally, the elevated burden observed here is largely confined to large singletons and known pathogenic CNVs, consistent with a global enrichment of CNVs under strong negative selection that possibly arose *de novo* or within the last few generations. This suggests that, in addition to substantial increases in sample size, alternative study designs that allow for discrimination of *de novo* CNVs will be fruitful for TS, as was recently shown by our group for likely-gene-disrupting variants identified by exome sequencing in TS trios (Willsey et al., 2017).

3.11 Methods

Sample ascertainment

TS cases (n=2,243) were ascertained primarily from TS specialty clinics through sites distributed throughout North America, Europe and Israel as part of an ongoing collaborative effort by the TSAICG (https://www.findtsgenes.org/) as described previously (Scharf et al.,

2013). Subjects were assessed for a lifetime diagnosis of TS, OCD and ADHD using a standardized and validated semi-structured direct interview (TICS Inventory) (Darrow et al., 2015). TS case samples were also obtained through web-based recruitment of individuals with a prior clinical diagnosis of TS who subsequently completed an online questionnaire that has been validated against the gold-standard TS structured diagnostic interview with nearly 100% concordance for all inclusion/exclusion criteria as well as high correct classification rates for DSM-IV diagnoses of OCD and ADHD (Darrow et al., 2015). Individuals for web-based screening were solicited through the Tourette Association of America mailing list as well as from 4 TS specialty clinics in the United States. Individuals with a history of intellectual disability, seizure disorder, or a known tic disorder unrelated to TS were excluded.

Additional cases (n=628) and ancestry-matched controls (n=544) were collected by the GGRI through 9 TS specialty clinics in Austria, Canada, France, Germany, Greece, Hungary, Italy and the Netherlands by expert clinicians using Tourette Syndrome Study Group criteria for Definite TS (DSM-IV TS diagnosis plus tics observed by a trained clinician) as well as DSM-IV diagnostic criteria for OCD and ADHD as described previously (Paschou et al., 2014). We did not conduct formal standardized assessments for other NDDs such as ID/DD, ASD, or SCZ/childhood psychosis; however, participants and their parents were asked about the presence of established or suspected diagnoses. 74.3% of TS cases were male, consistent with the 3- to 4-fold higher prevalence of the disorder in males compared to females (Robertson et al., 2017). The median age of cases was 17 (IQR, 12-31).

External control sets

Additional control subjects were taken from four external large-scale genetic studies consisting of individuals sampled from similar geographic locations, specifically selected

because intensity data was available and generated on the same Illumina OmniExpress platform as the TS cases and controls collected as part of this study:

1. Cardiff Controls (CC) (Green et al., 2010): UK Blood donors were recruited in Cardiff at the time of blood donation at centres in Wales and England. Although not explicitly screened for psychiatric disorders, these controls are likely to have low rates of severe neuropsychiatric illness, as blood donors in the UK are only eligible to donate if they are not taking any medications.
2. Consortium for Neuropsychiatric Phenomics (CNP) (Bilder et al., 2009): A collection of neuropsychiatric samples composed of patients with ADHD, bipolar disorder (BD), schizophrenia (SCZ), and psychologically normal controls, collected throughout North America as part of a large NIH Roadmap interdisciplinary research consortia centered at the University of California, Los Angeles (UCLA).
3. Genomic Psychiatry Cohort (GPC) (Pato et al., 2013): A large, longitudinal, population resource composed of clinically ascertained patients affected with BD, SCZ, their unaffected family members, and a large set of control samples with no family history of either disorder. Samples were collected at various sites throughout North America in a National Institute of Mental Health-sponsored study lead by the University of Southern California.
4. Wellcome Trust Case Control Consortium 2 (WTCCC2) (Power and Elliott, 2006): A subset of control samples from the National Blood Donors Cohort.

Sample genotyping

Samples selected for this study were all genotyped on the Illumina OmniExpress Exome v1.1 (Illumina, San Diego, CA, USA) in three separate batches (TS1-3), while samples from the CC, CNP, GPC, and WTCCC2 sets were genotyped on the Illumina OmniExpress 12v1.0

(Illumina, San Diego, CA, USA). A summary of the datasets, genotyping centers, and arrays used is provided in Table S1A. The OmniExpress Exome and OmniExpress arrays are identical except for the presence of exome-focused content on the former and additional intensity-only markers on the latter. We have observed that exome-specific assays in general exhibit a much higher variance overall in their LRR values. Therefore, in order to avoid detection biases due to this differential variance and to unequal probe density, only the SNP assays with common identifiers for all array versions across all datasets were used for QC and CNV detection, a total of 689,077 markers.

To ensure the generation of the most reliable SNP calls, intensity measures, and BAF, a custom cluster file was generated for each dataset separately and for each genotype batch when such information was available. Since the performance of Illumina's proprietary normalization and cluster generation process improves with the number of samples, we processed all of the raw intensity data available, regardless of clinical phenotype. An initial round of QC was carried out using Beeline v1.0 (Illumina, San Diego, CA, USA) to determine baseline call rates and LRR_SD for each sample using the canonical cluster file (*.egt) provided by the manufacturer for each array version. Any sample with a call rate < 0.98 or an LRR_SD > 0.30 was deemed a failed assay and removed (Pre-cluster QC, Table S1B). SNP clustering and genotype calling was then performed with only the passing samples for each dataset individually with GenomeStudio v2.0 (Illumina, San Diego, CA, USA)

SNP-based sample QC

We performed an initial round of QC based on SNP genotype data. All samples at this stage had a minimal call rate >0.98. Samples with a discordant sex status were excluded. For samples run in duplicate, we retained the assay with the higher call rate. We filtered autosomal SNPs for missingness, MAF, and deviation from Hardy-Weinberg equilibrium (HWE) before

pruning SNPs for linkage disequilibrium (LD) using the following options in PLINK v1.9 (Chang et al., 2015): "*plink --geno 0.02 --maf 0.01 --hwe 0.000001 --indep 50 5 1.5*". A total of 96,350 SNPs remained for Identity-By-State relatedness testing. For all pairs of subjects (and duplicate/repeated samples) with PI_HAT > 0.185, we removed the sample with the lower call rate, with the exception that if a control individual was related to a subject with a neuropsychiatric phenotype, it was explicitly removed. Additionally, we excluded 44 subjects that were identical (PI_HAT > 0.99) between the CNP and GPC cohorts.

Ancestry estimation

Following the exclusion of all clinical non-TS samples from external studies, genotype data for the remaining samples was combined with data from publicly available continental HapMap samples of CEU, YRI, and CHB/JPT ancestry genotyped on the OmniExpress array (Illumina, San Diego, CA, USA). Available European (EU) samples from the 1000 Genomes Project (http://www.internationalgenome.org/) genotyped on the Omni platform were also included to establish an appropriate calibration threshold for EU ancestry designation. We thinned our dataset randomly to 19,024 LD-independent markers for ancestry inference using fastStructure (Raj et al., 2014) with k=3. Samples were excluded if they contained > 0.0985 non-EU ancestry (Figure A-1.B). A final round of ancestry exclusion was performed by removing all samples outside of the median +/- 3SD on the first three ancestry principal components (PCs).

CNV calling

We created GC wave-adjusted LRR intensity files for all samples using PennCNV's *genomic_wave.pl* script, and employed two widely-used HMM-based CNV calling algorithms, PennCNV v2011-05-03 (Wang et al., 2007), and QuantiSNP v2.0 (Colella et al., 2007) to initially detect structural variants in our dataset. For PennCNV, we generated a custom population BAF file for each dataset separately before calling CNVs using emission probabilities defined in the

file *hhall.hmm*. QuantiSNP calls were generated from the GC-adjusted intensity files using the configuration file *levels-hd.dat*. A Perl script was used to merge concordant calls generated by both algorithms. CNVs were merged by taking the intersection of overlapping calls of the same CNV type (deletion or duplication). Additionally, adjacent CNV calls were merged if they were spanned by a CNV called by the other HMM algorithm. As HMMs have been shown to artificially break up large CNVs, we also merged CNV segments in the final concordant callset if they were of the same copy number and the number of intervening markers between them was less than 20% of the total of both segments combined using the PennCNV's *clean_cnv.pl* script. We repeated this joining process iteratively until no more merging of segments occurred.

Intensity-based sample QC

The PennCNV calling algorithm generates a number of array intensity-based metrics with regard to CNV assay quality. Intensity-based QC was conducted based on the distribution of all available assays and subsequently combined with the results from the SNP-based QC. To remove samples with data unsuited for CNV detection, we used empirically defined thresholds across several different metrics: WF, LRR_SD, and BAF_SD. Thresholds for were determined separately for each dataset by both manual examination of QC metrics and/or taking the mean +3x SD to determine outlying samples. Following intensity-based QC, all samples had an LRR-SD of <0.24, absolute value of WF < 0.04, and a BAF_SD < 0.04, well within established limits required for reliable CNV detection.

CNV-load sample QC

Although SNP and intensity-based QC removed most failed assays, we performed a final round of sample QC, removing eight additional samples with excessively high CNV load based on the total number of CNV calls (>45) or total CNV length (>10Mb). These thresholds

were determined empirically by visual inspection of distributions across all datasets combined. Our final dataset after QC consisted 6,527 samples: 2,434 TS cases and 4,093 controls.

Sensitivity assessment

We augmented a previously described method (Vacic et al., 2011) to investigate whether any difference in sensitivity to detect CNVs existed between cases and controls within the context of our study. Both HMM-based CNV callers we employed for genome-wide detection are univariate methods completely agnostic of intensity information across multiple samples and do not use known population frequency prior probabilities in their calling algorithms. Therefore, common CNVs act as an ideal proxy to evaluate the effectiveness to detect rare events accurately; they are detected in the same manner but are present at much higher frequencies, enabling an accurate estimation of the overall sensitivity of detection for rare events genome-wide.

We used the UCSC Genome Browser (https://genome.ucsc.edu) liftOver tool to translate a list of common HapMap3 CNVs to the hg19 reference. To match the thresholds used for our association tests in this study, we filtered the list of common CNVs to those that were >30 kbp in length. We reduced the number of markers required slightly to a minimum of 9 to ensure that an adequate number of events could be assessed. For each common CNV meeting these criteria, we examined the distribution of median-summarized normalized intensity measures within the CNV region across all study samples and retained only those loci that displayed no evidence of clustering intro different copy-number states. A total of 11 common CNV loci were retained for sensitivity analysis (Figure A-2).

We generated locus-specific genotyping calls in the following manner. First, we extracted the LRR intensity Z-scores for all probes in the region across all samples. The Z-scores for all probes spanning the CNV locus were then subjected to second round of

normalization across all samples. A GMM was fit to this distribution of Z-scores using the SciKit-learn Python package. The optimum number of clusters was automatically determined by minimization of the Bayesian Information Criterion (BIC) and corrected, when necessary, by manual adjustment. Individuals were assigned to a cluster only if the posterior probability of assignment exceeded 0.95.

Copy number state was inferred by examining the original LRR intensity values for samples within each cluster. We inspected for allele frequency differences between controls and cases for all clusters and found no significant difference (Fisher exact Test, Table A-2). We collapsed the clusters at each locus into CNVs of the same type (deletion or duplication). As this locus-specific genotyping method is more sensitive than HMM segmentation methods, we used the proportion of concordant of HMM-based calls as a proxy for genome-wide detection sensitivity. We found no significant difference in sensitivity to detect common CNVs between phenotypic groups at any of the 11 loci tested, either independently, or in concert (Fisher's exact test, Table A-3.A and B). Furthermore, the mean sensitivity for each sample was calculated and collectively assessed for any systematic difference between phenotypic groups. Considering duplications, deletions, or both in concert, we observed no significant difference in the sensitivity of segmentation calls between case and control groups (Welch's t-test, Table A-3.C).

Call filtering and delineation of rare CNVs

Calls were removed from the dataset if they spanned less than 10 markers, were less than 30kb in length, or overlapped by more than 0.5 of their total length with regions known to generate artifacts in SNP-based detection of CNVs. This included immunoglobulin domain regions, segmental duplications, and regions that have previously demonstrated associations specific to Epstein-Barr virus immortalized cell lines (Shirley et al., 2012). In addition, we removed calls that spanned telomeric (defined as 100kb from the chromosome ends),

centromeric regions, and gaps in the reference genome. We assigned all CNV calls a specific frequency count using PLINK v.1.07 (Purcell et al., 2007), with the option "*--cnv-freq-method2 0.5"*. Here, the frequency count of an individual CNV is determined as 1 + the total number of CNVs overlap by at least 50% of its total length (in bp), irrespective of CNV type. We then filtered our callset for CNVs with MAF < 1% (a frequency of 65 or lower across 6,527 samples). Furthermore, we removed calls if they shared more than a 50% reciprocal overlap in length to common CNVs regions derived from several large, publicly available SNP-array datasets, compiled by the Database of Genomic Variation and Phenotype in Humans using Ensembl Resources (DECIPHER; https://decipher.sanger.ac.uk/index):

1. 845 population samples from the Deciphering Developmental Disorders study.
2. 450 population samples from the 42M genotyped study.
3. 5919 population samples from the Affy6 study.

Conservatively, CNV regions were only considered common if present at a frequency of >10% within any individual dataset above.

<u>In-silico validation</u>

For each putative rare CNV, we generated two different metrics based on LRR intensity and BAF banding. To qualify CNVs based on intensity, we adopted a scoring methodology similar to the MeZOD method described elsewhere (Kirov et al., 2012), with modification. We observed that standardized intensity measures typically range from < -20 for homozygous deletions, [-6,-2.3] for heterozygous deletions, and > 1.3 for duplications. Because of the disproportionately large effect on intensity measures caused by homozygous deletions events in Illumina data, performing a second round of normalization across all samples within each putative CNV will skew the overall distribution when these events are present. Therefore, we only performed a single round of normalization of LRR intensity measures, within each sample.

Each CNV was scored by calculating the median of LRR intensity Z-scores (LRR-Z) for all probes within the region. To qualify CNVs based on BAF banding, we calculated the proportion of probes within the CNV region that showed evidence of a duplication event (BAF of [0.25-04] or [0.6-0.8]), and denoted this measure "BAF-D."

Based on thresholds established through manual inspection of large CNVs, we flagged deletions that had a LRR-Z > -2 or BAF-D > 0.02, and duplications with a LRR-Z < 1 or BAF-D < 0.1. To avoid differential missingness caused by subtle differences between datasets, we did not impose any hard cutoff for CNV exclusion based on these metrics. These thresholds were applied to flag CNV calls with marginal scores for manual inspection for the exclusion of obviously misclassified events. Through this in-silico validation process we discovered multiple instances of large CNV calls likely due to individual mosaicism (Figure A-3.A and B), and removed these events from subsequent analysis.

Furthermore, for each rare CNV call, we used distribution of summarized intensity information across all individuals. For each rare CNV region, we quantified proportion of samples whose LRR-Z metric fell outside of [-2.3, 1.3] and further inspected these regions manually. Putative rare CNV loci that showed substantial evidence for extensive polymorphism were subsequently scored for frequency using the GMM genotyping method described above (Figure A-3.C) and conservatively, only removed if shown to be variant in more than 10% of the samples across the entire dataset.

Annotation of rare CNVs

Rare CNVs were annotated for gene content according to RefSeq provided by the hg19 assembly of the UCSC Genome Browser. We only considered a CNV as "genic" if it overlapped any exon of a known protein-coding transcript (as designated by the RefSeq transcript accession prefix "NM"). The "gene count" of a CNV represents the total number of non-

redundant genes whose respective transcripts it overlaps. “Non-genic CNVs” represent all variants that are not genic according to the definition above.

In addition, all rare CNVs were assessed for clinical relevance in accordance with guidelines set forth by the American College of Medical Genetics (ACMG) (Kearney et al., 2011). This was accomplished through the use of the Scripps Genome Advisor (https://genomics.scripps.edu/ADVISER/); inspection of curated resources including ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), DECIPHER (https://decipher.sanger.ac.uk/index), and Online Mendelian Inheritance in Man (https://www.omim.org/); and followed by confirmation through review of primary literature. Conservatively, we only considered a CNV as “pathogenic” if directly supported by more than one primary publication. To screen for additional relevant pathogenic variants, we included variants that overlapped compiled lists defined in the literature (Malhotra and Sebat, 2012; McGrath et al., 2014). All non-pathogenic variants were automatically annotated by SGA: those classed as Category 2-4 were assigned as a variant of “unknown clinical significance”, and those assigned to Category 5 were considered “benign”. Note that, as only rare CNVs were considered, the number of “benign” variants is small. Rare genic CNVs were further annotated using pLI scores (http://exac.broadinstitute.org/). A CNV was marked as “constrained” if it overlapped any exon of a gene with a pLI score of > 0.9.

Global CNV burden analysis

We measured global CNV burden using three separate metrics: the number of rare CNVs (CNV count), the total length of all CNVs (CNV length) and the total number of genes intersected by CNVs (CNV gene count). To examine the effect of different covariates on different metrics of global CNV burden, we first fit a linear regression model for each burden metric, using the *glm* function in R:

Burden_metric ~ subject_sex + ancestry_PCs + LRR_SD

In the above model, ancestry_PCs included the first ten PCs derived from SNP data and, as various assay intensity quality metrics are highly correlated, LRR_SD was used as a single measure of assay quality. Across the entire dataset, only PC2 and LRR_SD were associated ($P<0.05$) with global CNV burden as measured by CNV count. To assess for a global burden difference between TS cases and controls, we fit the following general logistic regression model:

$$TS\_status \sim Burden\_metric + subject\_sex + PC1 + PC2 + PC3 + PC4 + LRR\_SD$$

Independent variables included burden metric, subject sex, and the first 4 ancestry PCs. Even though assay quality (LRR_SD) was associated exclusively with small CNV burden (<100kbp), it was included in all burden analysis, regardless of size. For all burden analyses, ORs, 95% CIs, and significance were calculated using the R function *glm (family="binomial")*. ORs were calculated by taking the exponential of the logistic regression coefficient.

To examine whether evolutionarily constrained CNVs are enriched in TS patients, we assumed a Poisson distribution of such rare events, and conducted a Poisson regression using the R function *glm (family="poisson")* with the same covariates as in the logistic regression above, including adjustment of subject sex, the first four ancestry PCs, and LRR_SD:

$$Constrained\_CNVs \sim TS\_status + subject\_sex + PC1 + PC2 + PC3 + PC4 + LRR\_SD$$

We tested large (>1Mb), singleton CNVs that affect conserved genes (pLI>0.9), stratified by CNV type, and compared the relative risk of such events to all large, singleton genic CNVs.

In Figure 3-1, global burden was analyzed separately for genic CNVs and non-genic CNVs. Since we were interested in comparing the relative contribution to TS risk by different measures of burden and across various categories, sizes, and frequency classes, the ORs presented in Figure 3-1 and Figure A-4 are calculated from standardized CNV burden metrics.

For Figure 3-2 and elsewhere, we calculated ORs using the unstandardized value of actual CNV counts, as this is directly interpretable.

Locus-specific tests of association

The segmental test of association was performed by quantifying the frequencies of case and control CNV carriers at all unique CNV breakpoint locations; the unique set of CNV breakpoints defines all locations genome-wide where the frequency of CNVs can differ between cases and controls. For gene-based association tests, we restricted our analysis to genic CNVs (CNVs that intersect an exon of any protein-coding transcript, as defined above) and quantified the frequencies of cases and control CNVs across each gene. Locus-specific p-values for both tests of association were determined by 1,000,000 permutations of phenotype labels, and genome-wide corrected p-values were obtained using the max(T) permutation method as implemented in PLINK v1.07, which controls for family-wise error rate by comparing the locus-specific test statistic to all test statistics genome-wide within each permutation. Association tests were conducted separately for deletions and duplications.

Sensitivity analysis of association Results

The segmental association test was repeated after carefully pair-matching each case with a control such that the global difference between each pair was minimized using GemTools (Lee et al., 2010). For the matched segmental association analysis, because of the drastic reduction in sample size, a genome-wide corrected alpha < 0.05 was used as a cutoff to indicate significance.

Chapter 4:

Identification of De Novo CNVs in Sporadic TS Trios Using WES

4.1 Abstract

In Chapter 3, we discovered the first two genome-wide significant loci in TS: deletions in *NRXN1* and duplications in *CNTN6*. Both confer a substantial increase in risk for the disorder. CNVs across both genes have been described in the literature to occur as *de novo* events. Moreover, we showed an increased burden for TS in large, singleton events and known, pathogenic CNVs. Collectively, these findings suggest that *de novo* variation may play a significant role in the etiology of the disorder. Here, we use 164 affected proband, unaffected parent trios to detect *de novo* CNVs using WES. Although our sample is too small to conclusively assess the risk conferred by such variants, we describe four *de novo* events that pass stringent QC procedures, all of which affect genes that are highly expressed in the brain and involved in neurodevelopmental processes. These results suggest that the *de novo* paradigm may be a fruitful means of future gene discovery in TS.

4.2 Background

WES has demonstrated to be an enormously effective means for elucidating causal mutations in a number of rare, mendelian disorders (Angelika et al., 2012; Bamshad et al., 2011; Gilissen et al., 2011). By targeting and resequencing only annotated exonic regions which, collectively, comprise only 1% of the genome, WES allows for the comprehensive characterization of rare, genic variation genome-wide without prior knowledge of the variants of interest in an extremely efficient and economical fashion when compared to WGS. In fact, WES has allowed for the precise identification of causal, highly penetrant coding mutations in large pedigrees where linkage, identity-by-descent, and other forms of genetic inquiry were previously able to identify region(s) likely harboring highly penetrant, pathogenic variants. Perhaps less initially obvious would be the way in which low-cost WES on a large scale would inform our understanding of complex, polygenic neuropsychiatric developmental disorders. In 2012, four landmark papers on the contribution of de novo SNVs and small indels to ASD were published (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012), largely based on the sequencing of simplex families derived from the Simons Simplex Consortium. *De novo*, likely gene disrupting (LGD) SNVs and indels were observed in approximately 20% of sporadic ASD cases, nearly twice the rate compared to control trios (Ronemus et al., 2014). This collective result was consistent with earlier ASD studies based on de novo CNVs (Marshall et al., 2008; Sebat et al., 2007). Since de novo LGD and large, likely pathogenic CNVs are, on average, exceptionally rare in a given individual, identification of recurrent mutations affecting the same gene or locus can provide strong statistical support for implicating specific genes important for disease pathogenesis (Sanders et al., 2015).

Work by our group has demonstrated that a similar contribution of *de novo* variation exists for TS. Through WES sequencing in 511 sporadic TS trios, we observed a 2.32-fold increase in the rate of de novo LGD variants compared to controls (Willsey et al., 2017). There

is reason to believe that this *de novo* paradigm extends to other types of rare variation as well, specifically CNVs. We previously identified two rare genic CNVs significantly associated with TS, deletions in *NRXN1* and duplications at *CNTN6*, each of which confers a substantial increase in disease risk, with *NRXN1* deletions being the largest risk-factor identified to date. Of note, CNVs at both loci have been described in cases with other neurodevelopmental disorders, including ASD, epilepsy, SCZ, and ID/DD, and have been known to occur among affected individuals as de novo events. We also observed a significant increase in global CNV burden that was largely attributable to large, ultra-rare singleton events and CNVs at known, clinically relevant, pathogenic loci. This suggests that this enrichment is due, in part, to variants under strong selective pressure and possibly de novo events.

Conditioned on such observations, here we sought to identify additional genes potentially involved in TS pathogenesis by identifying de novo CNVs in a set of 164 sporadic TS trios. We used XHMM (eXome-Hidden Markov model) (Fromer et al., 2012) to detect and genotype genic CNVs in WES data. After performing stringent call filtering by calibrating against expected transmission rates and removal of potential artifacts based on genomic context, we identified four high-confidence de novo events affecting genes highly expressed in brain and known to be involved in axon guidance, neurodevelopment, and cognitive function. Furthermore, our results provide preliminary support that expanded efforts to identify the complete spectrum of *de novo* variation in sporadic cases TS will allow for systematic identification of novel risk genes.

4.3 Sample sequencing and CNV calling

The trios analyzed here are a subset of samples collected for a previous study examining the role of rare, *de novo* SNVs and indels of TS for which WES had been performed (Willsey et al., 2017). All trios analyzed represent sporadic cases (proband) and unaffected

parents. Capture libraries and sequencing data were generated on complete sets of trios at two different sites. Exonic DNA was enriched using Agilent SureSelect All Exon v1.1 and sequenced at the Broad Institute (Broad, 135 trios), or using Nimblegen EZ Exome v3 and sequenced at the University of California, Los Angeles (UCLA, 29 trios). After sequence alignment using identical parameters, read coverage was quantified over all exonic targets (n=189,894 and 242,232 for Broad and UCLA datasets, respectively) and CNVs were identified and subsequently genotyped in all samples using XHMM. Given the substantial differences in target regions, capture size, and hybridization technologies, we performed PCA normalization of read depth data independently for each dataset (Figure 4-1). With the intent to perform stringent post-calling filtering, we set an initially low threshold for site discovery and genotyping (phred-scaled quality score, or QS > 5), and assessed the transmission rate across trios for different QS cutoffs in each dataset separately. Consistent with previous reports using similar capture technology, the Broad samples approached the theoretical parent-to-child transmission rate of 0.50 at a QS > 60 (Figure 4-2.A), whereas the UCLA samples required much more stringent filtering (Figure 4-2.B, QS=75).

4.4 Identification and validation of de novo events

Datasets based on different captures were very different with regard to overall CNV number and total CNV length and therefore, were processed independently. To identify putative *de novo* events of high confidence, we first filtered each callset for variants of relative rarity, defined as a MAF < 10% across all samples in each dataset based on a 50% reciprocal overlap. Events were considered *de novo* only if the QS of consistent genotypes (including non-calls) exceeded 60 (75 for the UCLA dataset). Annotating our putative *de novo* list with genomic context revealed that a large number of these events coincided with regions known to cause

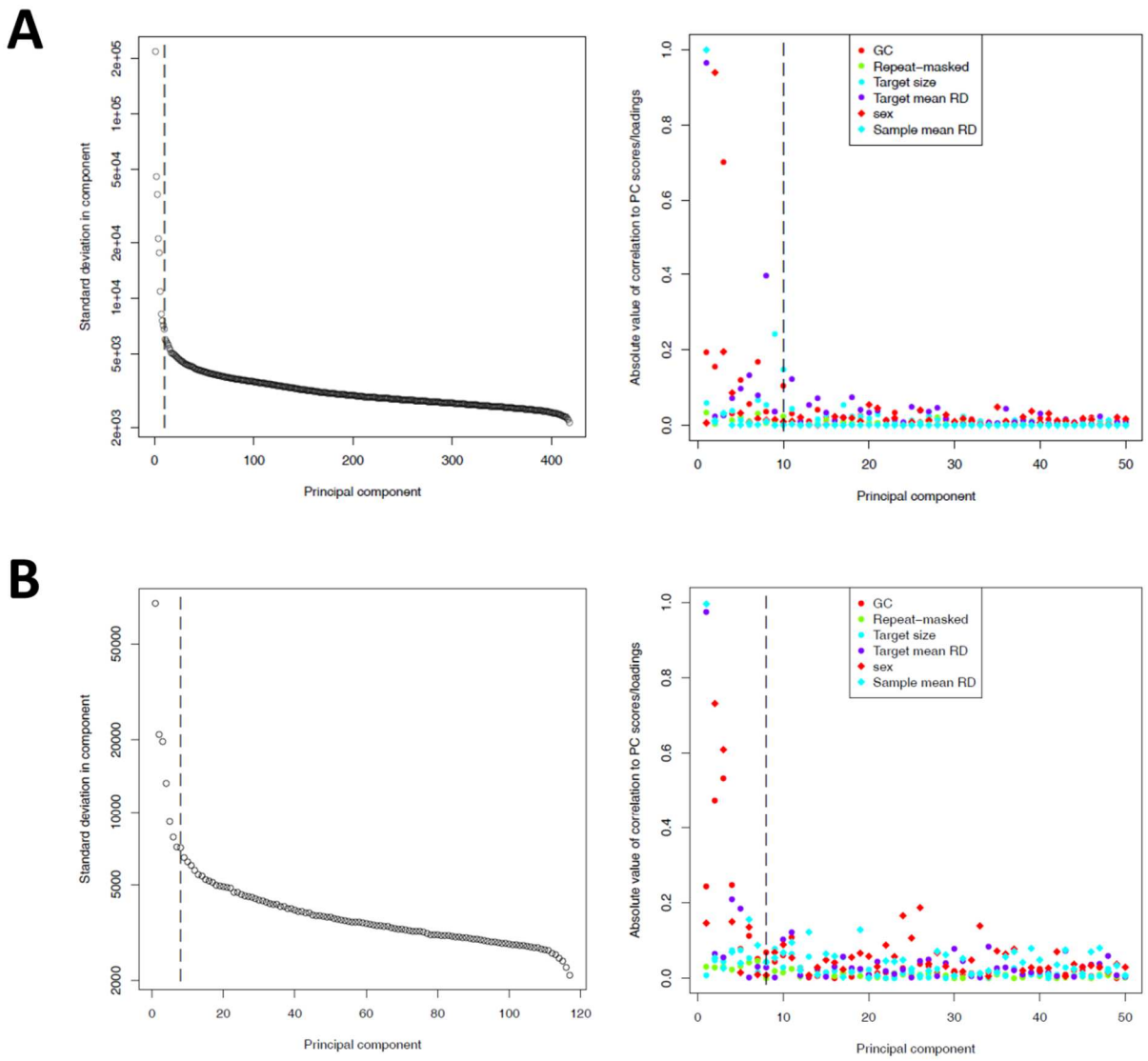


Figure 4-1. PCA normalization of read depth data.

A "scree" plot (left) displays variance in read-depth data plotted against each principal component, and the correlation PC score/loadings to known target and sample features is shown to the right, for the Broad (A) and the UCLA (B) datasets. The number of PCAs removed is indicated by the dashed line.

artifacts in CNV detection. Out of 53 events, 31 (58.5%) overlapped with segmental duplications by more than half of their length, and 21 (20.8%) occurred in VDJ recombination regions. 37 events were also observed in an unaffected individual (68.5%) (Table 4-1). After removing these

events and further restricting our consideration to CNVs > 1kb and spanning at least three consecutive exonic targets, 4 putative CNVs remained (Figure 4-1C). Microarray data with sufficient coverage was available for one of these *de novo* CNV carriers, confirming the presence of this variant (Figure 4-1.D).

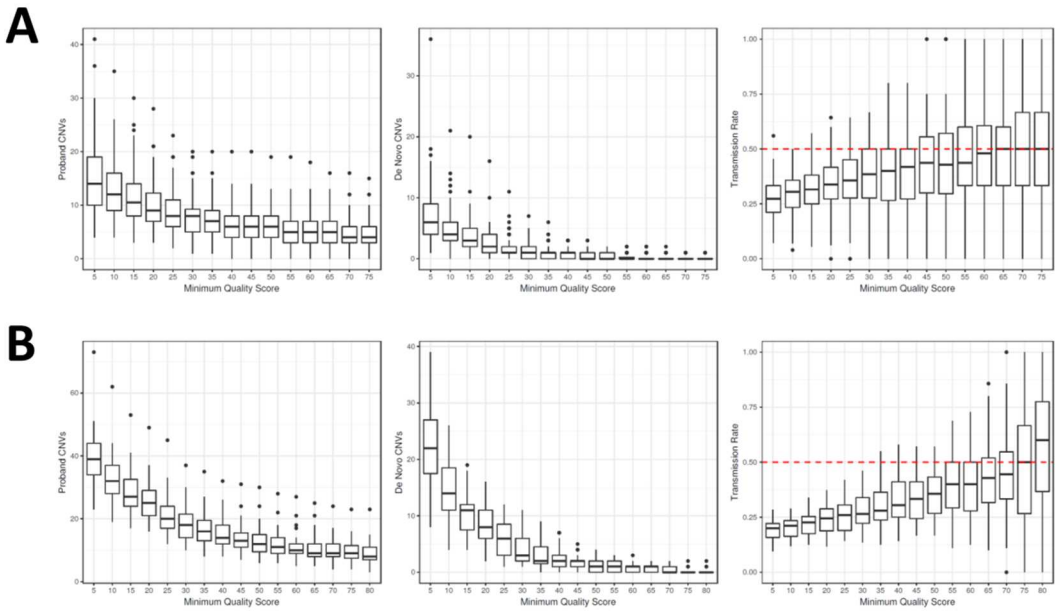


Figure 4-2. Identification of high-confidence de novo CNVs in 164 sporadic TS trios.
Titration plots displaying the number of called proband CNVs (left), number of de novo CNVs (middle), and the parent-to-child transmission rate (right) are shown for the Broad (A) and UCLA (B) datasets across various QS cutoffs.

4.5 Discussion

In this pilot study, we used WES data from 164 trio samples to perform genome-wide detection of exonic CNVs and identify *de novo* variants. After careful calibration of QS thresholds and further filtering of putative events, we resolved 4 high-confidence de novo CNVs in 164 trios (2.4%). Although additional statistical evidence from additional samples will be needed to confirm the involvement of these variants in TS, existing knowledge of the affected genes indicates that they all make plausible candidates.

The first of these (Figure 4-2.B) is a single gene duplication encompassing the *PLXNA1* gene. PLXNA1 is an axon guidance receptor that is widely expressed during early neurodevelopment, and extremely intolerant to loss-of-function mutations (pLI=1.0). Controlled expression of *PLXNA1* thought to be critical to the proper establishment of fine motor control in higher-order primates (Gu et al., 2017). The second is a large duplication encompassing *PRDM9*, *CDH10*, and *CHD9* (Figure 4-2.C). *CDH10* is highly expressed in the cortex, and intronic variants between *CDH9* and *CDH10* were the first common variants to be significantly associated with ASD through GWAS (Wang et al., 2009). The third is a single-gene duplication encompassing *BOD1* (Figure 4-2.D), a brain-expressed gene encoding a protein that localizes to synapses and neurons. Homozygous mutations in *BOD1* have been shown to perfectly segregate in a multi-generational family affected with ID (Esmaeeli-Nieh et al., 2016). The fourth is a 200kb duplication on 19q13.42 that affects multiple genes (Figure 4-2.E), most of which have high expression in the brain. One of these, *SHISA7*, directly associates with AMPA-type glutamate receptors, and is important for glutamatergic synaptic plasticity in the hippocampus and the functioning of contextual memory (Schmitz et al., 2017).

Although we do not know the precise extent or size of the CNVs detected by WES, the majority of the events were small (< 50kb), and near the limit typically observed for CNV detection using microarray. Several studies have now demonstrated an increased burden of smaller CNVs in the pathogenesis of ASD, highlighting the need to explore all classes of structural variation among neurodevelopmental disorders that technology makes accessible. A key question moving forward is to understand the extent to which, if any, de novo CNVs are enriched in TS compared to controls. This is something we were unable to currently address in our case-only dataset.

CHR	START	STOP	CNV	NUM_TARG	PROBAND	GENE(S)	PAR_WITH	CHILD_WITH	SEG_DUP
1	148016407	148025833	DEL	6	86708153	NBPF14	14	8	1.00
1	154919891	154920844	DEL	4	86707981	PBXIP1	1	0	0.00
1	202464393	202464854	DUP	2	TDT08-124-03	PPP1R12B	0	0	1.00
2	89156874	89513382	DEL	36	80247510	VDJ	8	8	0.78
2	89185088	89309954	DUP	14	1061704238	VDJ	4	3	0.60
2	89278455	89345782	DUP	8	86705341	VDJ	4	3	1.00
2	108479105	108496613	DUP	5	TDT10-043-03	RGPD4	0	1	1.00
2	179296822	179315140	DUP	7	86708604	PRKRA	16	3	0.00
2	228012077	228012201	DEL	2	86708150	COL4A4	0	0	0.00
3	126707504	126751681	DUP	30	1061707585	PLXNA1	0	0	0.00
4	9387287	9390803	DUP	4	TDT10-058-03	RP11-1396O13	3	1	1.00
4	190876126	190981682	DUP	11	TDT10-043-03	FRG1,FRG2	1	1	1.00
5	23509967	26988450	DUP	31	TDT01-044-03	PRDM9,CDH10,CDH9	0	0	0.00
5	173035268	173040260	DUP	3	91862223	BOD1	0	0	0.00
6	29910329	29913060	DUP	7	86708014	HLA-A	21	6	1.00
6	35755642	35765067	DUP	3	1061706277	CLPS	11	4	0.00
7	99930007	99950750	DUP	6	TDT09-012-03	STAG3L5P	1	0	1.00
7	141765388	141796215	DUP	12	86708586	MGAM	2	1	0.91
7	142457334	142460873	DEL	5	1061705451	PRSS1	1	0	1.00
9	69238228	69256891	DEL	3	80248737	CBWD6	4	4	1.00
11	55029703	55060048	DUP	9	TDT10-043-03	TRIM48	0	0	1.00
11	55370915	55406774	DEL	2	86705341	OR cluster	19	10	0.00
11	71567289	71574616	DUP	2	TDT10-058-03	CTD-2313N18	0	0	1.00
12	7942163	7967140	DEL	5	TDT08-102-03	NANOG,SLC2A14	1	0	0.39
12	10583714	10588011	DEL	4	58290903	KLRC2	10	2	1.00
12	11461102	11546921	DEL	7	TDT09-033-03	PRB cluster	1	1	0.74
12	104379321	104387292	DEL	6	TDT09-004-03	TDG,GLT8D2	2	3	0.23
14	22887302	22933315	DUP	16	TDT10-058-03	VDJ	0	1	0.00
14	22907940	22935586	DUP	13	TDT10-023-03	VDJ	0	1	0.00
14	106858873	106937404	DUP	5	TDT10-023-03	VDJ	0	0	0.18
14	107170322	107179262	DUP	3	80247515	VDJ	16	8	0.15
14	107170322	107211391	DUP	7	1061708028	VDJ	15	6	0.03
15	20643843	20666597	DEL	9	80247525	HERC2P3	6	4	1.00
15	20643843	20666597	DEL	9	80247821	HERC2P3	6	4	1.00
15	25475985	25488842	DUP	8	91862223	SNORD115 cluster	10	6	0.00
15	25488761	25496087	DEL	5	93421584	SNORD115 cluster	6	3	0.00
15	30659619	30665350	DEL	4	1061705143	CHRFAM7A	3	1	1.00
16	28617140	28620178	DUP	3	58289068	SULT1A1	10	2	1.00
16	53396126	53399915	DUP	5	TDT08-116-03	RP11-44F14	0	0	1.00
17	18513357	18541288	DEL	4	1061705781	TBC1D28	0	0	0.96
19	40389650	40406062	DUP	6	86708006	FCGBP	10	2	1.00
19	40391996	40406062	DUP	5	80247761	FCGBP	10	2	1.00
19	43372213	43683311	DUP	20	1061704789	PSG cluser	9	5	1.00
19	43514150	43766292	DUP	18	77348890	PSG cluster	6	5	0.97
19	54742744	54755732	DEL	13	TDT08-069-03	LILRB cluster	1	2	0.41
19	54802481	54818836	DUP	5	1061706277	LILRA cluster	5	3	0.18
19	55943552	56196292	DUP	62	TDT10-011-03	Many	0	0	0.00
19	58003481	58003825	DUP	4	1061704789	ZNF419	1	0	1.00
22	22609435	22664766	DUP	7	TDT10-043-03	VDJ	1	0	1.00
22	22899230	23040906	DUP	9	58289068	VDJ	0	0	0.46
22	24323137	24432631	DEL	6	80247510	CABIN1, GSTT cluster	18	4	0.52
22	41368423	41368675	DEL	2	TDT08-043-01	RBX1	0	0	0.00
22	42523447	42526795	DEL	7	77349658	VDJ	7	5	1.00

Table 4-1. Putative de novo CNV calls in 164 TS trios.

De novo CNV calls detected using XHMM after filtering events with a minimum QS necessary to obtain a median transmission rate across all available trios (See methods). Calls were annotated for genic content, and presence in other samples (PAR_WITH) or other probands (CHILD_WITH), and % overlap with segmental duplications (SEG_DUP).

A

CHR	START	STOP	SIZE (KB)	CNV	TARGETS	PROBAND	GENE(S)
3	126707504	126751681	44.2	DUP	30	1061707585	PLXNA1
5	23509967	26988450	3478.4	DUP	31	TDT01-044-03	PRDM9,CDH10,CDH9
5	173035268	173040260	4.9	DUP	3	91862223	BOD1
19	55943552	56196292	25.3	DUP	62	TDT10-011-03	Multiple

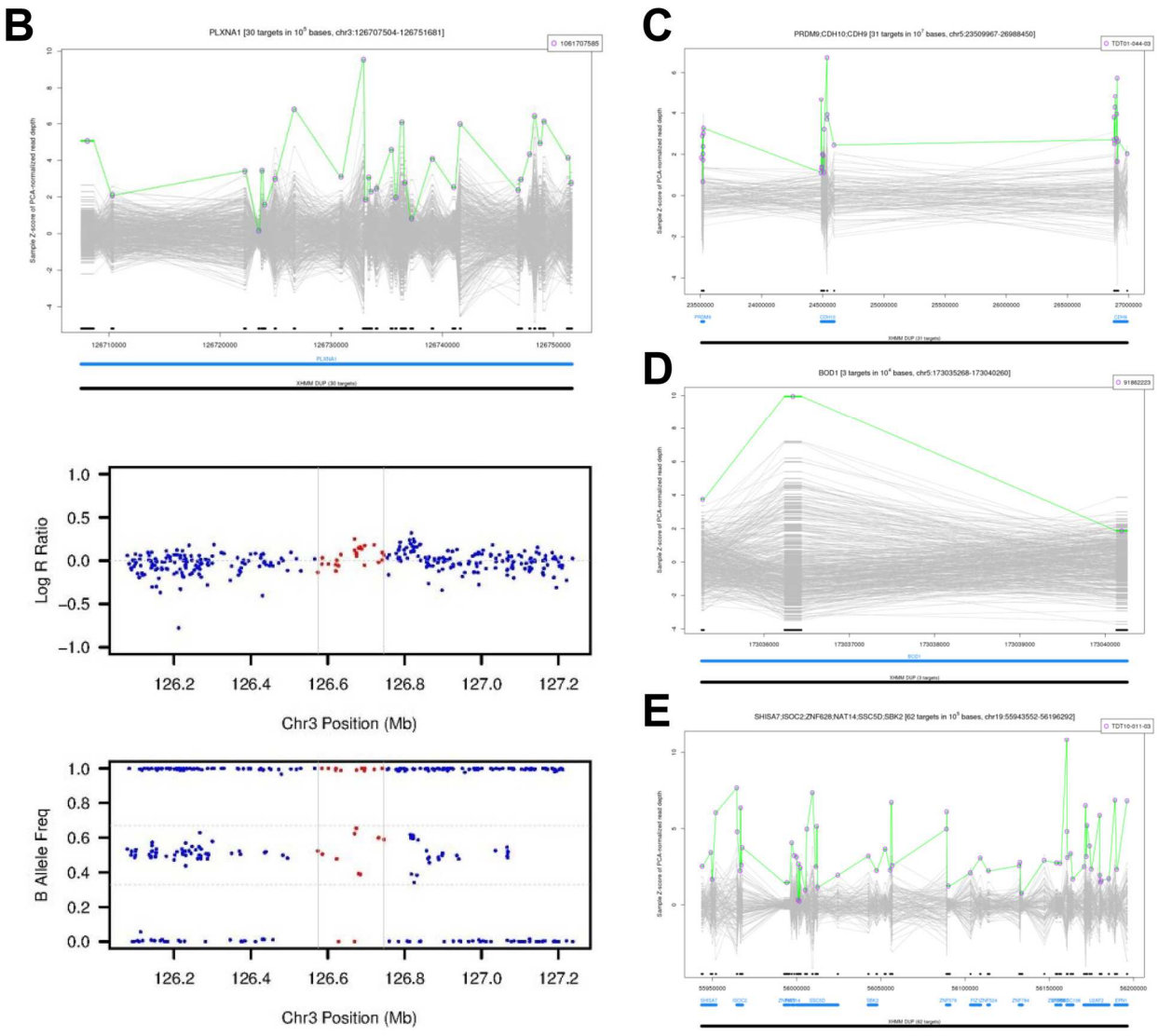


Figure 4-3. High-confidence de novo CNVs in 164 sporadic TS trios.

(A) De novo CNVs surviving post-call filtering. (B) Coverage plot for a single-gene disruptive duplication (top) in *PLXNA1* that was also confirmed by microarray analysis (bottom). (C) Locus-specific plots of normalized read depth displayed for additional putative *de novo* CNVs displayed in (A). Continuous duplication events are marked in green, with target locations indicated (pink circles). Genic content is shown below (blue).

4.6 Methods

Sample selection

A total of 170 complete parent-child trios were collected for sequencing through the TSAICG (https://www.findtsgene.org/). Each of the recruited trios consisted of an affected child who met a clinical definition for TS based on the DSM-IV-TR. Subjects and their parents were assessed for a lifetime diagnosis of TS, OCD and ADHD using a standardized and validated semi-structured direct interview (TICS Inventory) (Darrow et al., 2015). Only subjects whose parents did not meet a formal diagnosis for TS were included in this study. Further details regarding sample recruitment, ascertainment, and phenotyping are described elsewhere (Willsey et al., 2017).

Whole exome sequencing, variant calling, and quality control

For all individuals, DNA was derived from whole blood using standard protocols. Whole-exome capture and sequencing of DNA was performed for complete trios in two different ways. 139 trios were captured using the Agilent SureSelect v1.1 capture library and sequenced on the HiSeq 2500 sequencing platform (Illumina) at the Broad Institute. Another 31 trios were captured using the NimbleGen AllExon v3 and sequenced on an Illumina HiSeq 2000 at the University of California, Los Angeles (UCLA). Paired-end, 75bp sequencing was performed for all libraries at both centers. Reads were aligned to the hg19 using BWA-mem, duplicates were marked using Picard tools, and indel realignment, base score quality recalibration, and variant calling was performed on all samples simultaneously using the haplotype caller in GATK under the Broad Institute's "best practices" guidelines. Using genotypes derived from WES data, we excluded trios whose genotype calls did not reflect their described familial relationships, and samples with an excessive number of de novo variants. For failed samples, we removed the entire trio from the subsequent analysis. After genotype-based QC and removing samples with

low average coverage, we retained a total of 164 complete trios: 135 (97.1%) from the Broad samples and 29 (93.5%) from UCLA.

CNV detection using XHMM

We used XHMM for CNV detection in all trios passing genotype-based QC, largely according to the protocol outlined by the authors, with a few noted exceptions. Briefly, analysis-ready alignment files (*.bam) were first processed on all exome target intervals using GATK's "depth of coverage" module. We removed targets with extreme GC content (< 0.10 or > 0.90), as well as targets with a low-complexity fraction of > 0.25. Based on inspection of the distribution of mean read coverage across all samples, we also removed two outlier individuals with a mean coverage over all targets < 20. After inspection of scree plots, we explicitly removed the top 10 principal components for both datasets prior to normalization of read coverage. Both discovery and genotyping of CNVs was run across samples with a minimum QS threshold of 5.

Identification of rare, high-confidence de novo CNVs

We converted raw XHMM calls to PLINK format to filter out common CNVs based on a MAF > 10% using reciprocal overlap of 50%. Remaining CNVs were then input into PLINKSEQ along with pedigree information to calculate transmission frequencies at different QS cutoffs and identify putative de novo CNVs. Genotypes for all individuals (a total of 133 trios) were used to generate a titration plot of CNV transmission frequencies against QS. CNV calls were annotated for genic content, genotyping rate, overlap with regions known to cause spurious CNV calls (VDJ recombination regions and segmental duplication regions) as well as their presence or absence in other individuals within the dataset. High-confidence CNVs were defined as having a QS >60 (Broad) or >75 (UCLA) , < 0.50 overlap of their total length with segmental duplications, no overlap with VDJ recombination regions, and no additional calls observed in other parents.

For all putative de novo events, further screened CNV genotype calls in other individuals down to a QS > 5 to ensure that no additional samples harbored identical variants potentially filtered out by our strict QS thresholding.

Validation using auxiliary technology

In certain instances, we were able to confirm the presence or absence of a putative de novo CNV in proband samples, as a number of these were incorporated into previous GWAS and CNV studies and assayed using SNP microarrays (Huang et al., 2017; Scharf et al., 2013). As the reliable threshold for CNV identification is typically much larger for array-based detection, verification was performed by manual inspection of probe intensity plots over the putative CNV region. Details of the sample collection, genotyping, and array intensity processing are described in detail elsewhere (Huang et al., 2017).

Chapter 5:

Contribution of Common and Rare Variation to BP Susceptibility in Extended Pedigrees from Population Isolates

5.1 Abstract

Current evidence indicates that genetic risk for psychiatric disorders derives primarily from numerous common variants, each with a small phenotypic impact. The literature describing apparent segregation of BP in numerous multigenerational pedigrees suggests that, in such families, large-effect inherited variants might play a greater role. To evaluate this hypothesis, we conducted genetic analyses in 26 Colombian (CO) and Costa Rican (CR) pedigrees ascertained for BP1, the most severe and heritable form of BP. In these pedigrees, we performed microarray SNP genotyping of 856 individuals and high-coverage whole-genome sequencing of 454 individuals. BP1 individuals, compared to their unaffected relatives, had higher polygenic risk scores derived from independent GWAS, and displayed a higher burden of rare deleterious SNVs and rare CNVs across genes likely to be relevant to BP1. Parametric and non-parametric linkage analyses identified 15 BP1 linkage peaks, encompassing about 100 genes, although we observed no significant segregation pattern for any particular rare SNVs and CNVs. These results suggest that even in extended pedigrees, genetic risk for BP appears to derive mainly from variants of small to moderate effect.

5.2 Background

BP, consisting of episodes of mania and depression, has a heritability from twin studies estimated to be about 80% (Merikangas and Low, 2004). For BP, as for most other common disorders, SNP-based GWAS of large case/control samples have discovered many loci that contribute unequivocally to disease risk but that collectively explain only a small fraction of disease heritability; the most recent published BP GWAS, incorporating more than 20,000 cases and 30,000 controls has reported 30 genome-wide significant SNP-associations (Stahl et al., 2017). The hypothesis that rare SNVs and rare CNVs could explain a substantial proportion of this "missing heritability" (Manolio et al., 2009) has motivated the rapid growth of WES and WGS throughout biomedicine, including psychiatry. More than for other psychiatric disorders, however, sequencing efforts to identify variants with a high impact on BP risk have continued to focus on pedigrees (Cruceanu et al., 2013; Georgi et al., 2014; Goes et al., 2016; Ross et al., 2016). This focus reflects published descriptions, over several decades, of numerous extended families in which BP is observed across multiple generations; as would be expected if these pedigrees were segregating a relatively high-penetrance susceptibility variant.

Because the literature regarding the apparent segregation of BP in extended pedigrees is mostly anecdotal, we aimed to systematically characterize the genetic contribution to BP disease risk in a series of such families through evaluation of variants across the allele frequency spectrum. If rare variants contribute to this risk it is expected that they would be enriched in this sample which is, to our knowledge, the largest BP pedigree sample sequenced to date. Additionally, because a wide range of evidence indicates considerable etiological heterogeneity between BP1 and milder forms of BP, this study focused exclusively on families ascertained for multiple cases of BP1, a strategy which we reasoned would reduce the impact of such heterogeneity. In a further effort to reduce heterogeneity, we limited the dataset to

pedigrees derived from two Latin American populations that are considered closely related genetic isolates; the province of Antioquia in CO and the Central Valley of CR.

We collected microarray SNP data for 856 family members (as reported previously) (Pagani et al., 2016), and performed high-coverage WGS on 454 individuals, selected because their position in the pedigrees indicated that they would provide substantial information for imputing rare variation in the remaining family members. We analyzed these data to obtain high-quality genotypes for several categories of genetic variation, including SNVs, CNVs, and short tandem repeats (STRs). With this rich genetic information, we sought to evaluate the impact of both common and rare variants on BP1, focusing on two major questions about the genetic etiology of BP1. The first question relates to the overall genetic architecture of BP1 in these families. We attempted to address this question by characterizing the overall contribution of genetic variation to BP1 using both common and rare variants in a genome-wide burden analysis. For common variants, we calculated the polygenic risk scores (PRS) using the latest BP1 GWAS summary statistics (Stahl et al., 2017) and compared the polygenic burden of risk alleles in affected cases and related controls. For rare variants, we contrasted the burden of rare deleterious variants in a set of genes related to BP1 between affected cases and related controls, where deleterious variants are defined as stop-gain, stop-loss, splice-site, or missense variants predicted to be damaging by Polyphen-2 (Adzhubei et al., 2010). The second question we focused on is the identification of specific genetic loci segregating within families with BP. To address this question, we performed linkage analysis to test segregation of common variants and employed a new method to test segregation of rare deleterious variants in families.

5.3 Genotyping and sequencing of CO/CR BP1 families

Our study included 26 large extended families (11 CO and 15 CR) whose pedigree size varied from 12 to 355 subjects, with a range of 2 to 44 known BP1 individuals per pedigree

(Table 5-1). Methods used to assign diagnoses of BP1 in these pedigrees have been described previously (Fears et al., 2014) and are outlined briefly in Chapter 5.9, along with the criteria used to designate relatives as controls. We have performed SNP genotyping on 856 individuals, and after QC, we had 838 individuals including 206 BP1 subjects genotyped at 2,026,257 SNPs (Figure B-1). We used these genotype data for BP1 linkage analysis, estimation of inheritance vectors (IVs) imputation of WGS data, and for calling CNVs.

We performed high-coverage (36x) WGS on 454 individuals in 22 families (10 CO and 12 CR, Figure B-1). We called SNVs using the best practice pipeline of GATK HaplotypeCaller (DePristo et al., 2011), followed by variant-level QC using a classifier that employs sequencing quality information to filter out poorly sequenced variants, resulting in 20,396,290 autosomal SNVs. Genotype calls were further refined with Polymutt (Li et al., 2012), a pedigree-aware variant calling algorithm that corrected 99.83% of Mendelian inconsistent genotypes by changing mostly low quality genotype calls. We verified the ethnicity of founders using the principal component analysis (PCA) with samples from the 1000 Genomes Project (Genomes Project et al., 2015) (Figure B-2.A). After removal of WGS data from five individuals that failed QC (Figure B-1), we used the WGS data from the remaining 449 directly sequenced individuals to impute genome wide genotypes in 334 individuals for whom we had only microarray data. We performed pedigree-aware imputation using the GIGI software (Cheung et al., 2013), and measured imputation accuracy by comparing genotype concordance between microarray and imputed data for each individual. We found that, except for one individual, who we later excluded, the individuals whose genotypes we imputed displayed a genotype concordance rate > 97.93% between the microarray data and the GIGI imputation data. After imputation and QC,

Family ID	Ancestry	Pedigree Size	Males	Genotyped Individuals	Phenotyped Individuals	BP1 Individuals
CO10	CO	38	13	24	24	6
CO13	CO	24	10	20	19	5
CO14	CO	29	16	23	22	8
CO15	CO	27	12	21	21	5
CO18	CO	37	18	26	25	6
CO23	CO	48	20	32	31	9
CO25	CO	15	7	13	12	4
CO27	CO	58	31	35	35	9
CO4	CO	73	37	42	43	10
CO7	CO	149	72	111	112	29
CO8	CO	16	6	7	8	5
CR001	CR	46	25	20	7	8
CR004	CR	187	91	77	45	23
CR006	CR	37	22	13	8	4
CR007	CR	12	7	9	6	2
CR008	CR	30	15	17	13	7
CR009	CR	44	17	32	34	9
CR010	CR	30	15	17	12	4
CR011	CR	16	6	13	12	3
CR012	CR	35	16	26	22	5
CR013	CR	39	15	10	8	4
CR014	CR	26	14	8	3	5
CR015	CR	19	7	10	10	2
CR016	CR	24	14	18	19	4
CR201	CR	355	176	201	177	44
CR277	CR	25	11	13	10	4

Table 5-1. Description of families used in the current study.

Ancestry: the population isolate of each family; Pedigree size: the number of members within each family; Males: number of males in the family; Genotyped individuals: the number of family members with SNP genotype data; Phenotyped individuals: number of family members with endophenotype data; BP1 individuals: the number of individuals within each family with a definitive BP1 clinical diagnosis.

our WGS dataset consisted of 782 individuals including 190 BP1, 130 unaffected individuals, which we considered as controls, and 462 individuals with unknown disease status (Figure B-1). We used the BP1 and control data to perform burden analyses of common and rare variants and segregation analyses of rare variants.

We also performed genome-wide detection of CNVs using both microarray and WGS data. We applied a previously established pipeline for microarray-based detection (Chapter 2.2). Of 838 individuals passing SNP QC, 782 (93.3%, 189 BP1 subjects, 128 controls, and 465 unknown) also passed QC based on intensity metrics and overall CNV load. We detected a total of 5,437 CNVs (3,317 deletions and 2,120 duplications) after filtering for rare events > 5kb in length and spanned by a minimum of 10 probes. To detect CNVs from the WGS data, we used GenomeSTRiP software (Handsaker et al., 2011; Handsaker et al., 2015) to discover and genotype only bi-allelic deletions, as it has been demonstrated that such variants are far more amenable to imputation than other structural variant classes (Handsaker et al., 2015). After performing QC and merging of redundant sites, we genotyped 8,768 distinct deletion loci across the 449 WGS individuals passing both SNV and CNV-based QC, and used these calls to impute CNVs in the same set of individuals imputed for SNVs.

5.4 Characteristics of admixture in the CO/CR Extended Pedigrees

Individuals in our CO/CR pedigrees were from two Latin American populations that mainly consist of African, European, and Native American ancestries (Adhikari et al., 2016). We estimated genome-wide ancestry proportions in members of the CR and CO pedigrees using ADMIXTURE v1.22 (Alexander et al., 2009). We found that a majority of admixture proportions derived from the European ancestry followed by the Native American ancestry (Figure B-2.B). We also discovered that the admixture proportions were related to BP1 status in our pedigrees; risk of BP1 increased with an increasing proportion of European ancestry with log odds ratio (LOR) of 4.28 ($P=8x10^{-4}$) while we observed the opposite trend for Native American ancestry with LOR of -4.03 ($P=9.6x10^{-3}$) and African ancestry with LOR of -4.94 (P=0.026).

5.5 The effect of common variation on BP1 in CO/CR pedigrees

To determine the effect of common SNPs on BP1 in the CO/CR pedigrees, we calculated a PRS for each individual using the latest PGC GWAS summary statistics for BP1, consisting of 14,583 cases and 30,424 controls (Stahl et al., 2017). A higher PRS indicates a greater polygenic burden of risk alleles for BP1. We calculated the PRS at different GWAS p-value thresholds, where higher p-value thresholds used more common variants in the PRS calculation than did lower p-value thresholds. Results show that the mean PRS is higher in the 190 BP1 subjects compared to 130 controls, at GWAS p-value thresholds of 0.01 and 0.001 using linear mixed model (LMM) (P=0.001 and 0.007, respectively) and at a GWAS p-value threshold of 0.01 using generalized linear mixed model (GLMM) (P=0.003, Figure B-3). We also calculated Nagelkerke's R^2 from logistic regression and found that these PRS explain 1.5% of the phenotypic variance. This variance is noticeably smaller than that explained by PRS in the latest PGC BP GWAS data where the weighted average Nagelkerke's R^2 is 8%. We considered the possibility that this difference in the variance explained by the PRS might reflect population-level differences between the mostly European-descended PGC samples, and the Latin America pedigrees in our study, which had a substantial component of Native American and African ancestry. However, as more than 90% of the SNPs in the PGC BP GWAS were present in the CO/CR pedigrees, we consider it unlikely that this explanation alone explains the difference between the pedigree and population samples.

Although evidence over several decades delineated the distinctions between BP and SCZ, more recent studies have highlighted genetic overlaps between these syndromes (Cross-Disorder Group of the Psychiatric Genomics, 2013; International Schizophrenia et al., 2009; Stahl et al., 2017), which share symptoms in common. Notably, in GWAS data from large BP case/control samples, the PRS estimated from the PGC's SCZ GWAS results explains up to 2.5% of BP variance (Cross-Disorder Group of the Psychiatric Genomics, 2013). We contrasted the mean PRS from SCZ GWAS in BP1 affected subjects and related controls in the 22

pedigrees in our study; to calculate the PRS we used the latest available SCZ GWAS summary statistics for schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics, 2014), derived from 36,989 cases and 113,075 controls. In the CO/CR pedigrees, these PRS were not associated with an increased risk of BP1 at any of the GWAS p-value thresholds examined. The association of SCZ PRS with BP1 in the PGC but not in the CO/CR pedigrees may suggest differences in the characteristics of BP1 between these samples; in particular, this contrast may reflect the fact that we ascertained each of the pedigrees for multiple closely related cases of BP1.

5.6 The effect of rare variation on BP1 in CO/CR pedigrees

Identification of a gene-set implicated in BP1

To increase the statistical power to detect the effect of rare variants on BP1, we sought to limit the number of such variants to be evaluated by focusing on genes for which *a priori* information suggested their relevance to the phenotype. To identify such genes, we utilized three sources of information: (1) functional annotations of genome regions that have demonstrated increased heritability for BP1 in PGC GWAS data; (2) a list of the genes in the vicinity of genome-wide significant BP1 associations in the most recent PGC GWAS; and (3) a list of the genes sited under linkage peaks for BP1 in our analyses of these pedigrees. For each source, we identified genes that contain at least one putatively deleterious SNV in our dataset.

First, we evaluated the latest PGC BP1 GWAS summary statistics (Stahl et al., 2017) to identify cell-type specific promoter or enhancer regions in which BP1 heritability is enriched compared to the genome overall, using the methods and cell-type groups employed by Finucane et al. in a study of an earlier PGC GWAS with a broader definition of BP (Finucane et al., 2015). Consistent with their results, among the 10 cell-type groups that they had identified, we observed significant enrichment of heritability for BP1 only in the central nervous system

(CNS) cell type group (Figure 5-1A). The CNS cell type region comprised 457.5Mb of the genome, and contained 8,714 genes with at least one deleterious SNV in our dataset. Second, from clumping the PGC BP1 GWAS results using LD evaluated in founders of the CO/CR pedigrees, we identified a total of 15 genome-wide significant lead SNPs, with 72 genes in the 6.35Mb regions around these SNPs. Finally, in evaluating the regions defined by peaks from the BP1 linkage analysis, which we discuss in more detail below, we identified 99 genes sited within 1Mb of these peaks (Figure 5-1B). Ignoring overlap of genes appearing in multiple sources, the three sources yielded a gene-set of 8,757 unique genes for our rare variant analyses of BP1.

Burden of rare deleterious SNVs and CNVs in the gene-set for BP1

We defined rare SNVs using both an external source of allele frequency (Colombians in the 1000 Genomes database (Genomes Project et al., 2015) and Latinos in the ExAC database (Lek et al., 2016), and also allele frequency observed in the CO/CR families. We identified 25,072 rare deleterious SNVs in 8,236 of the 8,757 genes in our gene-set. For each individual, we computed the mean burden of these rare deleterious SNVs, then compared these means between BP1 individuals and related controls, while taking into account the proportion of European ancestry in each individual, and the mean burden of all rare SNVs in the gene-set. The latter procedure is necessary to account for individual differences in total amount of variation; an individual is likely to carry more rare deleterious SNVs if she/he carries more rare SNVs overall. We found that the mean burden of the rare deleterious SNVs was higher in BP1 individuals than in controls (P=0.047 using LMM). The risk of BP1, as indicated by the OR, increased by 1.26 for every one unit increase in quantile-normal transformed residual mean burden (P=0.048 using GLMM).

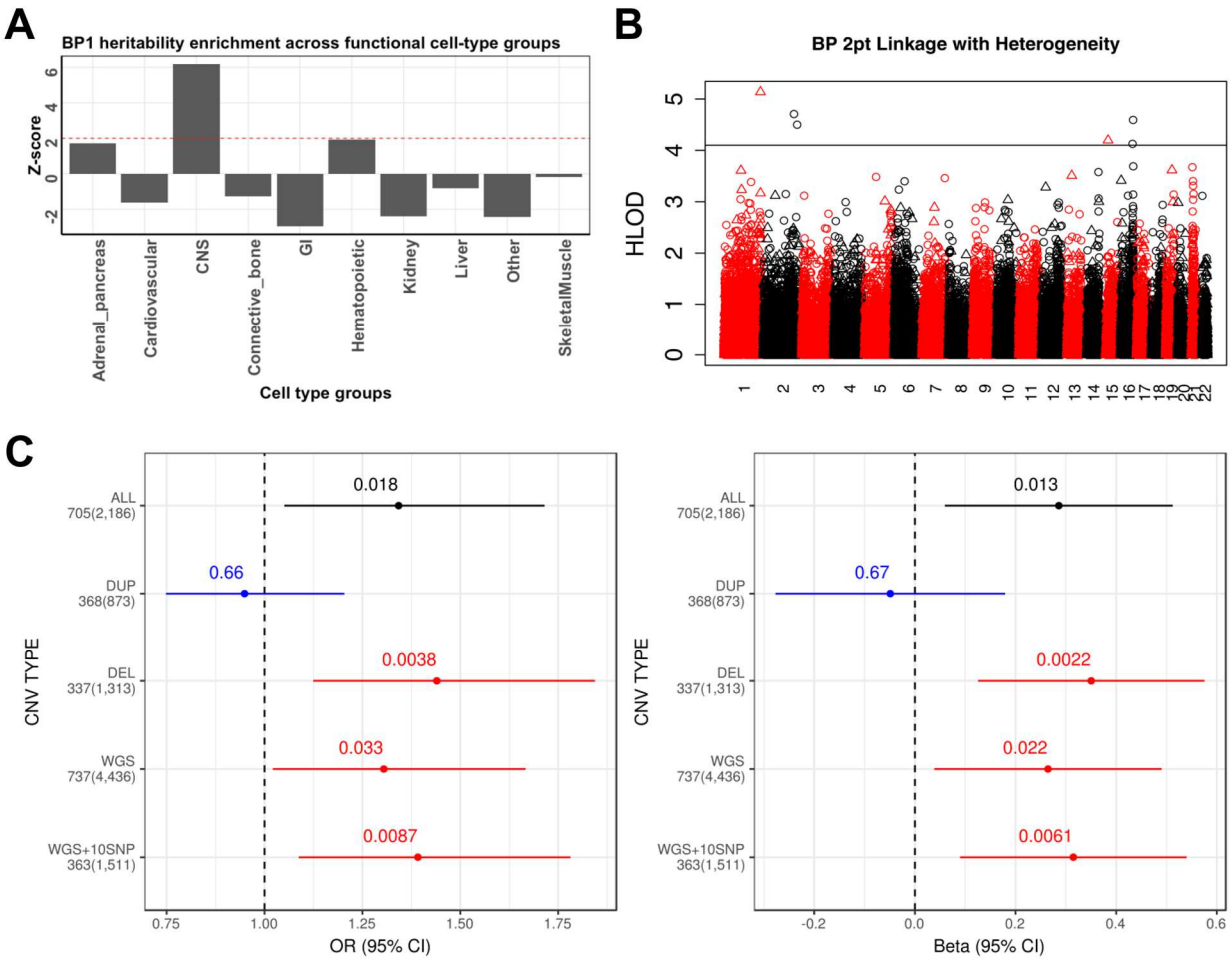


Figure 5-1. Burden analysis of rare CNVs across a BP1 gene set.

(A) We performed LDSC across different cell-type specific annotations to identify regions of enriched heritability among BP1. CNS-specific cell type annotation exhibited the greatest enrichment. (B) Shared regions by linkage analysis across BP1 pedigrees. Genes identified from (A) and (B) were combined with genes implicated in GWAS to construct a set 8,236 genes relevant to BP1. (C) Forest plots depicting the increased genic burden for BP1 subjects of rare CNVs affecting BP1 related genes, stratified by a detection method (array or WGS) and a CNV type. The number of CNVs affecting BP1 genes and total number of rare CNVs detected is displayed for each category. Enrichment was assessed using a GLMM (left) and a LMM (right) correcting for relatedness and other potential confounders. See Chapter 5.9 for details.

We also performed a burden analysis using rare CNVs. For CNVs called using microarray data, we restricted our analysis to variants greater than 5kb in length and spanned

by a minimum of 10 probes, and used frequency information from the Database of Genomic Variants (Zarrei et al., 2015) to define a set of rare variants. We identified 2,186 rare CNVs among 189 BP1 subjects and 128 related controls, and observed that the average number of such CNVs genome-wide did not differ between BP1 individuals and controls (6.86 and 6.95, respectively; P=0.67 with OR of 0.99 using GLMM). For each sample, we measured the CNV burden by enumeration of the gene count, defined as the number of genes in our BP1 gene-set intersected by rare CNVs carried by that individual. We measured enrichment of the CNV burden in BP1 individuals, under both LMM and GLMM, accounting for ancestry, relatedness, and factors known to affect global measures of genic CNV burden, including both the total number of CNVs and average CNV length (Raychaudhuri et al., 2010), as well as genotyping batch. The BP1 individuals had a higher mean burden of genes in our BP1 gene-set affected by rare CNVs than controls (P=0.013 using LMM), and this higher burden increased the risk of BP1 with OR of 1.34 (P=0.018 using GLMM, Figure 5-1.C). Stratifying our analysis by CNV type, we found that this increased burden was attributable exclusively to deletions ($P=2.2x10^{-3}$ using LMM and $P=3.8x10^{-3}$ with OR of 1.44 using GLMM). We repeated this analysis for deletions generated by imputation of CNV calls detected from WGS, filtering our rare deletions based on structural variant calls from Phase 3 of the 1000 Genomes Project (Sudmant et al., 2015). Rare deletions detected from WGS data (n=4,436 in 320 samples consisting of 190 BP1 subjects and 130 controls) were far more numerous than those detected from a subset of this sample used in the microarray analysis (n=2,186 in 317 samples consisting of 189 BP1 subjects and 128 controls). Similarly, although there was no difference in the overall genome-wide rate of rare deletions between cases and controls (13.71 and 14.08, respectively; P=0.45 with OR of 0.98 using GLMM), we observed an increased burden of genes in the BP1 gene-set affected by rare CNVs from WGS for BP1 subjects (P=0.022 using LMM), which increased the risk of BP1 with an OR of 1.3 (P=0.033 using GLMM). This burden was greater ($P=6.1x10^{-3}$ using LMM and $P=8.7x10^{-3}$ with OR of 1.39 using GLMM) when restricting our analysis to the subset of CNVs

covered by a minimum of 10 SNPs on the Omni2.5 array (n=1,511), thus demonstrating a consistent increase in gene count burden using different methods of detection.

5.7 Genetic Loci Segregating with BP1 in the CO/CR Extended Pedigrees

Linkage analysis using microarray and STR data

To identify loci that segregate with BP1 in the CO/CR pedigrees, we performed linkage analysis using both common SNPs from microarray data (209 BP1 subjects in 26 families) and STRs from WGS data (143 BP1 in 22 families). For these analyses, we considered individuals diagnosed with BP1 as affected, and designated the phenotype of all other individuals as unknown. For the linkage analysis with common SNPs, we used 99,446 independent bi-allelic SNPs with MAF > 0.35, while for the analysis with STRs, we used 85,813 bi- and multi-allelic STRs. We applied two-point parametric linkage analysis, using the Mendel software (Lange et al., 2013), to both SNP and STR data and estimated the HLOD (LOD with heterogeneity) scores. While there is not a single hard and fast approach for assessing significance in this setting, we report here the locations of HLODs greater than 4.1, which correspond to the LOD threshold for genome-wide significant linkage of 3.3 suggested by Lander and Kruglyak (Lander and Kruglyak, 1995). We also performed non-parametric linkage (NPL) analysis on these data using the Rapid software package (Abney, 2008), and estimated p-values via simulation.

Four microarray, SNPs displayed HLODs exceeding 4.1 (Figure 5-1B); two SNPs on 2q33 and 2q36 (kgp5681631 at 206.8 Mb and kgp10402924 at 228.3 Mb, with HLODs of 4.71 and 4.50, respectively) and two SNPs on 16q23 (kgp2684721 at 79.5 Mb and rs8058020 at 83.7 Mb, with HLODs of 4.13 and 4.59, respectively). There was no heterogeneity for the linkage results on 2q33 and 2q36, and on 16q23 the estimate of the proportion of linked families (denoted as a) was 0.90. The HLOD for two STRs exceeded the threshold of 4.1, one on 1q44 (247.1 Mb, HLOD=5.03, a = 0.72), and one on 15q14 (35.3 Mb, HLOD=4.32, a = 0.62). While

these results at five loci are suggestive, our simulation studies indicate that the evidence in favor of linkage is not as strong as it might appear. We carried out simulations under the null hypothesis of no linkage to BP1 anywhere in the genome using two different strategies. In one case, we kept fixed the phenotype information and used gene dropping to generate genotypes for markers with allele frequencies corresponding to the one observed in our sample; in the other case, we kept the observed genotypes and permuted the phenotype values (with appropriate consideration of family structure). Considering all sets of simulations, HLODs of 4.1 or greater were seen in 80% of replicate genome screens, suggesting that, in these pedigrees, 4.1 is likely not an appropriate threshold for declaring genome-wide significant linkage. In the NPL analysis, we approximated the empiric p-values using Rapid, and 13 SNPs and 156 STRs resulted in –log10(P) > 3.5. These SNPs and STRs were re-analyzed with 250,000 simulations to obtain a true empirical p-value; among those, eight SNPs and one STR had p<4.9e-05, which corresponds to the HLOD threshold > 4.1. We re-analyzed those eight SNPs and one STR with 1,000,000 simulations, and empiric p-values from this round of simulations ranged from 2e-06 to 3.9e-05, including two SNPs with $P=2x10^{-6}$, one on 16q23 (at 77.7 Mb) and one SNP on 21q21 (at 23.9 Mb).

Segregation of rare deleterious SNVs and CNVs within the gene set for BP1

To detect the segregation of rare variants with BP1 in the CO/CR pedigrees, we developed a statistical approach that quantifies the significance of the observed segregation pattern. Intuitively, it estimates the probability that we would observe the given segregation pattern of a rare variant or more extreme patterns under the null hypothesis of a random segregation; we refer to this as the segregation p-value. Our method used the imputation results to attempt to detect, with high confidence, founders who introduced a rare variant into a family. Given these founders, our method performed a gene dropping approach with random inheritance vectors (IVs) to find the proportion of IVs that generate the same or more extreme

segregation pattern than the observed pattern, which becomes a p-value. For SNVs, we used the same set of 25,072 rare deleterious variants in our gene-set used for the burden analysis, while for CNVs, we used 922 rare bi-allelic deletions called from WGS data in 685 genes present in the gene-set after QC. We filtered variants not shared between BP1 individuals and were able to identify founders who introduced rare variants into the family with high confidence for 86.1% and 80.0% of all rare SNVs and CNVs, respectively. We kept all of these variants for segregation analysis. For 25.4% of SNVs and 35.2% of CNVs analyzed in the segregation analysis, more than one founder introduced rare alleles into a family. In total, we analyzed segregation for 6,421 rare SNVs in 4,050 genes and 314 rare CNVs in 251 genes.

No segregation p-value for either SNVs or CNVs was significant after correction for multiple testing (Tables 5-2 and 5-3). The top gene in the SNV segregation analysis was *ACTR1B* ($P=5.18x10^{-4}$). *ACTR1B* had one rare variant (rs141238033, chr2:98275876) present in two families (CR004 and CR201). In family CR004, the rare variant was introduced by a founder whose BP status was unknown, as was that of his children. Hence, for this variant we only analyzed segregation in family CR201. This variant was predicted by Polyphen-2 to be a damaging missense variant, and it did not appear in the Colombian samples within the 1000 Genomes data (MAF of 0%) and it was very rare in the Latino samples of ExAC (MAF of 0.04%). It was enriched in the CO/CR pedigrees, with a MAF in all 449 sequenced individuals, not accounting for relatedness of 1.89% (47x over ExAC) and in family CR201 with 111 sequenced individuals of 6.76% (169x over ExAC). Accounting for relatedness (Boehnke, 1991) its MAF was 0.44% in CO/CR families (11x over ExAC) and 0.56% in family CR201 (14x over ExAC). *GOLPH3* was the top gene in the CNV segregation results ($P=7.89x10^{-4}$), with a single

Segregation Test	Location	Gene	Families	Rare Variants	P-value
Gene-level	2:98271420	ACTR1B	CR201	1	0.000518
	2:205409707	PARD3B	CO23	1	0.001046
	6:24803787	FAM65B	CO23	1	0.001338
	12:49418435	MLL2	CO7,CR004,CR012,CR016	4	0.001504
	12:53876078	MAP3K12	CR001,CR004	1	0.001513
	4:128553137	INTU	CR004,CR201,CR277	1	0.001716
	11:118063479	AMICA1	CR006	1	0.001958
	17:36921648	PIP4K2B	CO23,CR201	1	0.0022
	1:223281759	TLR5	CO7,CR010	2	0.002477
	6:117072450	FAM162B	CO14,CO7,CR001	1	0.00268
Variant-level	2:98275876	ACTR1B	CR201	1	0.000518
	1:228482059	OBSCN	CO7	1	0.000818
	6:90371215	MDN1	CO23,CR004	1	0.000975
	2:196681435	DNAH7	CO23	1	0.001046
	2:205986458	PARD3B	CO23	1	0.001046
	12:49434801	MLL2	CR004	1	0.001241
	6:24848287	FAM65B	CO23	1	0.001338
	12:53876573	MAP3K12	CR001,CR004	1	0.001513
	4:128626819	INTU	CR004,CR201,CR277	1	0.001716
	11:118067537	AMICA1	CR006	1	0.001958
Family-level	6:117083172	FAM162B	CO7	1	0.000377
	2:98275876	ACTR1B	CR201	1	0.000518
	1:228482059	OBSCN	CO7	1	0.000818
	3:140401440	TRIM42	CR012	1	0.000972
	2:196681435	DNAH7	CO23	1	0.001046
	2:205986458	PARD3B	CO23	1	0.001046
	12:49434801	MLL2	CR004	1	0.001241
	12:53876573	MAP3K12	CR004	1	0.001241
	6:24848287	FAM65B	CO23	1	0.001338
	17:36935737	PIP4K2B	CR201	1	0.001602

Table 5-2. Results from the segregation analysis of rare, deleterious SNVs.

Segregation test: the type of segregation test performed (see Chapter 5.9); Location: the hg19 genomic location of the variant; Gene: genes affected by the rare variant; Families, families that carry the rare variant(s); Rare variants: the number of rare variants used for each test.

rare CNV (DEL_P0095_217, chr5:32161816-32162478) appearing only in family CO27. Its MAF among all sequenced individuals in the CO/CR pedigrees, not accounting for relatedness, was 0.7795% and 14% in family CO27 with 25 sequenced individuals; accounting for relatedness its MAF was 0.218% in all families and 2.94% in family CO27. Neither of the above two rare variants segregated perfectly with BP1 status in the pedigrees in which they were present. In

GOLPH3, we also observed a rare damaging SNV (rs192633198, chr5:32126511) in a different family (CO7) from the family carrying the rare CNV (CO27). The segregation p-value of this SNV in family CO7 was 0.12, but it segregated perfectly, having been inherited from a founder whose three offspring included two BP1 individuals carrying the rare allele and one unaffected individual who did not carry it.

Segregation Test	Location	Gene	Families	Rare Variants	P-value
Gene-level	5:32161816	GOLPH3	CO27	1	0.000789
	21:40594759	BRWD1	CO7	1	0.008174
	7:1495145	MICALL2	CO23,CR201	1	0.014247
	18:72672166	ZNF407	CR004	1	0.014854
	3:147155005	ZIC1	CO15,CO7	1	0.025583
	3:183433723	YEATS2	CR010	1	0.031282
	10:18670518	CACNB2	CO18	1	0.031311
	1:1223477	SCNN1D	CO4	1	0.034854
	3:52614424	PBRM1	CO14,CO18,CR008	2	0.040371
	22:43974737	EFCAB6	CO23,CO7	2	0.040386
Variant-level	5:32161816	GOLPH3	CO27	1	0.000789
	21:40594759	BRWD1	CO7	1	0.008174
	22:43974737	EFCAB6	CO23	1	0.012922
	7:1495145	MICALL2	CO23,CR201	1	0.014247
	18:72672166	ZNF407	CR004	1	0.014854
	3:147155005	ZIC1	CO15,CO7	1	0.025583
	3:183433723	YEATS2	CR010	1	0.031282
	10:18670518	CACNB2	CO18	1	0.031311
	12:99818192	ANKS1B	CO27	1	0.032181
	1:1223649	SCNN1D	CO4	1	0.034854
Family-level	5:32161816	GOLPH3	CO27	1	0.000789
	21:40594759	BRWD1	CO7	1	0.008174
	22:43974737	EFCAB6	CO23	1	0.012922
	18:72672166	ZNF407	CR004	1	0.014854
	3:147155005	ZIC1	CO15	1	0.015621
	6:16259314	GMPR	CR004	1	0.031155
	6:159419156	RSPH3	CO7	1	0.031242
	3:183433723	YEATS2	CR010	1	0.031282
	3:52669549	PBRM1	CO18	1	0.031311
	3:52669566	PBRM1	CO18	1	0.031311

Table 5-3. Results from the segregation analysis of rare CNVs.

5.8 Discussion

The studies described here enabled us to document the contribution of both common and rare variants to BP1 risk, in the CO/CR pedigrees that we genotyped and sequenced. Our finding of elevated BP1 PRS scores, in BP1 affected individuals compared to controls, indicates that in these pedigrees, as in case/control samples, BP1 risk derives substantially from the polygenic effect of common SNPs. This result is consistent with what has been observed in extended pedigrees that apparently segregate non-psychiatric common disorders, for example familial dyslipidemias (Ripatti et al., 2016). It will, however, be important to understand why the magnitude of the polygenic contribution is so much smaller in the CO/CR pedigrees compared to the cases from the PGC. The dissimilarity between the pedigree and case/control series in terms of sample size and ethnicity offers one explanation for this divergence (Martin et al., 2017). Another possibility, however, is that BP1 in individuals from extended pedigrees is simply less polygenic than in population samples.

In support of the idea that BP1 in extended pedigrees may be etiologically distinct from this diagnosis in case/control samples, we found that common risk variants for SCZ do not appear to contribute to BP1 risk in our samples. This result contrasts with the observation in the PGC that the SCZ PRS is an equivalently strong predictor of BP1 risk as the BP PRS. This discrepancy relates to one of the most important uncertainties in our understanding of severe mental illness, the genetic relationship between BP and SCZ. The separation between BP and SCZ was for many decades a bedrock principle of psychiatric nosology, based on both the classically distinct longitudinal trajectories of these syndromes as well as evidence from genetic epidemiology studies suggesting that they do not co-segregate in families. More recently, however, several lines of evidence indicate a shared genetic architecture between SCZ and BP, particularly with respect to BP1 (Cross-Disorder Group of the Psychiatric Genomics, 2013; International Schizophrenia et al., 2009; Stahl et al., 2017). Efforts are now underway in multiple

datasets to examine the relationship between BP and SCZ at a finer-grained level than that of syndromic diagnosis, for example in terms of the presence or absence of individual symptoms or the distribution of specific quantitative traits (Sanders et al., 2017). Our results suggest that it may be particularly informative to include the rich phenotypic data available from the CO/CR pedigrees in such studies.

To date, there are no convincing reports of discoveries of high-impact BP susceptibility variants, or even loci, from either pedigree or case/control sequencing studies. Such studies have, almost certainly, all been underpowered for such discoveries. Many pedigree-based sequencing studies have, for example, described single families in which candidate variants were possibly segregating with disease (Cruceanu et al., 2013; Georgi et al., 2014; Goes et al., 2016; Ross et al., 2016). The dataset that we describe here is also underpowered for rare variant discovery. However, the comprehensive genotype data that it contains for nearly 800 individuals from 22 families, provide an opportunity for more complete evaluation than has previously been possible of the contribution of rare variants to BP1 within pedigrees, and for a more rigorous examination of the segregation of rare variants with the disorder.

In assessing the contribution of rare variants to BP1, we found, in a set of 8,757 genes selected based on hypothesized relevance to this disorder, a higher burden of rare variants (deleterious SNVs as well as CNVs) in affected individuals compared to related controls. We chose this set of genes because they corresponded to regions where BP1 heritability was enriched and where BP1 GWAS hits and our linkage peaks resided. Ament et al. used a gene-set related to neuronal excitability and similarly observed enrichment of rare risk variants for BP among those genes (Ament et al., 2015). Previous studies have reported no evidence for a global enrichment of rare CNVs in BP individuals (Green et al., 2016; Grozeva et al., 2010). In contrast to these studies, however, which were restricted to analyses of CNVs > 100kb, our use of higher density arrays and WGS allowed us to examine much smaller events. Our results

suggest that the impact of CNVs on BP burden may be restricted to those that affect specific sets of genes, or possibly derive mainly by smaller events < 100kb.

We applied, to this dataset, a new statistical approach that we developed to calculate a p-value for segregation of rare variants. We did so because existing methods used for this purpose (Bureau et al., 2014; Qiao et al., 2017; Sul et al., 2016) are not scalable to very large pedigrees such as those in our dataset. These methods also make the simplifying assumption that only one founder has introduced a given rare variant into the pedigree. While this assumption is reasonable for small or moderate-sized families, we have observed in large pedigrees (n>70) many instances in which more than one founder has introduced a rare variant into the family. Our method relies on accurate imputation of rare variants, which is made feasible by a family imputation approach that achieves a higher call rate and accuracy for rare alleles than population-based imputation approaches (Cheung et al., 2013). This approach, however, requires reliable estimation of IVs in each family, and if only a small number of subjects in a family are genotyped at markers used for estimating IVs, imputation results might not be very accurate.

Although the method we developed worked well, in practice, we did not detect rare variants with strong evidence of segregation in the CO/CR pedigrees. The top gene that we detected through SNV segregation analysis is *ACTR1B*. It has one damaging missense rare SNV that is very rare in the Latino population of ExAC, but has a relatively high MAF in one of our pedigrees, with the rare allele occurring at about 170X increased frequency compared to reference samples. A previous CNV analysis using WTCCC2 SCZ data found an association between the large duplication around *ACTR1B* and SCZ, but it was not replicated (Morris et al., 2014). *GOLPH3* is the top gene from the CNV segregation analysis, which contains one rare and relatively small deletion appearing only in one family. A previous linkage study for SCZ found a linkage peak on the 5p13 region that included this gene, but did not support association

between *GOLPH3* and SCZ (Ritter et al., 2012). One limitation in the segregation analysis for rare variants is that while we were able to identify founders who introduced rare variants to the family for a majority of variants, for some variants we could not identify those founders with high confidence, and hence did not test them. Duplications and CNVs from microarray data did not generate an allele for each chromosome, and hence they were not tested as our method relies on imputation that requires genotype information. Lastly, we considered only deleterious coding SNVs for both burden and segregation analyses of rare variants; we did not analyze non-coding SNVs or indels because it is a challenge to measure their deleteriousness (Ionita-Laza et al., 2016; Kircher et al., 2014).

In conclusion, our study demonstrates the polygenic genetic architecture of BP1 in a well characterized and large series of extended pedigrees, reflecting the action of a combination of many common and rare variants (including both SNVs and CNVs) with small or moderate effect sizes. Additionally, unlike in BP case/control samples, common SCZ risk alleles do not contribute to BP1 risk in these families. Finally, although our new method makes it feasible to rigorously evaluate rare variant segregation in large pedigrees, our inability to identify BP1-associated coding variants suggests a highly-polygenic mode of inheritance and/or the possibility that non-coding variants may play also play a role for BP1 risk in these pedigrees.

5.9 Methods

Ascertainment and Phenotypic Assessment of Study Sample

The study sample consisted of members of 26 pedigrees, 15 from CR and 11 from CO, each ascertained based on multiple members who had previously been clinically diagnosed with BP1. Diagnostic interviews were conducted using the Mini International Neuropsychiatric Interview (M.I.N.I) (Sheehan et al., 1998) and a Spanish version of the DIGS; the CR clinical team developed this translated and validated (Palacio et al., 2004) version and made it publicly

available at https://www.nimhgenetics.org/interviews/digs_3.0_b/. Interviews were performed by bilingual psychologists or psychiatrists extensively trained in the use of these instruments. Among CR subjects newly recruited for this study, only those who answered positively to M.I.N.I questions related to mood or psychotic symptoms were targeted for the complete DIGS interview. All available CO family members received a DIGS interview. Six research psychiatrists or psychologists, at the University of California, San Francisco (VR), University of California, Los Angeles (CEB, JGF, JM), University of Antioquia (CLJ), and Rutgers University (JE), who are experts in the diagnosis of mood disorders reviewed all available clinical material for the cases (DIGS interview, medical records, hospital notes), with at least one expert rater reviewing each case. Prior to beginning Best Estimate (BE) procedures each diagnostician established reliability (diagnostic agreement of 90% or better) with the Chair of the BE group (VR). Individuals designated as BP1 had a BE diagnosis of BP1, unipolar mania, or schizoaffective disorder, bipolar type. Control subjects were those who went through the complete psychiatric evaluation and were found to have no mental illness, as well as those who answered negatively to all M.I.N.I. (Sheehan et al., 1998) questions related to mood or psychotic symptoms, and were > 60 years of age. Written informed consent was obtained from all participants. Institutional Review Boards at participating institutions approved all study procedures.

DNA Extraction and Microarray Genotyping QC

DNA was extracted from whole blood using standard protocols. Illumina Omni 2.5 chips were used for genotyping of 856 participants, in three batches. A subset of samples was repeated in each batch to enable concordance checks. A total of 2,026,257 SNPs were polymorphic and passed all QC procedures, including the evaluation of call rate, testing for Hardy Weinberg equilibrium (HWE), and Mendelian error. For linkage studies we used 99,446 SNPs with MAF>0.35, that had been LD-pruned to have r^2<0.5. During QC procedures, allele

frequency calculations, calculations of HWE, and estimates of LD were performed using only unrelated (founder) individuals. Further pedigree-wide Mendel checks were performed on the set of 99,446 SNPs used in linkage analysis; 0.7% of markers had 1 or more errors in Mendelian inheritance in these additional checks and all data for the marker in the family that generated the error was set to missing. Individual-level QC checks included verifying the pedigree structure by comparing theoretical kinship with empirical estimates, assessing missing rate, and verifying that the genetic sex agreed with the reported sex. The final array genotype data set included data on 838 subjects in 26 families.

Estimation of global admixture proportions

We generated estimates for all 838 pedigree members with SNP array genotype data. The reference populations were the CEU (n=112) and YRI (n=113) from HapMap (International HapMap, 2003; International HapMap et al., 2007), as well as 52 Native American samples from Central or South America. The Native American samples were the Chibchan-speaking subset of those used in Reich et al. (Reich et al., 2012) selected to originate from geographical regions relevant to CR/CO and to have virtually no European or African admixture. In total the admixture analysis used 57,180 LD-pruned SNPs and 1115 individuals. We compared the proportion of European ancestry between BP1 subjects and controls using both a linear mixed model (LMM) and a generalized linear mixed model (GLMM), taking into account relatedness of subjects using a kinship matrix.

Linkage Analysis

Overview: Several of the 26 pedigrees are large and complex, with multiple inbreeding or marriage loops. Such pedigrees pose a significant challenge to analysis, as standard multipoint approaches that employ the Lander-Green algorithm fail in this setting. Possible solutions would include trimming and cutting pedigrees into smaller units, however we found

that applying this approach in these pedigrees led to a reduced capability to resolve phase, and therefore resulted in substantially reduced power. We also considered MCMC approaches (Garner et al., 2001), however given the density of markers in a genome-wide setting, such approaches are exceedingly slow, and we found it difficult to evaluate if we had reached convergence. We evaluated linkage in our pedigrees using a two-point parametric linkage on both SNPs and multi-allelic STRs identified from the WGS data using LobSTR (Gymrek et al., 2012), as well as a non-parametric method proposed specifically for use in large, complex pedigrees (Abney, 2008).

Two-point Parametric Linkage Analysis: We used the software Mendel (Lange et al., 2013) to estimate allele frequencies in the 26 pedigrees, performing these estimates separately for the CR and CO samples, and accounting for pedigree relationships. Two-point parametric linkage with heterogeneity was done in Mendel for autosomes, the very large pedigree CR201 had to be broken into four non-overlapping sections in order to do this analysis. Our parametric model used disease frequency to be 0.003, and penetrance parameters: P(BP1|DD)=0.9, P(BP1|DN)=0.81 and P(BP1|NN)=0.01; where "D" represents the disease allele and "N" represents the normal allele. Individuals are considered either affected with BP1 or phenotype unknown, that is, we designate no individuals as controls. This is the same model employed in McInnes et al (McInnes et al., 1996) and Herzberg et al (Herzberg et al., 2006). The null hypothesis of no linkage was evaluated using the LOD (logarithm of odds) score, which is the log 10 of the ratio of the likelihood of linkage, given the estimated recombination fraction, to the likelihood given the null value for the recombination fraction (0.5). Traditionally, a LOD score of 3 (P=0.0001) was considered significant linkage, however Lander & Kruglyak (Lander and Kruglyak, 1995) suggested LOD=3.3 ($P=4.9x10^{-5}$) was a more appropriate threshold. The HLOD (LOD with heterogeneity), tests 2 parameters: the recombination fraction and proportion of linked pedigrees. The HLOD score has traditionally been evaluated using a combination of chi-

square distributions with 1 and 2 degrees of freedom (Abreu et al., 2002). By this metric, an HLOD of 3.6-3.7 corresponds to p~0.0001; using the Lander and Kruglyak p-value threshold of 4.9E-05 corresponds to an HLOD of 4.1.

Non-parametric Linkage Analysis: We used Rapid (Abney, 2008) for non-parametric analysis of allele-sharing of SNPs among BP1 individuals in our pedigrees. Rapid is designed for use in large complex pedigrees, where estimation of identity by descent (IBD) can be computationally demanding. Rapid evaluates significance using an approximation to the empiric p-value, which Abney et al.(Abney, 2008) states is conservative. After the first genome-wide scan to identify promising results using the approximation, we re-analyzed markers with -log10(P) > 3.5 with 250,000 simulations to obtain a true empiric p-value. Markers with p<4.9e-05 were then evaluated with 1,000,000 simulations to obtain a more accurate p-value.

Whole genome sequencing data

Subject Selection: We used ExomePicks to identify the subset of subjects to sequence that would enable maximum opportunity to impute sequence variants into the remaining genotyped individuals in the pedigrees. Selection in ExomePicks was done without regards to phenotype, when possible BP1 subjects were selected over non-BP siblings. ExomePicks selected 487 of the 838 genotyped subjects as candidates for sequencing. Sequencing was done in two rounds, and due to budget constraints, not all families were sequenced: no individuals from three CR families and one CO family were sequenced. In total 454 CR/CO subjects were sequenced, including 144 BP1 (Figure B1).

WGS and Variant Calling: Illumina performed WGS of 454 individuals with 36x overall coverage. To improve accuracy of variant calls, we re-aligned and re-called the WGS data using the GATK best practices (DePristo et al., 2011). We converted the original BAM files to FASTQ files, and used the Churchill pipeline (Kelly et al., 2015), which is an efficient and scalable

implementation of the GATK best practices pipeline. We used the HaplotypeCaller of GATK (version 3.5-0-g36282e4). First, each individual was called separately to generate a GVCF file, and all individuals were joint-called using the GenotypeGVCFs tool in GATK. The variant calling was performed in the high performance cluster at UCLA called the Hoffman2 cluster.

Individual-level QC of WGS Data: Before checking sequencing quality of each individual, we first removed variants that failed Variant Quality Score Recalibration (VQSR) in the GATK pipeline and we set each genotype whose genotype quality (GQ) score was ≤ 20 to missing. Additionally, all multi-allelic SNVs were excluded. In the remaining variants, for each sequenced subject we assessed genotype missing rate, the number of singletons and WGS genotype concordance with array genotypes, verified agreement between reported sex and sex determined from X chromosome markers, and compared empirical estimates of kinship with theoretical estimates. We also performed principal component analysis (PCA) using 1000 Genomes (1KG) Phase 3 as a reference panel (Genomes Project et al., 2015). We used EIGENSTRAT (Price et al., 2006) for PCA, and included only founders from WGS data, and used independent common SNVs present in both BP1 WGS and 1KG Phase 3.

Variant-level QC of WGS data: In doing variant-site QC, rather than using fixed thresholds for missingness, HWE, etc. to filter out poor quality variant sites, we used logistic regression to predict the probability of variant sites being of good or poor quality, with prediction based on mean and standard deviation of variant sequencing depth and genotype quality and a fraction of individuals meeting certain thresholds for depth and quality. We trained the regression model on a set of variants deemed to be of obvious good and poor quality using several QC measures such as genotype missing rate, HWE p-value, and the number of Mendelian errors. We assessed the accuracy of the model using cross validation.

Genotype refinement using a pedigree-aware variant calling algorithm: We used the pedigree-aware variant calling method Polymutt (Li et al., 2012) to refine genotypes of variant sites that passed QC. This method takes a VCF file with genotype likelihood generated from GATK and pedigree structure as input and generates a VCF file that refines genotype calls based on their likelihood and pedigree structure. To use Polymutt, some of large pedigrees had to be trimmed to remove inbreeding loops and reduce pedigree complexity. The use of Polymutt to refine genotype calls reduces the possibility of identifying *de novo* mutations, but results in a dataset with nearly zero Mendelian inheritance errors (MEs). Polymutt also imputed all sporadic missing genotypes.

Imputation of sequenced sites into genotyped subjects: The 454 individuals in 22 of the 26 CO/CR families who underwent WGS were selected with the goal of imputing the rest of 334 family members who were genotyped on the Omni 2.5 chip, but not directly sequenced (Figure 6-1). We used the imputation software GIGI (Cheung et al., 2013), an approach designed to impute genotypes on large extended pedigrees. First, we obtained a genetic map of variant sites using the Rutgers genetic map interpolator (Matise et al., 2007). We then used MORGAN (Thompson, 2011) to obtain pedigree IVs based on independent common SNP array genotype data of individuals in 22 families with WGS data. GIGI used these IVs to impute WGS variants identified in the directly sequenced individuals into the pedigree members with only SNP genotype data. GIGI also imputed sequence data for pedigree members who were neither sequenced nor genotyped in the Omni 2.5 chip. GIGI imputed each family separately, and we further divided each chromosome into 10 Mb intervals to perform efficient imputation by utilizing parallelization in the high performance cluster. After imputation, GIGI generated the probability of each imputed genotype and we used the threshold-based calling with the default threshold to call genotypes for the rare variant burden analysis while we used the most likely genotype calls for the rare variant segregation analysis. For the threshold-based calling approach, only

genotypes with probability in excess of 80% were called, genotypes that did not meet this threshold were set to missing. After the GIGI imputation, there were 782 individuals who were either sequenced or imputed with the high quality in 22 pedigrees. Among them, 190 are BP1 and 130 are controls, and they were analyzed in analyses of polygenic risk, and burden and segregation for rare SNVs and CNVs.

Identification and QC of STRs: We detected STRs with the lobSTR software (Gymrek et al., 2012) (version 4.0.0) that uses sequencing data to call STRs. The BAM files aligned with BWA-MEM that were generated during the variant calling process were used as input to lobSTR, which then generated VCF files for STR loci. To perform QC on STR calls, we developed a filtering strategy based on the violation of Mendelian inheritance as well as comparison between lobSTR calls and STR data previously detected with electrophoresis for one family. This filter removed 1) monomorphic STRs, 2) STRs with repeats of one nucleotide, 3) STRs with ambiguous repeating nucleotides, 4) STRs with call rate < 95%, coverage < 5x, or Q-score < 0.95.

Annotation of SNVs: Discovered SNVs were mapped to UCSC knownGene (Rosenbloom et al., 2015) and GENCODE V.19 (Harrow et al., 2006) transcripts, and damaging missense SNVs were predicted by PolyPhen-2 (Adzhubei et al., 2010). To identify rare SNVs, we used both external and internal sources of allele frequency. For external allele frequency information, we used allele frequency in 1000 Genomes (Genomes Project et al., 2015) (1KG) Colombians (CLM) and ExAC (Lek et al., 2016) Latino population (AMR). If a variant in our dataset is present in 1KG CLM or ExAC AMR, it is rare if its MAF is < 1% in either dataset. If a variant is not present in both 1KG CLM and ExAC AMR, it is rare if its MAF is < 10% in our dataset where MAF is estimated from all sequenced individuals. We defined deleterious SNVs as variants that are stop-gain, stop-loss, splice-site, and missense variants predicted to be damaging by PolyPhen-2.

Identification and annotation of rare CNVs

Array-based CNV detection: We adapted a previously established pipeline for array-based CNV detection (Huang et al., 2017). Briefly, after reclustering array data by genotyping batch, we generated a consensus callset based on two separate calling algorithms (Colella et al., 2007; Wang et al., 2007). CNVs were only included in if they were called by both algorithms and of the same relative type (CNV gain or loss). We removed outliers based on several intensity-based metrics (Supplementary Figure 4), merged fragmented CNV calls, removed CNVs in regions prone to false-positives, and further removed outliers based on global CNV load. Out of 838 samples passing SNP-based QC, 780 (93.1%) had intensity data suitable for CNV detection. We restricted our analysis to CNVs spanning at least 10 SNPs and > 5kb in length.

WGS-based CNV detection: We used GenomeSTRiP (Handsaker et al., 2011; Handsaker et al., 2015) to both discover and genotype deletion CNVs across all samples with WGS. We removed three individuals with an excessive number of CNV calls who also failed SNV QC (Figure 5-S3A) and two individuals with possible sample mix-ups before genotyping deletions across all remaining individuals. Chromosome level plot of CNVs detected after outlier removal exhibited typical peaks on chromsomes 2 and 14, corresponding to CNVs found in VDJ recombination regions (Figure 5-S3B). Inspection of available trios (n=67) demonstrated low levels of Mendelian inconsistencies across these candidate sites (0.0026±0.00067, Figure 5-S3C). We filtered genotyped sites for redundancy and quality, generating a final WGS deletion callset consisting of 8,768 distinct CNV loci, and estimated a genome-wide FDR of 0.018 for deletions ≥ 5kb using an intensity rank-sum test against available SNP data. Finally, we removed monomorphic CNVs, Mendelian inconsistent genotypes, and genotypes with high missing rate > 5% prior to performing imputation of CNVs using the same imputation pipeline for SNVs.

Annotation of rare CNVs: As with SNVs, we applied the similar method to classify rare CNVs in both array and WGS CNV datasets. We used two different sources of information, dependent on the platform, to exclude variants that are common in the general population. For array-based CNV calls, we extracted frequency information from Database of Genomic Variants (DGV) Gold Standard Variants (Release 2016-05-15) (Zarrei et al., 2015). Since the DGV is a consolidated resource comprised of many individual studies, we only considered frequency information for CNVs estimated from datasets with a minimum of 1000 distinct samples. For the WGS deletion call set, we used structural variant calls from Phase 3 of the 1000 Genomes Project (Sudmant et al., 2015), and extracted frequency information for all deletions in the AMR continental group. We annotated all CNVs in both the array-based and WGS datasets with frequencies from these public datasets based on a 50% reciprocal overlap, and retained CNVs with MAF < 1%. CNVs without frequency information were retained if their MAF is < 10% across all available samples in our dataset (regardless of phenotype).

Identification of genes for burden and segregation analyses for rare variants

We highlighted genes for burden and segregation analysis of rare variants from three sources: (1) genes in regions that demonstrated evidence of enrichment of BP1 heritability, as evaluated in the PGC BP1 GWAS data set (summary GWAS statistics); (2) genes near PGC BP1 GWAS peaks; (3) genes within 1Mb of our linkage peaks. We only consider genes where there is at least one deleterious SNV in our BP1 dataset.

We identified regions with enrichment of BP1 heritability using PGC GWAS summary statistics and 220 epigenetic profiles that originated from 100 individual cell types or tissues. The epigenetic annotation came from cell-type specific histone modification (ChIP-seq) data generated by the NIH Roadmap Epigenome Project (Bernstein et al., 2010). H3K4me1, H3K4me3 and H3K9ac data were post-processed by Trynka et al. (Trynka et al., 2013). Peaks

were called using MACS v.1.4. For each cell-type and specific histone mark, start and end of the peaks were determined. H3K27ac data were post-processed separately by Hnisz et al. (Hnisz et al., 2013), also using MACS v.1.4. The 220 cell type-specific annotations were then divided into 10 groups by taking a union of the cell type-specific annotation within each group, following Finucane et al (Finucane et al., 2015). For each of the 10 tissue groups, the genome was annotated with the start and end of regions identified as active promoters/enhancers. Within all regions marked for a specific tissue group, we evaluated the possible enrichment of BP1 heritability using PGC BP1 GWAS summary statistics and Stratified LD Score Regression (Finucane et al., 2015). For tissue groups that demonstrated significant BP1 heritability in the PGC, we then highlighted for further analysis in our BP1 families all genes within regions marked by the annotation process.

We also specifically targeted genes near genome-wide significant association signals in the PGC BP1 GWAS. The PGC BP1 GWAS summary statistics were clumped in PLINK, using our BP1 genotype data as the LD reference (founder genotypes only). Clumps were formed in windows of 250 Kb and using an r^2 threshold of 0.1. Among the resultant clumps, if the lead SNV was genome-wide significantly associated to BP1 in the PGC data ($p < 5e-08$), we determined the physical extent of the SNVs that were in the same clump as the lead SNP, and considered for further analysis all genes within such regions.

Lastly we targeted genes within 1 Mb of linkage peaks (both parametric and non-parametric) identified in our BP1 pedigrees. Linkage peaks were defined as parametric HLOD > 4.1, or non-parametric p-value < 4.9e-05.

Polygenic Risk Score Analysis

We calculated the polygenic risk score (PRS) in our sample using summary statistics for autosomal SNVs from the PGC GWAS of BP1, and SCZ. SCZ GWAS results were downloaded

from (https://www.med.unc.edu/pgc/results-and-downloads), and BP1 GWAS results were obtained with permission. We used PRSice (Euesden et al., 2015) with both sets of GWAS data. We excluded A/T and G/C SNVs and SNVs in the MHC region on chromosome 6. Our WGS data were LD clumped, and we retained from the GWAS summary statistics the most significant SNV for each clump. All PRS were estimated at five p-value thresholds, and we used GLMM to assess the odds of being BP1 as a function of increasing mean PRS, for each GWAS p-value threshold. We also used LMM, modeling PRS as a dependent variable and disease status as a predictor. Relationships among individuals were captured with a variance component to account for kinship (Chen et al., 2016), and global admixture proportions of European ancestry were included as a fixed effect in both LMM and GLMM. When estimating Nagelkerke R^2, we used a logistic regression model without taking into account relationships as it was not straightforward to estimate R^2 in the GLMM framework. We also did not include the global admixture proportions when calculating Nagelkerke R^2 because we were interested in variance of BP1 explained only by the PRS.

Comparison of the burden of deleterious SNVs and CNVs in BP1 and Controls

For each subject, we calculated the mean burden of rare deleterious SNVs using the PLINK --score option. This corresponds to the fraction of deleterious alternative minor alleles at those SNVs that each individual has. Missing genotypes were imputed by PLINK when computing the mean burden. We also calculated the mean burden of all rare SNVs in genes described as above including variants that are not deleterious. The mean burden of rare deleterious SNVs was regressed on the mean burden of all rare SNVs using LMM and included a kinship matrix to account for relatedness among individuals. The residuals of the mean burden of rare deleterious SNVs after LMM were then quantile-normal (QN) transformed. We compared the QN transformed residuals between BP1 subjects and controls using both LMM and GLMM,

taking into account relationships among individuals and admixture proportions of European ancestry.

To test for an increased CNV burden in BP1 subjects compared to controls, we tabulated three measures for each subject: the total number of genes within our gene-set affected by CNVs (gene count), the total number of CNVs regardless of genic content (total CNV count), and the average size of all CNVs (average CNV size). Genes were considered affected by a particular CNV based on a strict overlap with annotated gene boundaries. To examine for global differences in CNV rate, we used a GLMM to compare total CNV counts between BP1 subjects and controls, accounting for subject relatedness and genotyping batch. For the CNV burden across our BP1 gene-set, similar to the mean burden of deleterious SNVs, gene count was first regressed on total CNV count, average CNV size, and relatedness between individuals using LMM. The residuals were then QN transformed, and we used both LMM and GLMM to compare the transformed residuals between BP1 subjects and controls while accounting for relationships among individuals and global estimates of European ancestry. As gene count burden is influenced heavily by overall CNV detection rate (Raychaudhuri et al., 2010), including total CNV count, average CNV size, and genotyping batch as covariates within the model is important to control for any case-control differences that may arise from variation in assay quality. Burden analysis for CNVs from WGS data was conducted in the same manner (without genotyping batch as a covariate). Although GenomeSTRiP genotypes individuals across all CNV loci with high efficiency, we note that inclusion of the total CNV count corrects for varying rates of imputation efficiency across samples.

Segregation Analysis for rare SNVs and CNV

Given a rare variant in a family, we developed a statistical approach that computes a p-value to estimate the probability of having the observed segregation pattern or more an extreme

segregation pattern. Our segregation statistic (S_{rare}) is the sum of the number of affected (BP1) individuals with a rare variant and the number of unaffected (control) individuals without a rare variant. We assume that we know founders who introduced the rare variant into a family, and we denote these founders as F^{rv}. We will discuss later how we can identify F^{rv}. Given a S_{rare} statistic for a certain segregation pattern and F^{rv}, we can compute a p-value of S_{rare} by enumerating random IVs assuming that F^{rv} have the variant and by finding the proportion of IVs that generate the same or larger S_{rare} values. We called this the "Family-level" p-value as we computed this p-value for each rare variant in each family. In addition to the family-level p-values, we computed a "Variant-level" p-value that meta-analyzed p-values across different families for the same rare variant. We combined p-values using the Fisher's method (the sum of log p-values). The direction of effect was consistent across different families as we were only interested in the enrichment of rare variants among affected individuals and depletion among unaffected individuals, and we were not interested in the opposite direction. We also computed a "Gene-level" p-value that meta-analyzed p-values across different rare variants and families in a gene. We used Fisher's method as well, and it is important to note that because of LD, a founder may have two or more rare variants in a gene with the same S_{rare} statistic and p-value. Because meta-analysis assumes independence among p-values, we included only one p-value from the same founder if there were multiple same p-values from this founder in a gene.

To detect F^{rv}, founders who introduced the rare variant into a family, we used the results of GIGI imputation. GIGI generated the probability of imputed genotypes for everyone in a family, even those who were not genotyped with the microarray (including all founders). Assuming bi-allelic variants, GIGI generated three probabilities for three genotypes (1/1, 1/2, and 2/2 where 1 and 2 are two alleles). We had two strategies to identify F^{rv}. In the first strategy, if the highest genotype probability among the three probabilities of a founder was > 0.8, we considered this founder to have a high-quality genotype. If all founders in a family had high-

quality genotypes, we included this rare variant for analysis and identified F^{rv} who had a rare variant. We used the second strategy when all founders except a pair of top founders had high-quality genotypes. A pair of top founders was a couple who did not both have parents in the pedigree structure, and they were not usually sequenced or imputed well. In some cases, GIGI assigned about 0.5 probability of having a rare variant to both top founders. This indicated that a rare variant was inherited from one of the couples, but GIGI was not able to accurately identify which founder introduced the variant. Assuming that both top founders were neither sequenced nor imputed well, and they did not contribute to the S_{rare} statistic, the segregation p-value would be the same regardless of which of the founders carried the rare variant. Hence, we randomly assigned the rare variant to one of the top founders in the second strategy. We ignored rare variants in which there was more than one pair of top founders in a family who did not have the high-quality genotypes as well as rare variants in which founders carried two alleles of a rare variant. We analyzed rare variants with low genotype missing rate (< 5%) that at least two affected individuals shared in a family as we were not interested in rare variants present in one or zero affected individuals.

Chapter 6:

Specific Enrichment of Genic Rare CNV Burden in Bipolar Disorder

6.1 Abstract

In the previous chapter, we explored the contribution of several classes of genetic variation on bipolar disorder in 26 large, multi-generational and phenotypically well-characterized pedigrees from two population isolates from Central and South America. We found limited evidence for the presence of highly penetrant variant(s) of large effect, consistent with decades of research on the genetics of bipolar disorder in both family and population-based studies. The lack of any clearly segregating variants suggest that even in large, multi-generational families with a high incidence of BP, the mode of inheritance is complex and highly polygenic. In agreement with this observation, affected individuals displayed a significantly elevated burden of rare deleterious SNVs distributed across a large set of BP-related genes. We demonstrated a likewise enrichment for rare CNVs detected using complementary methods based on both SNP microarray and WGS data. Here, we aim to recapitulate the latter findings in a larger North American BP sample of 2,424 unrelated individuals of European descent by analyzing rare CNVs detected genome-wide using microarrays. While there was no evidence of an increased genome-wide burden of rare CNVs in terms of CNV rate, size, or proportion, cases and controls differed significantly in terms of where rare CNVs were located. CNVs in BP cases affected more genes, particularly in the subset that is both brain-expressed and extremely intolerant to loss of function mutations in the general population. Additionally, we provide evidence for an increased female CNV burden for rare CNVs in BP.

6.2 Background

BP and SCZ are both debilitating, lifelong psychiatric disorders with high heritability, age of onset in late adolescence, and some degree of shared symptomatology, including anhedonia, cognitive impairment, and psychosis. Although clinically, these two diseases represent exclusive entities, research conducted over the last several decades, particularly in the field of molecular genetics, has pointed towards at least some overlap in disease etiology between the two. The relatedness between these two disorders has long been observed clinically; the term "schizoaffective psychosis" was first assigned by Kasanin in the 1930's to describe patients who exhibited symptoms of both, but with a general disease course that was less severe than SCZ alone (Abrams et al., 2008).

Early family studies (Gershon et al., 1982; Maier et al., 1993) observed that SCZ and BP tended to "breed true" within families, supporting the notion that the genetic liabilities for each disorder were distinct (Craddock et al., 2009). These studies, however, were limited in sample size. In one of the largest epidemiological studies ever conducted (Cross-Disorder Group of the Psychiatric Genomics et al., 2013), based on 2 million families taken from the Swedish national public register, Lichtenstein et al. demonstrated a significant elevation in risk for both SCZ and BD in offspring and siblings of individuals affected with either disorder. The relative risk for SCZ in a sibling of an individual affected with BP and vice versa were 3.9 (95 % confidence interval [CI] = 3.4 - 4.4) and 3.7 (95% CI = 3.2 - 4.2) respectively, suggesting at least a partial overlap in the genetic liability for SCZ and BP.

A growing number of GWAS have been performed in BP to date. The largest of these incorporated a primary sample in excess of 50,000 individuals and identified a total of 30 genome-wide significant loci (Stahl et al., 2017). These significant loci harbored a number of genes important for neuronal function, including ion channel proteins CACNA1C, SCNA2, and

SLC4A1, and synaptic proteins ANK3 and RIMS1. Notably, 8 out of the 30 genome-wide significant loci in BP are also significantly associated with SCZ (Schizophrenia Working Group of the Psychiatric Genomics, 2014), indicative of a sharing of genetic risk due to common variants associated with either disorder. The shared genetic liability between SCZ and BP extends far beyond these singly identified, significantly-associated risk loci from GWAS. Empirical evidence for the sharing of genetic risk factors between SCZ and BP had previously been obtained on a genome-wide level, from both genetic risk profile scoring (International Schizophrenia et al., 2009) and estimates of SNP-based co-heritability (Cross-Disorder Group of the Psychiatric Genomics, 2013).

There is overwhelming support for a significant contribution of rare CNVs to the genetic etiology of SCZ. The contribution of rare CNVs in the pathogenesis of SCZ has been demonstrated both genome-wide, as measured by and enrichment of global CNV load, and also through the identification of individual, significantly associated CNV loci. In a study from the International Schizophrenia Consortium, Stone et al. reported a subtle but highly significant 1.15-fold increase in the rate of rare CNVs > 100kb among SCZ cases compared to controls (International Schizophrenia, 2008). The largest CNV study in SCZ performed to date, involving 41,321 subjects, reaffirmed this enrichment of global rare CNV burden, and furthermore, provided strict genome-wide significant support for 8 distinct CNV loci that each confer substantial risk for the disorder (Marshall et al., 2016). Collectively, about 2.5% of SCZ patients carry such CNVs.

Given the clinical, epidemiological, and genetic overlap between SCZ and BP, a number of studies have explored if rare CNVs play a similar role in BP. Zhang et al. first reported that singleton deletions >100kb in length were significantly enriched (1.31-fold) in BP cases compared to controls (Zhang et al., 2009). However, in a larger sample from the WTCCC, Grozeva et al observed no apparent increase in the rate of rare (<1% MAF) or singleton CNVs

for BP (Grozeva et al., 2010). In fact, paradoxically, they reported trend for a reduced burden of rare CNVs among BP patients compared to controls, a finding which they again report in a larger study of the same BP samples compared to a larger set of controls (Grozeva et al., 2013). A trend of reduced burden has also been reported in separate, University College London BP sample (McQuillin et al., 2011). In the largest CNV survey performed to date, Green et al. compared the rates of large (>500kb) CNVs in 2,591 new BP cases to 6,882 SCZ samples and 8,842 controls drawn from prior and publicly available data. While the rate of large, rare CNVs among SCZ patients was significantly enriched compared to controls, they observed no significant difference for BP (Green et al., 2016). They did not observe a reduction of burden, as reported by other groups; however, as cases and controls were collected and genotyped separately on different platforms and drawn from several studies, they did not attempt to compare rates for smaller events, since these are more susceptible to batch effects.

If there is a conclusion to be drawn from the collective results from CNV studies in BP, it is that in contrast to SCZ, the evidence for the involvement of CNVs in BP is much less clear. Here, we investigate the role of rare CNVs in an independent sample of 2,366 individuals (1,353 cases and 1,013 controls) detected using SNP genotyping arrays. A key strength of the current study is that all samples (including controls) were recruited and evaluated under uniform methods within a single study, and collectively genotyped on the same platform at the same genotyping center, minimizing potential confounding due to ascertainment differences, population stratification, and of particular importance in the context of CNV studies, genotyping batch.

6.3 Sample collection, QC, and CNV calling

All samples used in this study are derived from a subset of samples collected as part of the GPC, an NIMH initiative to ascertain and enroll a large clinical cohort of SCZ and BP

patients along with geographically matched, non-psychiatric controls (Pato et al., 2013). This study focused on the contribution of rare structural variation in BP patients of self-reported European (EU) ancestry. CNVs were surveyed genome-wide on Illumina Omni Express v1.0 SNP genotyping chips with all patient DNA derived from whole blood. We performed extensive sample QC based on SNP genotype information, metrics of array intensity assay quality, and CNV calls (See Chapter 6.8 for details). We verified continental ancestry based on PCA alongside continental HapMap samples and further excluded ancestral outliers (Figure C-1.A and B). Our final sample consisted of 1,013 controls with no medical history of psychosis and 1,353 BP cases; 821 with a diagnosis of BP1, 47 with bipolar disorder, type II (BP2), and 485 with schizoaffective disorder, bipolar type (SAB) (Figure 6-1.A). Critically, the distributions of case and control samples were virtually indistinguishable across multiple intensity metrics (Figure 6-1.B), demonstrating that there was no difference in assay quality between phenotypic groups. Although there was a slight difference between cases and controls in genetic loading across the first principal component of ancestry (P = 0.0497, Kolmogorov-Smirnov test, Figure 6-1.C), calculation of the genomic inflation factor based on all available SNP data, without correction, showed minimal evidence for population stratification (λ_{gc} = 1.021, Figure C-1.C). We used a previously established pipeline to call CNVs (Chapter 2.2), and for compatibility with previous reports, we restricted our analysis to rare CNVs present in less than 1% of all samples, spanned by a minimum of 10 probes and with a minimum size of 100kb. In total, we resolved 1,707 CNVs (957 duplications and 750 deletions).

A

Diagnosis	Individuals	Males	% Male	Psychotic Features	% Psychosis
All Cases	1,353	696	51.4	1,134	83.8
Bipolar I	821	393	47.9	624	76.0
Bipolar II	47	15	31.9	25	53.2
Schizoaffective Disorder	485	288	59.4	485	100.0
Controls	1,013	438	43.2	n.a.	n.a.

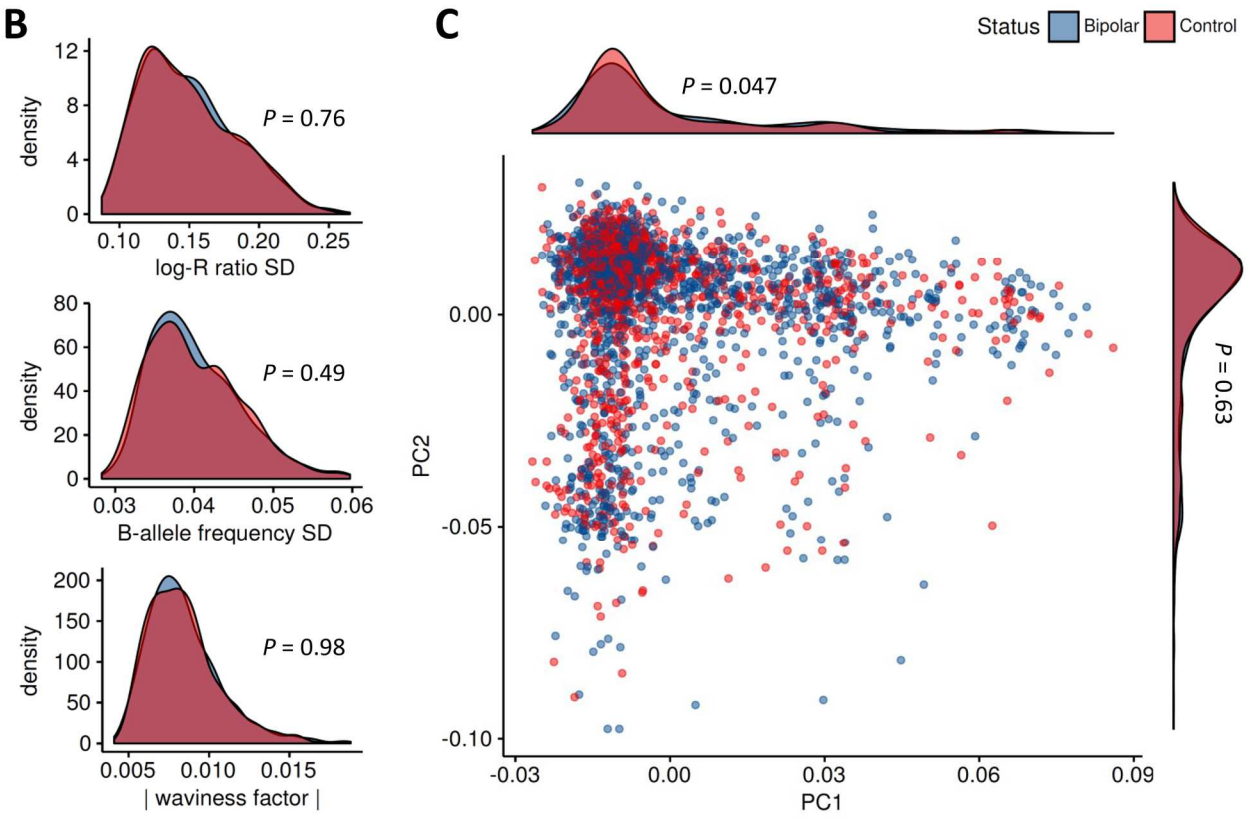


Figure 6-1. Sample characteristics of 1,353 BP cases and 1,013 controls.

(A) Clinical and demographic information for all individuals passing quality control used for the analysis in this study. (B) Density plots of the BP case and control samples across three different intensity metrics exhibit no differences in assay quality between phenotypic groups. (C) Scatterplot of the first two principal components calculated using SNP genotypes. There is a slight imbalance between case and control samples on PC1, however the λgc calculated using all available SNP data showed no evidence for population stratification (Figure C-1.C). Blue, BP samples; Red, control samples; P-values calculated using a two-sided Kolmogorov–Smirnov test,

6.4 Burden analysis: CNV event type and size

We first examined for global differences between BP subjects and controls in terms of both the overall rate of rare CNVs and the proportion of CNV carriers. Considering all rare events, there was no difference in CNV rate (0.705 CNVs per BP case vs. 0.704 per control; two-sided, empirical P=0.982, controlling for subject sex) or proportion (0.494 vs 0.480; empirical P=0.501) (Table 6-1.A). Similarly, no significant difference was observed when stratifying our analysis by event type and size, or when alternatively measuring genome-wide CNV burden in terms of average CNV size or aggregate CNV length per individual (Table 6-1.B). Additionally, we obtained similar results when we modeled global CNV burden under regression framework, both with and without the inclusion of potential confounding covariates (including subject sex, ancestry, and a metrics for CNV assay quality; data not shown). In an attempt to replicate a previous report of a significantly elevated rate of singleton deletions in BP (Zhang et al., 2009), we repeated these analyses restricted to events observed only once across our entire dataset. We saw no evidence for enrichment of singletons in either cases or controls (Table C-1). Taken together, we conclude that in our current sample, there is no appreciable difference in genome-wide burden in terms of overall CNV rate, proportion, and size, or the occurrence of singleton events.

6.5 Burden analysis: genic content

We next explored whether rare CNVs in BP differed with regard to their genomic context, specifically in the total number and type of genes they potentially affect. To concentrate on CNVs with an interpretable biological impact for BP, we limited our consideration to (i) all known protein coding genes (*all protein coding*, n=19,188) and specific subsets thereof, including (ii) genes without extreme intolerance to loss of function mutations (PLI < 0.9, *unconstrained*

A

	CNV RATE			PROPORTION OF CARRIERS		
CNV TYPE	BIP (n=1,353)	CTRL (n=1,013)	P-value	BIP	CTRL	P-value
100-500kb	787 (0.582)	604 (0.596)	0.679	0.433	0.428	0.835
DEL	361 (0.267)	289 (0.285)	0.412	0.232	0.240	0.663
DUP	426 (0.315)	315 (0.311)	0.887	0.260	0.260	1.000
>500kb	167 (0.123)	109 (0.108)	0.982	0.116	0.101	0.501
DEL	44 (0.033)	27 (0.027)	0.487	0.031	0.027	0.541
DUP	123 (0.091)	82 (0.081)	0.228	0.088	0.078	0.238
ALL CNVs	954 (0.705)	713 (0.704)	0.982	0.494	0.480	0.501
DEL	405 (0.299)	316 (0.312)	0.612	0.254	0.254	1.000
DUP	549 (0.406)	397 (0.392)	0.609	0.321	0.316	0.822

B

	AVERAGE KB			TOTAL KB		
CNV TYPE	BIP	CTRL	P-value	BIP	CTRL	P-value
100-500kb	206.2	202.5	0.531	276.0	281.6	0.621
DEL	192.5	187.4	0.495	220.7	221.7	0.929
DUP	217.8	220.0	0.792	262.8	261.9	0.947
>500kb	1024.0	1183.0	0.196	1084.0	1247.0	0.205
DEL	1193.0	1198.0	0.984	1260.0	1198.0	0.800
DUP	960.8	1164.0	0.156	985.2	1201.0	0.133
ALL CNVs	347.0	358.4	0.676	500.9	515.6	0.711
DEL	302.0	268.1	0.317	360.7	335.5	0.554
DUP	384.2	437.3	0.228	485.3	517.3	0.519

Table 6-1. Global burden analysis of rare CNVs: rate, proportion, and size.

(A) Burden analysis comparing the total number of CNVs and average CNV rate (in parenthesis), and also in terms of the proportion of CNV carriers, stratified by CNV type and event size. (B) Burden analysis by CNV size, comparing either the average length of CNVs or the total length of all CNVs between BP cases and controls. P-values determined empirically by performing 100,000 label-swapping permutations, controlling for subject sex

genes, n=14,753); (iii) genes expressed in brain (*brain expressed* genes, n=10,346); (iv) genes extremely intolerant of loss of function mutations (PLI > 0.9, *constrained* genes, n=3,488); and (v) genes both expressed in the brain and extremely intolerant to loss of function mutations (*brain expressed + constrained*, n=2,891). We quantified genic CNV burden as the total number of genes impacted by rare CNVs per individual, and compared cases and controls using both linear and logistic regression, correcting for potential confounds including subject sex, ancestry, and differences in CNV detection rate. Despite having no demonstrable increase in CNV rate or size, we observed a significant excess in CNV burden when measured in terms of the affected number of genes per individual. BP cases exhibited a significant excess of protein coding genes affected by rare CNVs (0.446 genes per individual; 95% CI = 0.188-0.703; $P = 7.1 \times 10^{-4}$) and, concomitant with this general increase in genic burden throughout the genome, a significant excess was also observed in all subsets of protein coding genes examined (Figure 6-2.A). In particular, BP cases exhibited the most significant excess in the smallest fraction of genes considered, those that were both highly intolerant to loss of function mutations and also expressed in the brain (excess of 0.083 brain expressed + constrained genes / individual; 95% CI = 0.040-0.125; $P = 1.3 \times 10^{-4}$). The P values determined from linear regression were in good agreement with those obtained by performing 100,000 permutations of phenotype labels ($P = 1.6 \times 10^{-3}$ for all protein coding genes and $P = 4.2 \times 10^{-4}$ for brain expressed + constrained genes, Table C-2), demonstrating that our regression model was well-calibrated.

The excess of 0.446 protein coding genes/individual translated into a modest but significant 5.1% relative enrichment of affected genes in BP cases compared to controls (OR = 1.051, 95% CI = 1.021-1.084, $P = 1.1 \times 10^{-3}$, logistic regression, Figure 6-2.A). The enrichment over controls was higher for the set of constrained genes (OR = 1.286, 95% CI = 1.104-1.505, $P = 1.5 \times 10^{-3}$) and greatest overall for the set of constrained genes also expressed in the brain (OR = 1.390, 95% CI = 1.175-1.659, $P = 1.8 \times 10^{-4}$).

Given the significant genome-wide enrichment for BP observed across all protein coding genes, we conducted an additional analysis, adjusting for this genome-wide excess (see Chapter 6.8). Although this approach substantially diminishes P values and is likely conservative (as CNVs often span multiple genes), these corrected ORs serve to distinguish whether the enrichment of particular subset of genes is specific or merely reflective of the global increase in genic burden observed for BP in general. Following this stringent correction, only the set enrichment of brain expressed + constrained genes remained significantly enriched (corrected OR = 1.278, 95% CI = 1.047-1.565, P = 0.017), and the set of unconstrained genes showed signs of depletion (corrected OR = 0.828, 95% CI = 0.704-0.972, P 0.022, Figure 6-2.A).

Separating across diagnostic categories, this similar pattern of genic enrichment was observed in both BP1 (Figure 6-2.B) and SAB patients (Figure 6-2.C). Notably, despite a reduction in sample size, genic enrichment was both greater and more significant when considering only the subset of BP1 cases compared to all affected individuals at large. BP1 patients had, on average, 46% more brain expressed + constrained genes affected by rare CNVs compared to controls (OR = 1.463, 95% CI = 1.212-1.781, P = 1.0×10^{-4}), and this enrichment remained highly significant even after accounting for the general genome-wide enrichment for all protein coding genes (OR = 1.396, 95% CI = 1.129-1.737, P = 9.8×10^{-3}). Although the small number of BP2 individuals in our dataset prohibits any definitive conclusion, this clinical subgroup, by contrast, showed no trend of enrichment in either constrained or brain-expressed + constrained genes (Figure C-2), and was excluded from subsequent analyses.

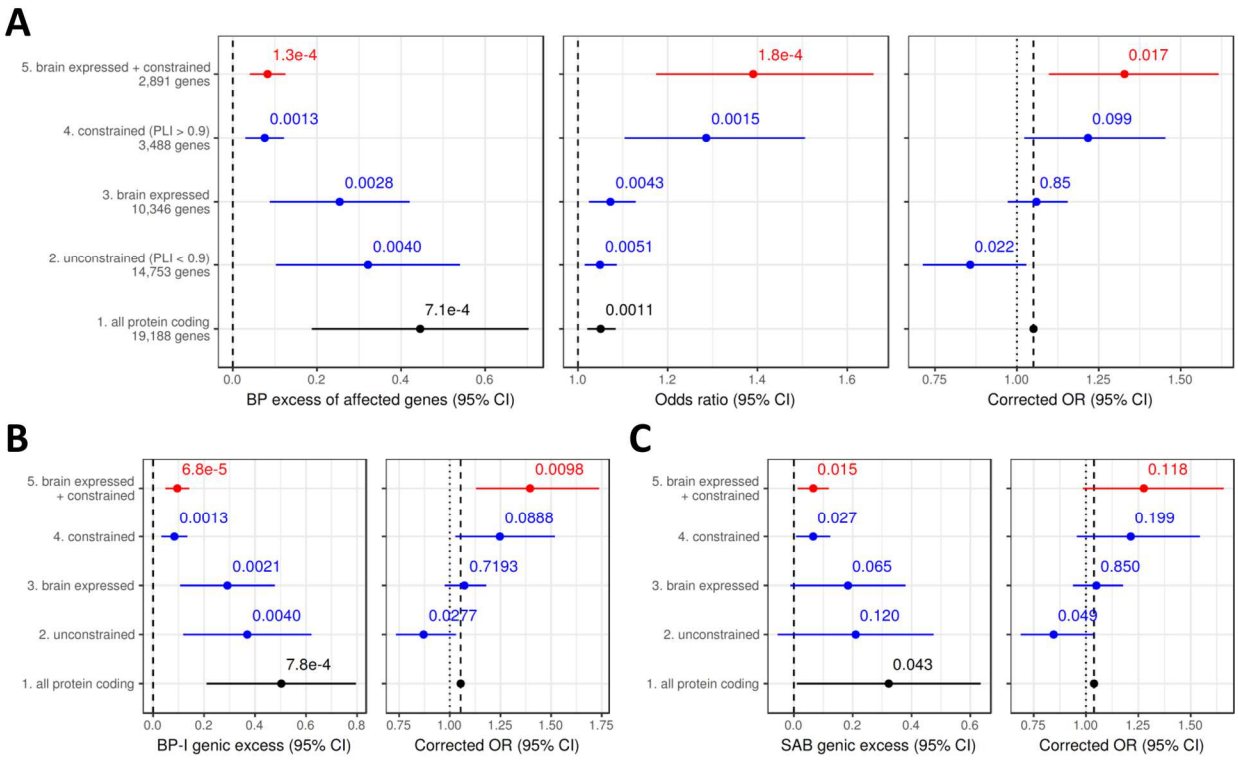


Figure 6-2. Genic burden of rare CNVs among sets of protein coding genes.

(A) Genic burden analysis comparing the number of genes affected by rare CNVs between all BP cases and controls for different sets of protein coding genes. Excess of affected genes (left) and odds ratio (middle) calculated using linear regression and logistic regression, respectively, controlling for subject sex, ancestry, and genome-wide rate of all rare CNVs. Corrected OR (right) shows odds ratios calculated with including the total number of affected protein coding genes as a covariate, and represents a conservative estimate of the enrichment of each subset of protein coding genes beyond the genome-wide excess that is observed in BP cases (see Chapter 6.8). Genic excess and corrected ORs for (B) BP1 and (C) SAB alone. BP2 alone did not exhibit a similar pattern of genic enrichment (Figure C-2).

6.6 Evidence for increased female CNV burden in BP

When fitting our models for genic burden, we observed that subject sex was often a significant predictor. Since an increased female burden of large and/or pathogenic CNVs has been reported in other disorders as well as in the general population, we decided to explore this further. We first evaluated the genic burden for the set of brain expressed + constrained genes, comparing BP1 and SAB individuals collectively to controls separately for female and male

subjects. We observed that, while overall, genic enrichment was significant for both sexes (female genic OR = 1.483, 95% CI = 1.184-1.883, P = 8.5 x 10^{-5}; and male genic OR = 1.396, 95% CI = 1.049-1.889, P = 0.025; Figure 6-3.A), genic burden differed between sexes when we partitioned our data by CNV size (100-500 kb and > 500 kb). In females, rare CNVs within both categories contributed substantially to the overall genic enrichment signal. By contrast, for males, the increase in genic burden was largely attributable to smaller CNVs < 500 kb (OR = 1.595, 95% CI = 1.061-2.449, P = 0.028) and there was no significant enrichment for CNVs > 500 kb (OR = 1.072, 95% CI = 0.754-1.534, P = 0.693).

Next, we partitioned our CNVs into discrete categories with increasing levels of presumed pathogenicity, and compared the overall rate of events between BP cases and controls of the same sex (Figure 6-3.B). Overall, we observed a trend of increased burden measured in terms of the number of CNVs that correlated with increasing pathogenicity, but only among females. CNVs > 500kb affecting a brain expressed + constrained gene were significantly more numerous among BP females (OR = 1.767, 95% CI = 1.059-3.032, P = 0.033) and known pathogenic CNVs exhibited a striking four-fold increase compared to female controls (OR = 3.97, 95% CI = 1.604-11.953, P = 5.9 x 10^{-3}).

6.7 Discussion

Although the sample size of the current study is limited compared to some previous CNV studies in BP, our uniform ascertainment, data generation, QC, and analytical procedures resulted in a dataset that showed minimal evidence for selection bias or case-control differences in assay quality, making it ideal for evaluating the contribution of rare CNVs in aggregate. We found that while the average number, occurrence of carriers, or size of rare CNVs did not substantially differ between cases and controls, the location of these rare events did; specifically, in the number and type of genes affected. Rare CNVs among BP cases affected, on

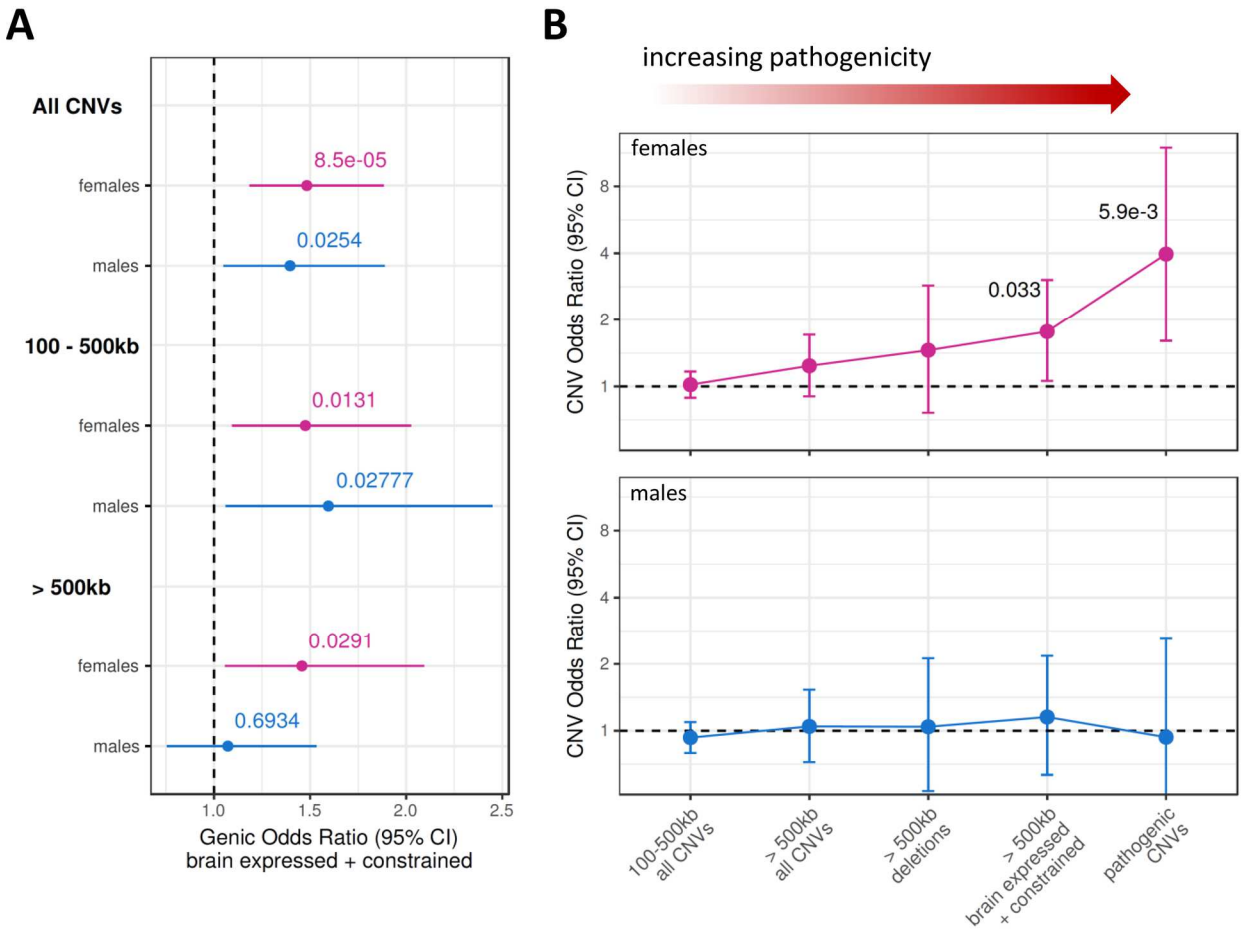


Figure 6-3. Gender differences in BP in for rare CNVs affecting highly constrained, brain-expressed genes and known, pathogenic events.

(A) Comparison of genic burden for CNVs affecting brain expressed + constrained genes between case and controls, stratified by subject sex and CNV size. Female BP cases exhibit a uniform pattern of genic enrichment compared to female controls. For male subjects, the increased burden observed among BP cases is largely due to CNVs < 500kb. (B) Burden analysis comparing the number of CNVs observed in BP cases compared to controls, stratified by presumed pathogenicity of events. Female BP subjects show a trend of increasing burden that correlates with presumed pathogenicity, with a significant enrichment of large CNVs that affect brain expressed + constrained genes (OR = 1.767, 95% CI = 1.059-3.032, P = 0.033) and known pathogenic CNVs (OR = 3.97, 95% CI = 1.604-11.953, P = 5.9 x 10^{-3}). Male subjects, by contrast, do not exhibit this pattern.

average, 0.446 more protein coding genes per BP individual compared to controls. While this represents a significant enrichment, the effect size (OR of 1.051 per protein coding gene) is

comparably modest to what has been consistently reported for both ASD and SCZ, although methodological differences make a definitive comparison across disorders difficult.

Partitioning this genic burden across different subsets of protein coding genes, we determined that the source of this genic enrichment was not distributed randomly across all protein genes throughout genome, but rather concentrated among a specific subset: genes that are both highly intolerant to disrupting mutations and also expressed in the brain. BP1 cases displayed the strongest genic enrichment among this subset, even after conservatively correcting for the general enrichment across all protein coding genes (OR = 1.396). A likewise concentration of rare variation, including both CNVs and also likely gene-disrupting SNVs, has now been observed across similar subsets of genes within several other neuropsychiatric disorders, including ASD, ID, and SCZ.

Although an increase in burden across brain expressed + constrained genes presented in both males and females with BP, there was a distinct gender difference in terms how this genic burden distributed across rare CNVs. In females, CNVs across all size ranges contributed, whereas in males, this genic burden was mostly limited to rare CNVs < 500 kb in length. Consistent with this, we also observed a gender difference in the occurrence of rare CNVs when partitioned according to pathogenicity. BP females exhibited a higher mutational burden of large, rare CNVs, particularly in those well-established as definitive pathogenic events. An increased female load of large, and/or known pathogenic CNVs has been previously been reported for both ASD and SCZ (Han et al., 2016; Jacquemont et al., 2014; Levy et al., 2011), lending credence to the hypothesis of a female protective role as a possible cause of the male:female bias observed in these disorders. The gender differences observed here represent a novel finding for BP, and one of particular interest since, in contrast to ASD and SCZ, BP is largely acknowledged to exhibit no gender-specific bias in overall disease prevalence (Barnett and Smoller, 2009). Limited epidemiological evidence does suggest, however, that psychosis, a

symptom shared with SCZ and also prevalent among cases in this study, does display some gender imbalance with regard to both transmission and prevalence (Jongsma et al., 2018; Nordentoft et al., 2006), and we suspect that the high incidence of this symptom among BP individuals in our study could contribute to the differences observed here. Clearly, larger, well-phenotyped samples in BP will be necessary to explore these possibilities further.

6.8 Methods

Study samples and genotyping

The BP samples used in this study are a subset of samples collected as part of a large NIMH initiative to ascertain and enroll a large clinical cohort of SCZ and BD patients alongside non-psychiatric controls. Details regarding subject recruitment, patient assessment, and exclusion criteria have been described in detail elsewhere (Pato et al., 2013). Briefly, GPC participants and geographically-matched controls were recruited from a North American population at clinical sites primarily among 8 different states, including California, Georgia, Massachusetts, North Carolina, New York, Texas, and Ohio. Both cases and controls were subjected to an initial screening questionnaire, and likely cases were additionally interviewed using a validated diagnostic instrument. Control individuals with a potential history of psychosis or mania were excluded.

Genotyping, CNV calling, and sample QC

DNA was extracted from whole blood for all subjects using standard protocols, and genotyping was performed by the Broad Institute (Boston, MA) on the OmniExpress v1.0 array (Illumina) according to manufacturer's protocols. Array data was generated on an initial set totaling of 3,194 samples. To improve the accuracy of intensity measures, we excluded poorly performing samples with an initial call rate < 0.98 (n=71), and generated new cluster definitions for all SNPs across based on all remaining samples before extracting genotype calls, LRR and

BAF values in GenomeStudio 2.0 (Illumina). Samples with a discordant sex status or excessive heterozygosity were removed. Genotype data was combined with publically available HapMap data generated on the same array (Illumina), and a total of 109,867 LD-thinned markers were included for PCA analysis to exclude non-EU samples and inspect for cryptic relatedness using PLINK v1.09 (Chang et al., 2015). For all pairs of samples with a pi-hat value > 0.2, we excluded the sample with the lower genotyping call rate.

To detect CNVs genome-wide, we used a previously established pipeline (Chapter 2.2). Briefly, we performed segmentation on all with all using two HMM-based CNV calling algorithms, PennCNV v2011-05-03 (Wang et al., 2007), and QuantiSNP v2.0 (Colella et al., 2007). To improve specificity, only CNVs detected by both algorithms and of the same relative type (gain or loss) were retained. We excluded samples with poor intensity measures by removing outliers (mean +- 3SD or by manual inspection) with regard to several different metrics generated by PennCNV. We also removed samples with excessive CNV load in terms of total number of CNV calls made or aggregate CNV length, as determined by either calling algorithm independently. Our final sample, after QC using SNP genotype data, intensity metrics, and overall CNV load, consisted of Adjacent CNV calls of the same type were iteratively merged if the number of intervening probes between each segment was less than 20% of the combined probe density. CNVs that spanned centromeres, or overlapped by more than 50% of their total length with telomeres, segmental duplications, or VDJ recombination regions were also removed. We restricted our analysis to rare CNVs (defined as a MAF < 1% across all samples based on a 50% reciprocal overlap) at least 100 kb in length and spanned by a minimum of 10 probes, and performed in silico validation on all rare CNVs as previously described (Huang et al., 2017).

Burden Analysis

To perform burden analysis by gene count, we first annotated all rare CNVs by the number of overlapping genes contained within each of the following five gene sets:

1. To generate a list of known protein coding genes (all protein coding genes), we used BioMart (http://grch37.ensembl.org/biomart/martview/) to filter for genes in Ensembl v91 with the gene type “protein coding.” To enable compatibility between external resources, we additionally filtered for genes with an HGNC symbol identifier (n=21,252 genes). Removal of genes present on unplaced or alternate contigs and the mitochondrial genome resulted in a list of 19,188 genes.
2. To create a list of genes without an extreme intolerance to loss of function mutations (unconstrained genes), we downloaded from ExAC release 3.1, and selected corresponding gene symbols with a pLI score < 0.9. We used this to subset the list of all protein coding genes (1), resulting in 14,590 genes.
3. To obtain the subset of protein coding genes expressed in the brain (brain expressed genes), we selected all genes from Supplement 1 of Fagerberg et al. with an FPKM (fragments per kilobase of exon model per million mapped reads) value > 5. We then subset this list from (1) to obtain 10,346 genes.
4. For the list of genes extremely intolerant to loss of function mutations (constrained genes), we selected for genes with a pLI score >= 0.9 and used this to subset our list of all protein coding genes, resulting in 3,488 genes.
5. The list of brain expressed + constrained genes was generated by taking the intersection of (3) and (4) above, resulting in 2,891 genes.

For each set, to estimate the excess of genes affected by rare CNVs or odds ratios for BP using the following linear and logistic regression models, respectively:

$$genes_i = \beta_0 + \sum_{x=1}^{3} \beta_x PCx_i + \beta_4 sex_i + \beta_5 length_i + \beta_6 status_i + \epsilon$$

$$\log\left(\frac{P_{i,case}}{1-P_{i,control}}\right)=\beta_0+\sum_{x=1}^{3}\beta_x(PCx_i)+\beta_4 sex_i+\beta_5 length_i+\beta_6 genes_i+\epsilon$$

where for individual i, $genes_i$ is the number of genes in the gene set of interest affected by rare CNVs, PCx_i is the x^{th} principal component of ancestry derived from SNP genotype data, sex_i is the individual's gender, $length_i$ is the aggregate length of all rare CNVs (without regard to genic content), $status_i$ is the individual's affection status, and ε represents an error term.

To calculate corrected ORs for each gene set, we added an additional term to the model that included the total number of protein coding genes affected by all rare CNVs within an individual. Including this extra term, in effect, captures the genic enrichment for the gene set of interest beyond the general increase observed among all protein coding genes for individuals with BP.

Appendix A: Supplementary Material for Chapter 3

A. Studies and genotyping								
GENOTYPING BATCH	**ARRAY**		**CENTER**		**PHENOTYPES**			
CC	OmniExpress v 1.0		Cardiff		Control			
CNP	OmniExpress v 1.0		Broad		Control/Clinical			
GPC	OmniExpress v 1.0		Broad		Control/Clinical			
WTCCC2	OmniExpress v 1.0		Cardiff		Control			
TS1	OmniExpress Exome v 1.1		UCLA		Control/TS			
TS2	OmniExpress Exome v 1.1		UCLA		Control/TS			
TS3	OmniExpress Exome v 1.1		UCLA		Control/TS			
B. QC summary								
QC STEP	**CC**	**CNP**	**GPC**	**WTCCC2**	**TS1**	**TS2**	**TS3**	**TOTALS**
Sample Genotypes	1,146	1,511	3,197	960	1,152	2,160	136	10,262
Pre-cluster QC	1,141	1,510	3,126	870	1,148	2,152	135	10,082
Sex Concordance	1,141	1,510	3,125	870	1,146	2,149	135	10,076
Replicates/Loading Control	1,141	1,491	3,081	870	1,134	2,143	134	9,994
Cryptic Relatedness	1,106	1,430	2,914	855	1,121	2,110	133	9,669
Clinical Phenotype	1,106	1,268	1,342	855	1,121	2,110	133	7,935
EU Ancestry	1,101	646	1,232	842	1,076	2,001	129	7,027
Heterozygosity	1,089	644	1,223	837	1,069	1,986	129	6,977
Intensity QC	1,068	634	1,143	810	959	1,805	116	6,535
CNV Load QC	1,067	634	1,141	808	958	1,803	116	6,527

Table A-1. Sample genotyping and QC summary.

(A) Summary of included studies and genotyping information. Sample phenotypes, genotyping platform, and genotyping center for different datasets collected for this study are shown, separated by study. (B) Summary of quality control procedures by study. The number of samples remaining within each batch after each successive quality control step is shown. Study abbreviations: Cardiff Controls (CC), Consortium for Neuropsychiatric Phenomics (CNP), Genomic Psychiatry Cohort (GPC), Wellcome Trust Case-Control Consortium (WTCCC2) and TS cases and controls collected for this study (TS1-3).

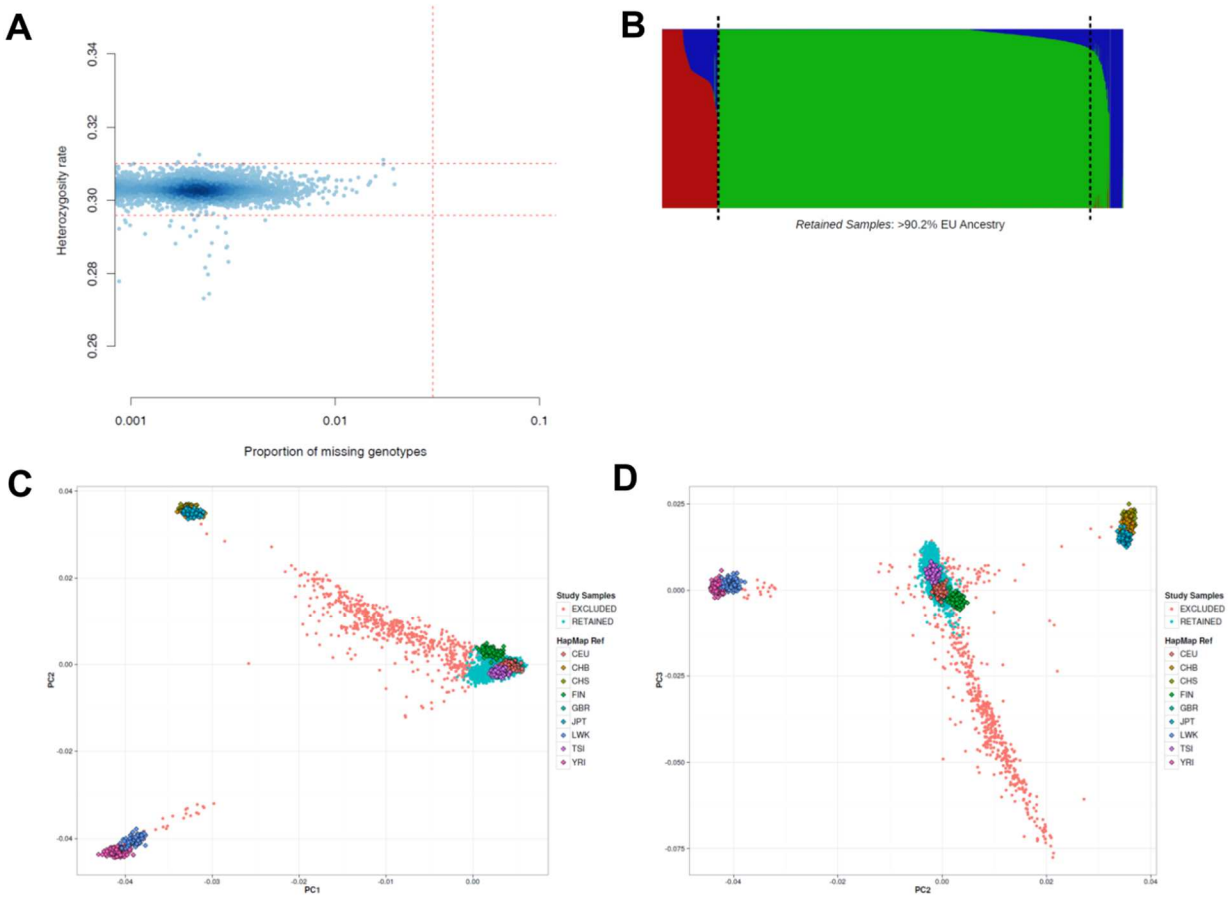


Figure A-1. SNP-based quality control and ancestry determination.

(A) Exclusion of sample outliers based on heterozygosity, mean +/- 3 SD (red dotted lines). (B) Exclusion of non-European samples based on ethnicity estimation using fastStructure with HapMap continental groups and K=3 clustering. Samples with > 9.85% non-EU ancestry were excluded. This threshold was calibrated against the maximum of reference HapMap/1000 Genomes European groups CEU, GBR, and TSI. The results of principal component (PC) analysis for the cohort and reference groups are plotted along (C) PCs 1 and 2 and (D) PCs 2 and 3. Retained samples and excluded samples are shown in cyan and pink, respectively. CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; CHB, Han Chinese in Beijing, China; CHS, Southern Han Chinese; FIN, Finnish in Finland; GBR, British in England and Scotland; JPT, Japanese in Tokyo, Japan; LWK, Luhya in Webuye, Kenya; TSI, Toscani in Italia; YRI, Yoruba in Ibadan, Nigeria.

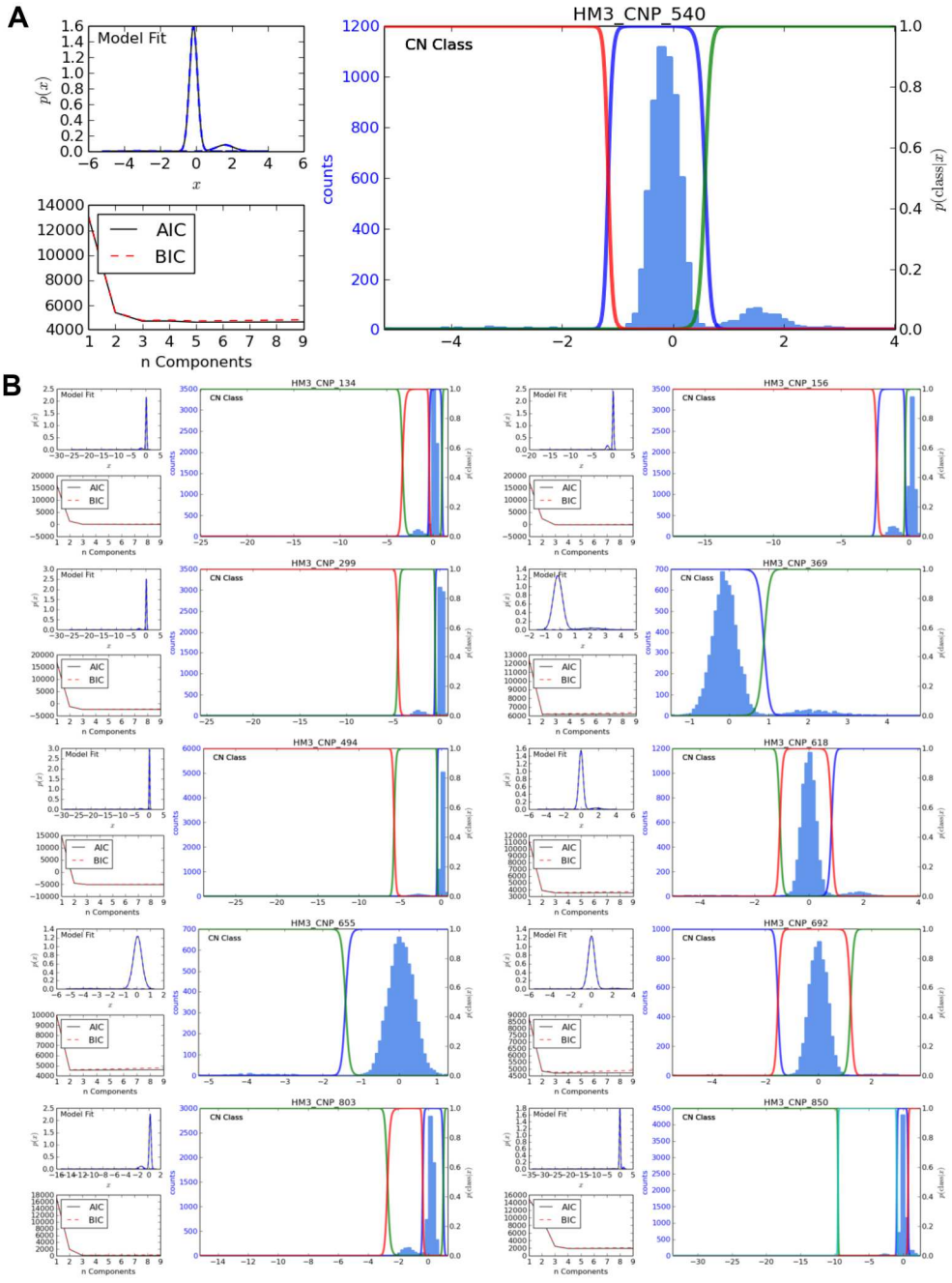


Figure A-2. GMM clusters of common HapMap3 CNVs.

(A) A representative GMM cluster plot for locus HM3_CNP_540. Subplots for each CNV depict, counter-clockwise: the best-fit model, Akaike and Bayesian Information Criterion metrics calculated for GMM fitting 1-9 components, and the posterior probability for CNV cluster assignment (colored lines) overlaying the distribution of median summarized intensity values for all samples across region calculated using the best-fit model. (B) GMM plots for the 10 additional HapMap3 CNV loci that were used to critically evaluate sensitivity between cases and controls (Tables A-2 and A-3).

GMM calls at common HapMap3 CNVs								
CNV_ID	CLUSTER	CLUSTER_LRR	GMM_COPY	CTRL_CALLS	CASE_CALLS	CTRL_FREQ	CASE_FREQ	p-value
HM3_CNP_134	1	-13.17478812	0	7	4	0.002	0.002	1.0
HM3_CNP_134	2	-1.544234008	1	296	191	0.072	0.078	0.4
HM3_CNP_156	1	-1.141264855	1	517	315	0.126	0.129	0.7
HM3_CNP_156	2	-10.96495207	0	18	13	0.004	0.005	0.6
HM3_CNP_299	1	-2.148959932	1	275	142	0.067	0.058	0.2
HM3_CNP_299	2	-20.27406193	0	4	3	0.001	0.001	0.7
HM3_CNP_369	1	2.201328191	3	234	137	0.057	0.056	0.9
HM3_CNP_494	1	-2.7402128	1	196	100	0.048	0.041	0.2
HM3_CNP_494	2	-23.74078464	0	1	1	0.000	0.000	1.0
HM3_CNP_540	1	1.648736036	3	392	263	0.096	0.108	0.1
HM3_CNP_540	2	-3.301784945	1	32	17	0.008	0.007	0.8
HM3_CNP_618	1	-3.470922513	1	44	24	0.011	0.010	0.8
HM3_CNP_618	2	1.817743156	3	167	91	0.041	0.037	0.5
HM3_CNP_655	1	-3.645609914	1	45	34	0.011	0.014	0.3
HM3_CNP_692	0	-4.175170396	1	10	11	0.002	0.005	0.2
HM3_CNP_692	2	2.34262532	3	47	32	0.011	0.013	0.6
HM3_CNP_803	1	-10.64502833	0	15	14	0.004	0.006	0.2
HM3_CNP_803	2	-1.347106673	1	417	258	0.102	0.106	0.6
HM3_CNP_850	1	-31.82000658	0	1	1	0.000	0.000	1.0
HM3_CNP_850	2	-2.649145422	1	74	49	0.018	0.020	0.6
HM3_CNP_850	3	1.410233747	3	165	101	0.040	0.041	0.8

Table A-2. GMM clustered genotype calls at common HapMap 3 CNVs.

For sensitivity analysis, all 6,427 samples used in this study were genotyped across 11 common Hapmap3 CNVs using a locus-specific GMM-based clustering method. CNV_ID, HapMap3 accession number; CLUSTER_ID, Arbitrary identifier assigned by the clustering algorithm; CLUSTER_LRR, The mean value of all median-summarized standardized intensity values (LRR-Z) for all samples assigned to the cluster; GMM_COPY, Inferred copy-number state. Call frequencies (FREQ) for 4,093 controls (CTRL) and 2,434 TS cases (CASE) reflect the proportion of GMM-based genotype calls with >0.95 posterior probability of cluster assignment. There was no significant difference in CNV genotype frequency between phenotypic groups at any of the 21 non-reference genotype calls across all 11 loci (2-sided Fisher's exact test).

A. Sensitivity analysis by locus									
CNV_ID	CNV_TYPE	GMM_TOTAL	GMM_CTRL	GMM_CASE	HMM_CTRL	HMM_CASE	CTRL_SENSE	CASE_SENSE	p-value
HM3_CNP_134	DEL	498	303	195	300	194	0.99	0.995	1.0
HM3_CNP_156	DEL	863	535	328	531	321	0.993	0.979	0.11
HM3_CNP_299	DEL	424	279	145	279	145	1.000	1.000	1.0
HM3_CNP_369	DUP	371	234	137	208	122	0.889	0.891	1.0
HM3_CNP_494	DEL	298	197	101	197	101	1.000	1.000	1.0
HM3_CNP_540	DUP	655	392	263	391	261	0.997	0.992	0.57
HM3_CNP_540	DEL	49	32	17	32	17	1.000	1.000	1.0
HM3_CNP_618	DEL	68	44	24	44	24	1.000	1.000	1.0
HM3_CNP_618	DUP	258	167	91	166	90	0.994	0.989	1.0
HM3_CNP_655	DEL	79	45	34	45	34	1.000	1.000	1.0
HM3_CNP_692	DEL	21	10	11	10	11	1.000	1.000	1.0
HM3_CNP_692	DUP	79	47	32	47	32	1.000	1.000	1.0
HM3_CNP_803	DEL	704	432	272	428	272	0.991	1.000	0.16
HM3_CNP_850	DEL	125	75	50	75	50	1.000	1.000	1.0
HM3_CNP_850	DUP	266	165	101	164	98	0.994	0.970	0.15
B. Overall sensitivity across common CNVs									
CNV_TYPE	GMM_TOTALS	GMM_CTRL	GMM_CASE	HMM_CTRL	HMM_CASE	CTRL_SENSE	CASE_SENSE	p-value	
DEL+DUP	4758	2957	1801	2917	1772	0.986	0.984	0.53	
DEL	3129	1952	1177	1941	1169	0.994	0.993	0.81	
DUP	1629	1005	624	976	603	0.971	0.966	0.65	
C. Group-wise sensitivity analysis across individuals									
CNV_TYPE	CTRL_SENSE	Std. Error	CASE_SENSE	Std. Error	p-value				
DEL+DUP	0.989	0.002	0.983	0.003	0.15				
DEL	0.996	0.001	0.991	0.002	0.14				
DUP	0.973	0.005	0.967	0.007	0.46				

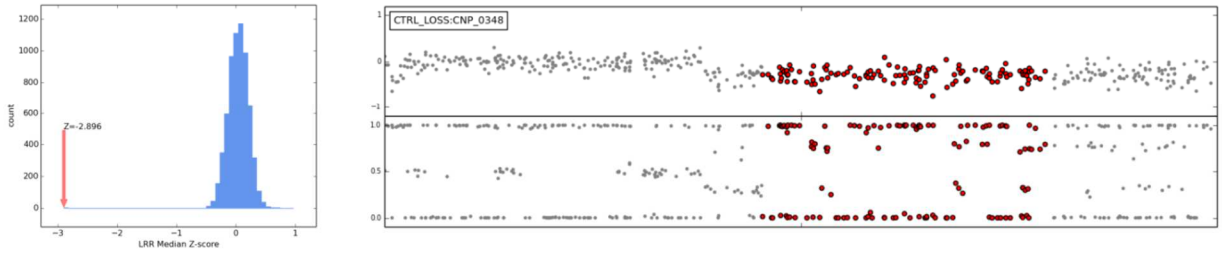
Table A-3. Sensitivity analysis of consensus HMM segmentation calls.

(A) Comparison of CNV detection sensitivity between cases and controls for each locus individually. The sensitivity of HMM calling for each locus was defined as the number of concordant HMM calls divided by the total number of non-reference genotypes determined through GMM-based clustering, a more sensitive, locus specific method. GMM genotypes were collapsed into calls of the same class (CNV_TYPE: DEL, all deletions; DUP, all duplications). P-values were calculated using a 2-sided Fisher exact test. (B) Overall sensitivity across all loci, stratified by CNV_TYPE. P-values were calculated using a 2-sided Fisher's exact test, comparing concordance rates between cases and controls. (C) Group-wise comparison of sensitivity between cases and controls based on the sensitivity calculated for each individual. No significant difference was observed between phenotypic groups whether considering deletions, duplications, or both in concert P-values calculated using a 2-sided Welch's *t*-test, comparing the average sensitivity by individual between phenotypic groups.

A

SAMPLE_ID	CHR	BP1	BP2	COPY	#SNPS	START_SNP	END_SNP	LRR-Z	BAF_{del}	BAF_{dup}	OUTLIER-Z
CNP_0348	6	67801176	67887156	1	21	rs9363696	rs16899159	-2.96	0.62	0.38	0.00015
CNP_0348[1]	**6**	**67907952**	**68586809**	**1**	**120**	**rs12197620**	**rs9354637**	**-2.90**	**0.76**	**0.16**	**0.00015**
CNP_0348	6	68707131	69142008	1	96	rs4707250	rs9363918	-2.93	0.73	0.17	0.00015
WT_0533[2]	**10**	**47375657**	**47703869**	**3**	**48**	**rs28599894**	**rs4434935**	**2.48**	**0.33**	**0.63**	**0.09**
CC_0852	10	47375657	47703869	3	48	rs28599894	rs4434935	2.36	0.33	0.60	0.09
WT_0866	10	47375657	47703869	3	48	rs28599894	rs4434935	2.33	0.38	0.60	0.09
TS_0457	10	47375657	47703869	3	48	rs28599894	rs4434935	2.29	0.39	0.60	0.09
TS_1843	10	47375657	47703869	3	48	rs28599894	rs4434935	2.05	0.39	0.60	0.09

B



C

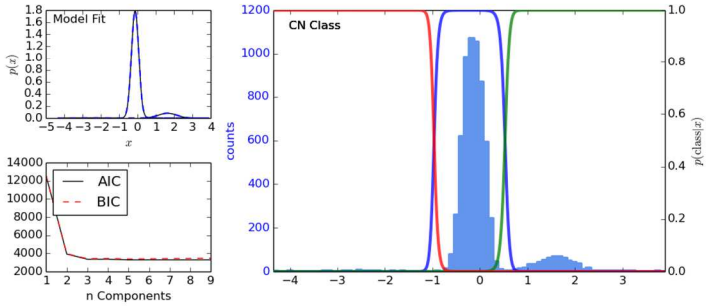


Figure A-3. *In silico* validation of CNV calls.

(A) Representative CNVs scored with various CNV validation metrics. Abbreviations: median summarized intensity measures across a CNV locus, standardized by sample (LRR-Z), proportion of probes with a BAF banding pattern indicative of a duplication event (BAF-D), proportion of samples with LRR-Z scores indicative of a polymorphic event (OUTLIER-Z). (B) Example of a large singleton mosaic event flagged for exclusion indicated as (1) in (A). This CNV on chromosome 6 was detected as three separate CNVs. The largest CNV call exhibits an LRR-Z score of -2.86 (left, red arrow), indicative of a deletion, but shows a clear BAF-banding pattern of a duplication event (right), with a BAF_{dup} score of 0.16. This is indicative of a mosaic event, where only a proportion of cells from sample CNP_0348 harbor the deletion event. (C) Example of a polymorphic CNV on chr10:47,375,657-47,703,869 misclassified as a rare event due to reduced sensitivity, indicated as (2) in (A), with an OUTLIER-Z score of 0.09. Genotyping by GMM-based clustering indicated that this misclassified rare event has a MAF of 0.12.

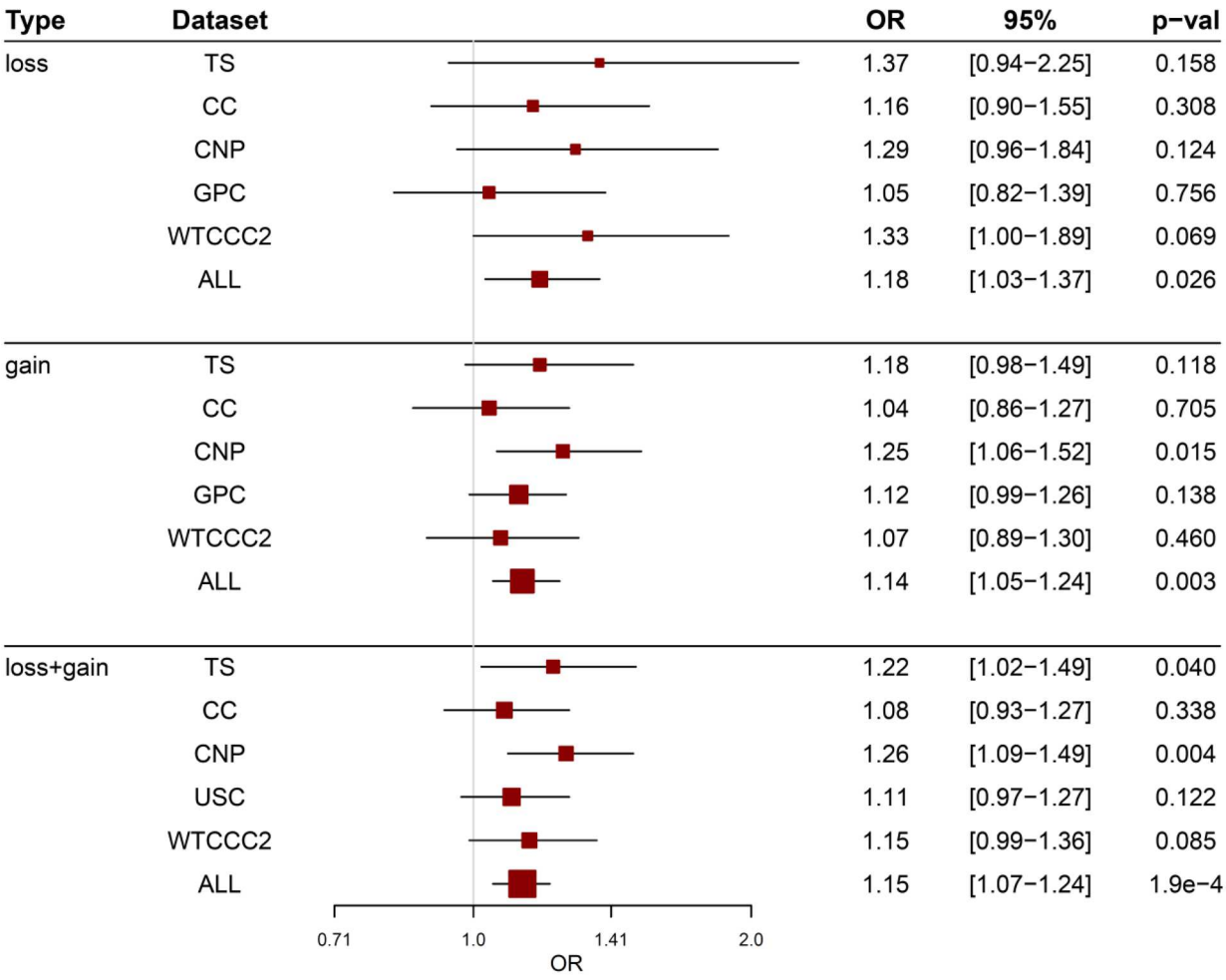


Figure A-4. Elevated CNV burden is consistent across datasets.

We assessed for increased CNV burden using different metrics and found that total CNV length was most significantly associated with an increased risk for TS (Figure 2). To ensure that the enrichment signal was not driven by a single dataset, here we repeated the assessment of burden by total CNV length, examining all TS samples compared to each of the control sample sets individually and to all control samples together. An increased burden is consistent across all datasets, and additionally when stratified by CNV type: loss (deletions); gain (duplications) and loss + gain (both deletions and duplications). Red boxes, ORs; Black lines, 95% CIs. TS, controls collected and genotyped alongside TS cases; CC, CNP, USC, WTCCC2, control samples taken from external datasets. ORs calculated from logistic regression using a standardized burden metric. ORs > 1 represent an increase in risk for TS per unit of CNV burden.

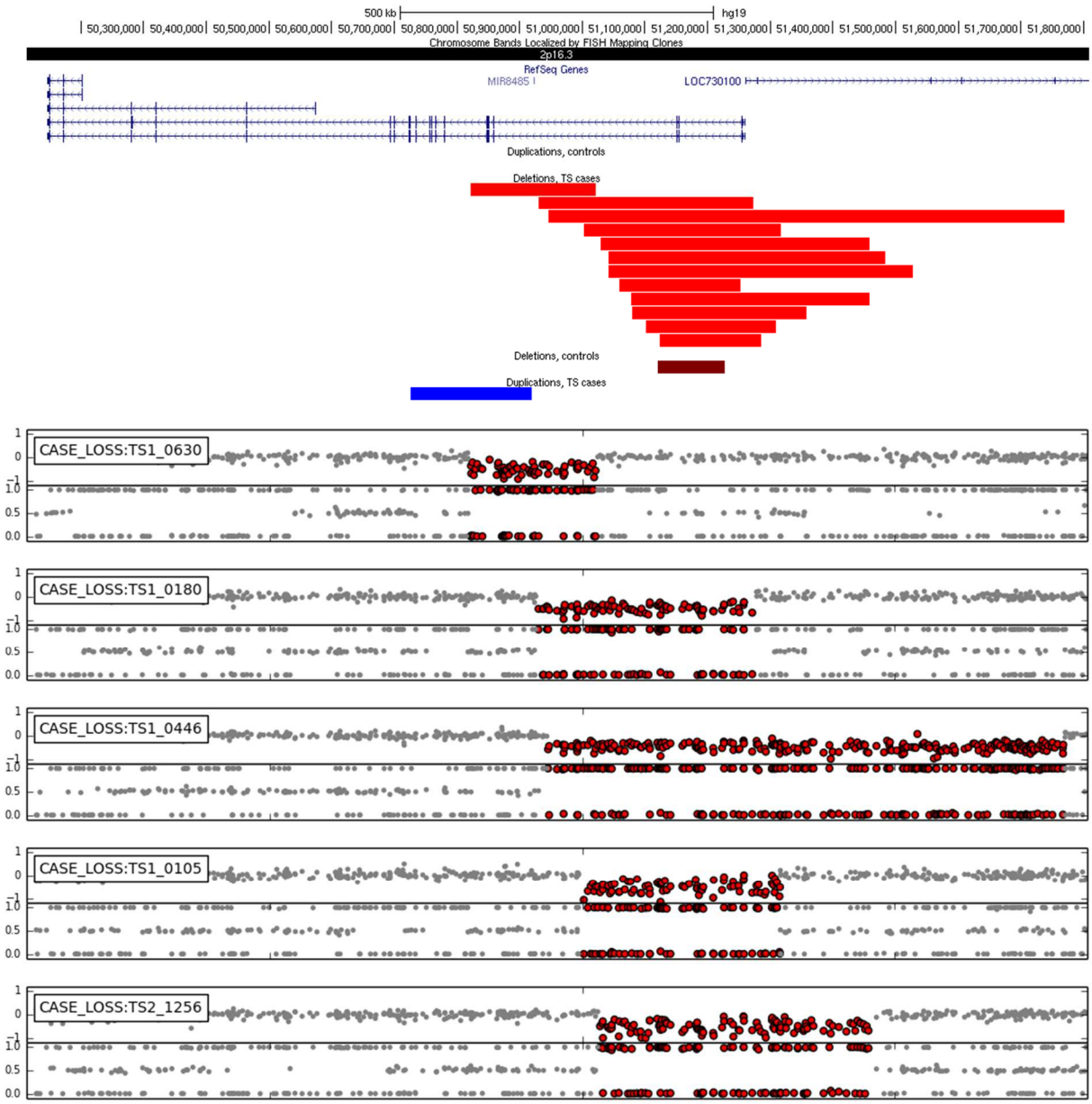


Figure A-5. Exonic CNVs affecting *NRXN1.*

UCSC genome browser track (top) depicting all exonic *NRXN1* CNVs > 30kb identified in this study: 12 heterozygous case deletions (red), one control deletion (dark red) and a single case duplication (blue). Probe-level plots of LRR intensity and BAF for the first five exonic *NRXN1* CNV carriers depicted are shown beneath in the same order. Colored probes (red) indicate the location of called deletions (red).

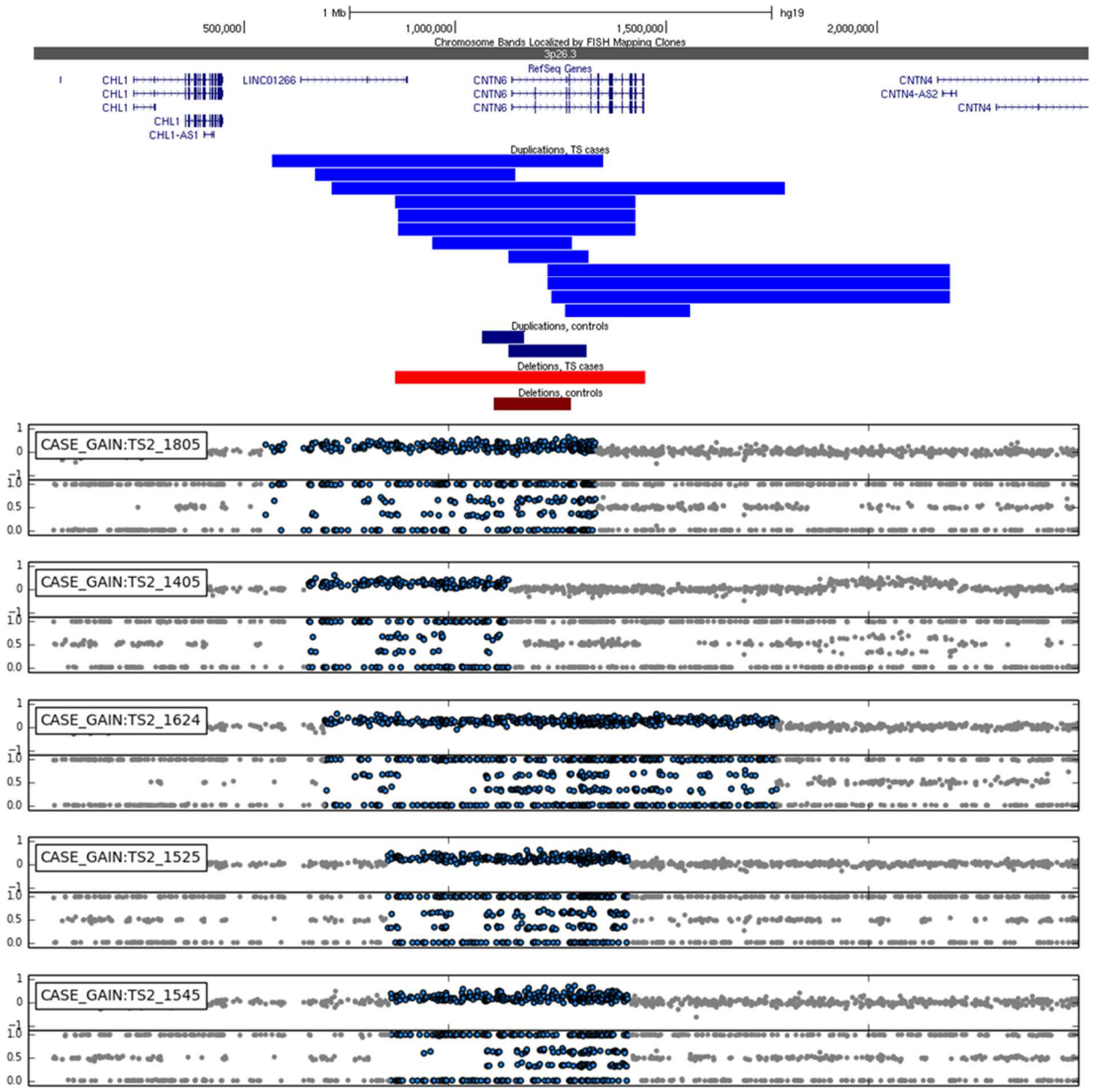


Figure A-6. Exonic CNVs overlapping *CNTN6*.

UCSC genome browser track (top) displaying heterozygous genic duplications in TS cases (blue) and controls (dark blue) followed by deletions (red). Probe-level LRR and BAF plots for the first five CNVs portrayed in the browser track spanning *CNTN6* are shown below in the same order. Colored probes (blue) indicate the location of called duplications.

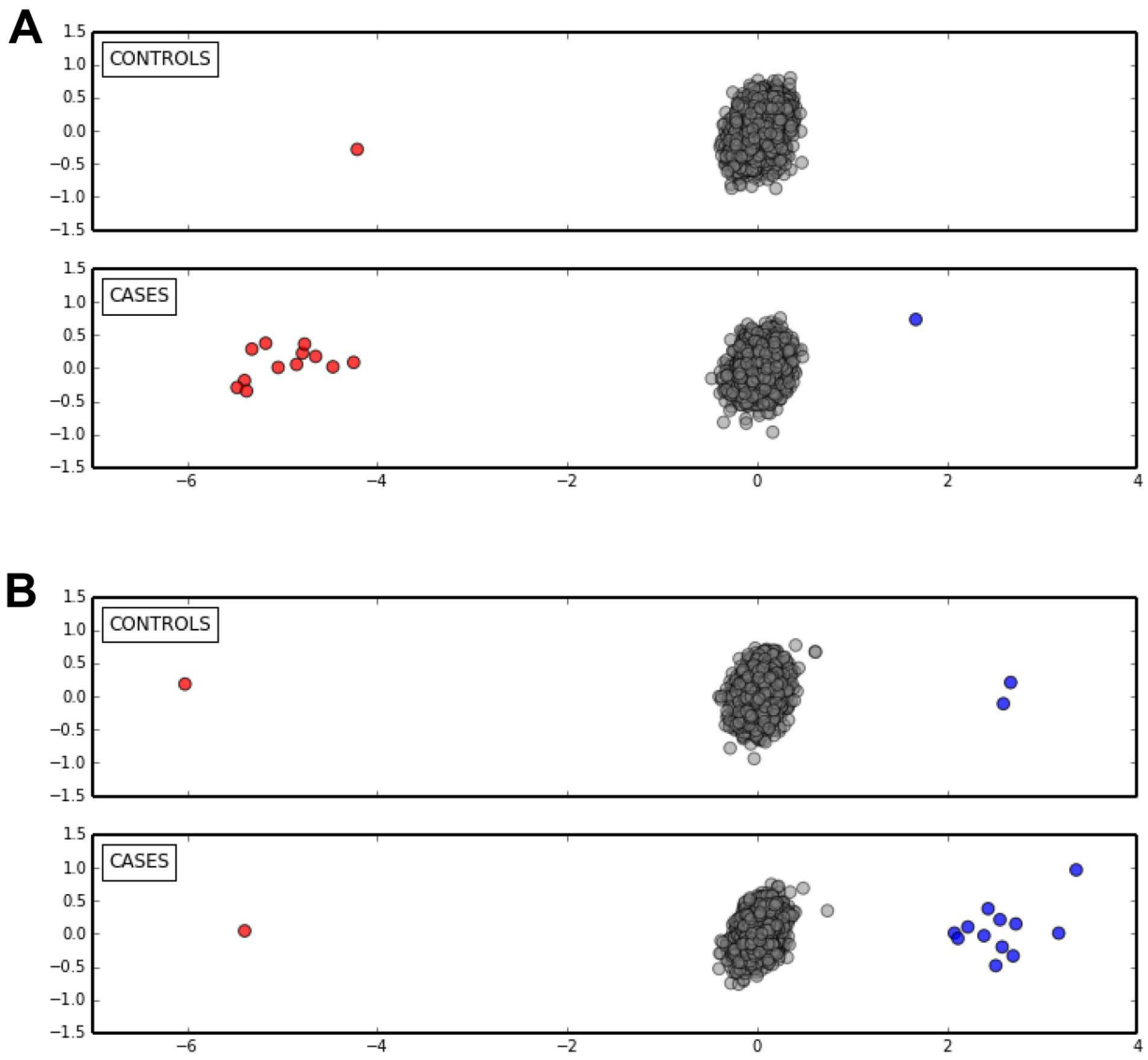


Figure A-7. Confirmation CNV calls and locus sensitivity using intensity-based outlier detection. For each CNV locus, the median summarized normalized probe intensities across a CNV call is plotted against the median value of common of CNV-invariant probes flanking the region. For individuals without a CNV, intensity values are taken from the intersection of detected events. Distinct separation of carriers from non-carriers indicates high sensitivity of CNV calling at both (A) *NRXN1* and (B) *CNTN6*.

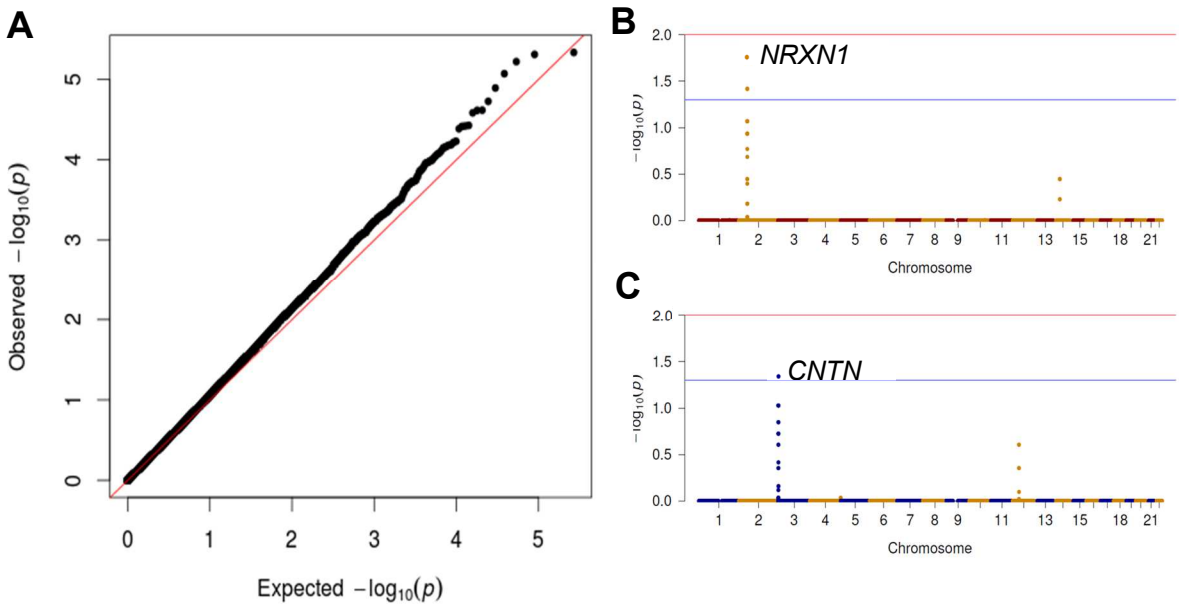


Figure A-1. Examination of CNV association test for population-specific effects.

To verify the robustness of our results to population stratification, we pair-matched each case subject with exactly one control such that the global difference between all pairs is minimized using Gem Tools (Lee et al., 2010). (A) The SNP-based λgc of the resultant dataset (1996 cases and 1996 controls) was an acceptable 1.082. Manhattan plots of segmental association results demonstrate that (B) deletions in NRXN1 and (C) duplications in CNTN6 are significant with an $\alpha < 0.05$ (blue line). Deletions and duplications were analyzed separately. The -log10 (p-value) displayed is empirically corrected for FWER genome-wide using the max(T) method with 1,000,000 permutations.

Appendix B: Supplementary Material for Chapter 5

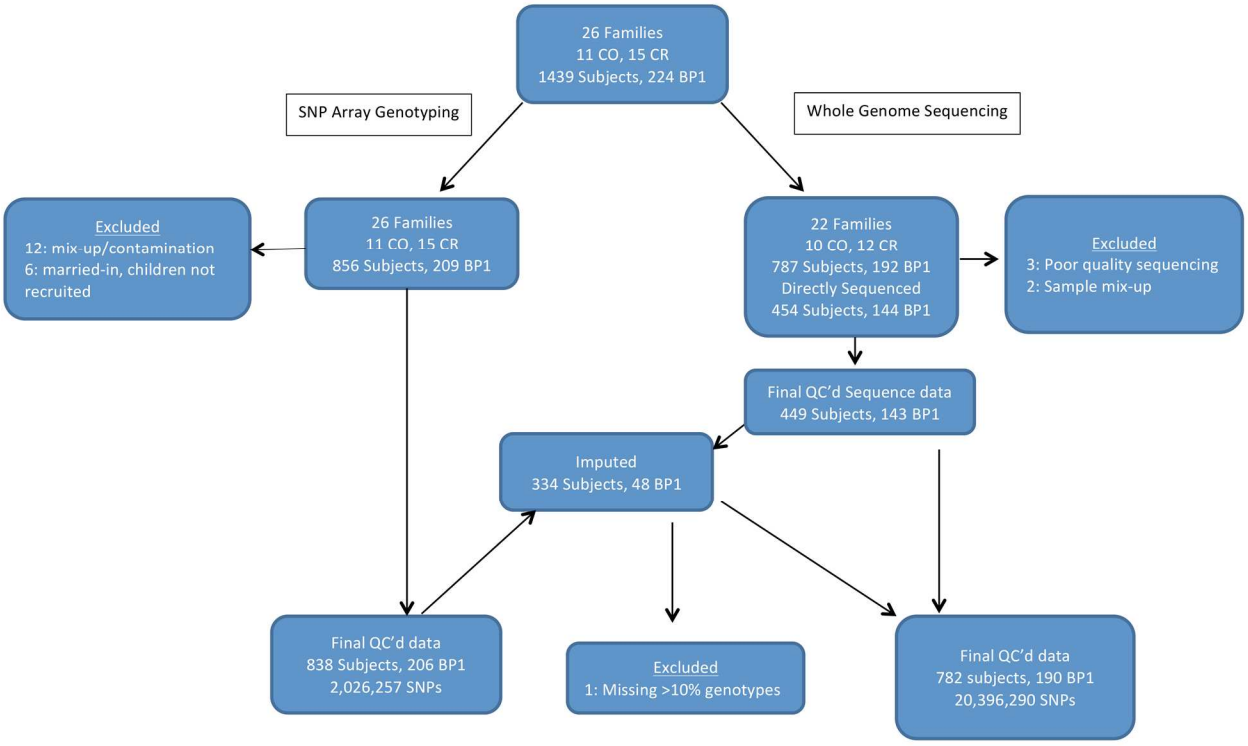


Figure B-1. Overview of genetic data and quality control for samples used in this study.

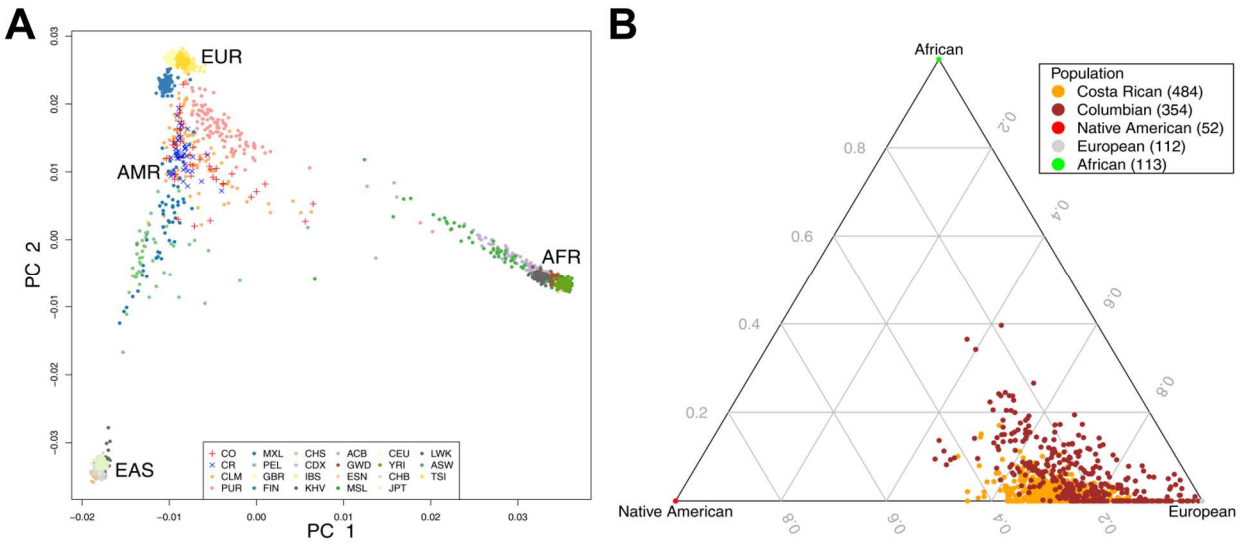


Figure B-2. Ancestry analysis of BP1 subjects.

Plot of PC1 and PC2 for CO/CR pedigree members against HapMap groups. EUR: Europeans; AFR: Africans; EAS: East Asians; AMR: populations from the Americas. (B) Admixture of CO/CR pedigree members across the continental subgroups outlined in (A).

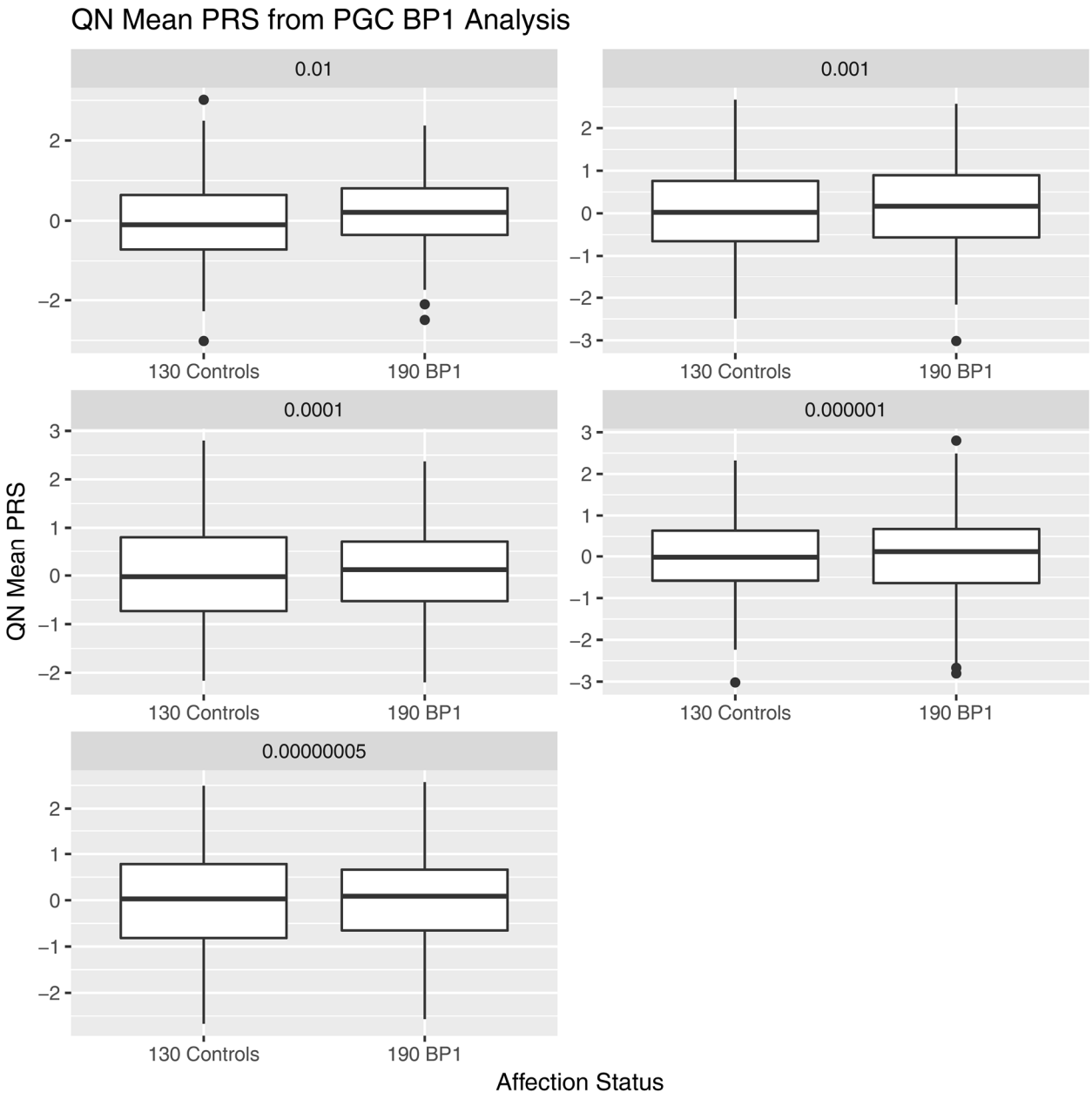


Figure B-3. PRS analysis in BP1 pedigrees.

Comparison of PRS estimated from PGC BP1 GWAS summary statistics in BP1 subjects and controls. PRS is computed at different GWAS p-value thresholds of the PGC BP1 GWAS.

.

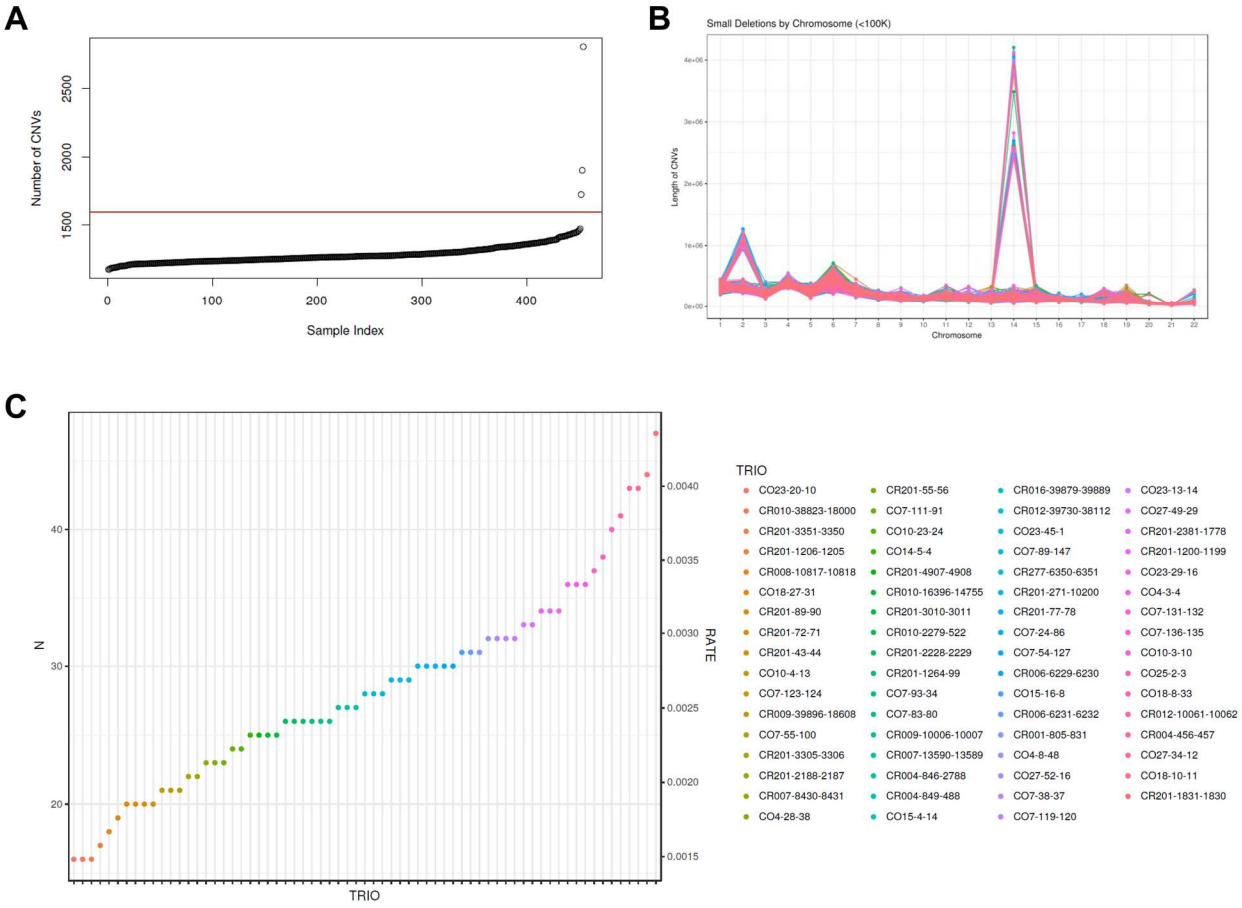


Figure B-4. Quality control for CNVs detected by WGS.

Total number of CNVs detected by individual for all samples used for the CNV discovery phase in GenomeSTRiP. Red line indicates threshold for outlier removal. (B) Chromosome level plot of total CNV load by sample after outlier removal in (A). Peaks on chromosomes 2 and 14 correspond to CNVs found in VDJ recombination regions. (C) Mendelian error rate across all genotyped sites for 67 complete trios present in our dataset.

Appendix C: Supplementary Material for Chapter 6

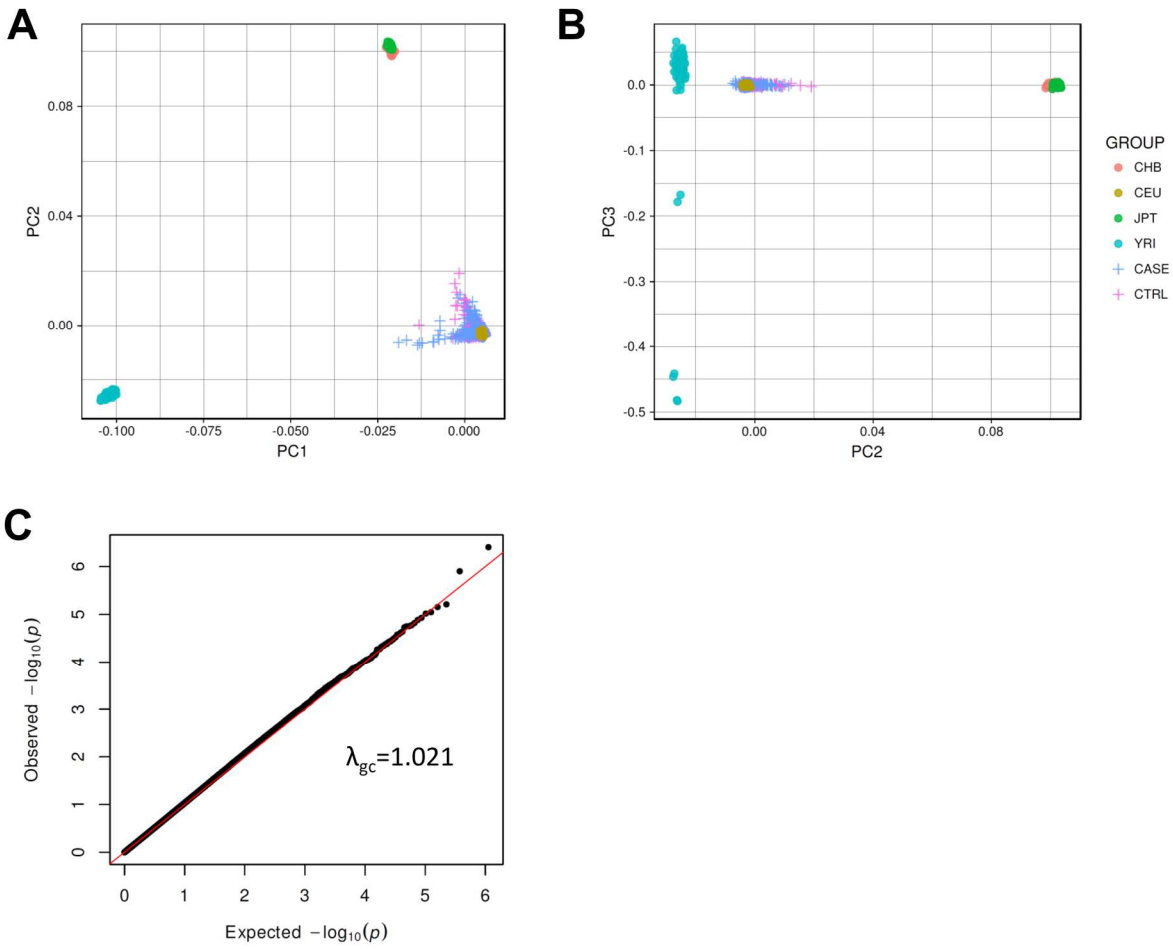


Figure C-1. Population QC of BP samples.

Plots of selected BP samples and controls alongside continental HapMap samples on (A) PC1 and PC2 and (B) PC2 and PC3. (C) Case:control association testing of selected samples on all available autosomal SNPs demonstrates minimal evidence of population stratification (λ_{gc} = 1.021).

	SINGLETON CNV RATE			PROPORTION OF CARRIERS		
CNV TYPE	BIP (n=1,353)	CTRL (n=1,013)	P-value	BIP	CTRL	P-value
100-500kb	330 (0.244)	263 (0.260)	0.471	0.433	0.428	0.335
DEL	181 (0.134)	139 (0.137)	0.870	0.123	0.123	1.000
DUP	200 (0.315)	160 (0.311)	0.590	0.260	0.260	0.462
>500kb	167 (0.123)	109 (0.108)	0.982	0.116	0.101	0.501
DEL	44 (0.033)	27 (0.027)	0.487	0.031	0.027	0.541
DUP	123 (0.091)	82 (0.081)	0.228	0.088	0.078	0.238
ALL CNVs	334 (0.247)	261 (0.258)	0.632	0.208	0.223	0.389
DEL	190 (0.140)	136 (0.134)	0.704	0.128	0.121	0.658
DUP	225 (0.166)	174 (0.172)	0.774	0.143	0.155	0.433

Table C-1. Global burden analysis of the rate and proportion of singleton CNVs.

Burden analysis comparing the total number of singleton CNVs and average singleton CNV rate (in parenthesis), and also in terms of the proportion of singleton CNV carriers, stratified by CNV type and event size.

Gene Set	Case Rate	Ctrl Rate	Case/Ctrl Ratio	P-value	Case Proportion	Ctrl Proportion	Case/Ctrl Ratio	P-value
all protein coding	1.762	1.355	1.300	**0.0032**	0.371	0.334	1.112	**0.0132**
unconstrained (PLI < 0.9)	1.368	1.076	1.271	**0.0086**	0.324	0.306	1.058	0.1654
brain expressed	0.934	0.703	1.328	**0.0080**	0.285	0.266	1.076	0.1322
constrained (PLI > 0.9)	0.273	0.203	1.341	**0.0062**	0.160	0.132	1.206	**0.0224**
brain expressed + constrained	0.235	0.159	1.479	**8.4e-4**	0.145	0.109	1.334	**0.0028**

Table C-2. Genic burden analysis of rare CNVs.

Genic burden analysis comparing the average number genes within each gene set affected and the proportion of such CNV carriers. This analysis is similar to the regression analysis performed in Figure 6.2, except by using permutation. P-values determined empirically by performing 100,000 label-swapping permutations, controlling for subject sex.

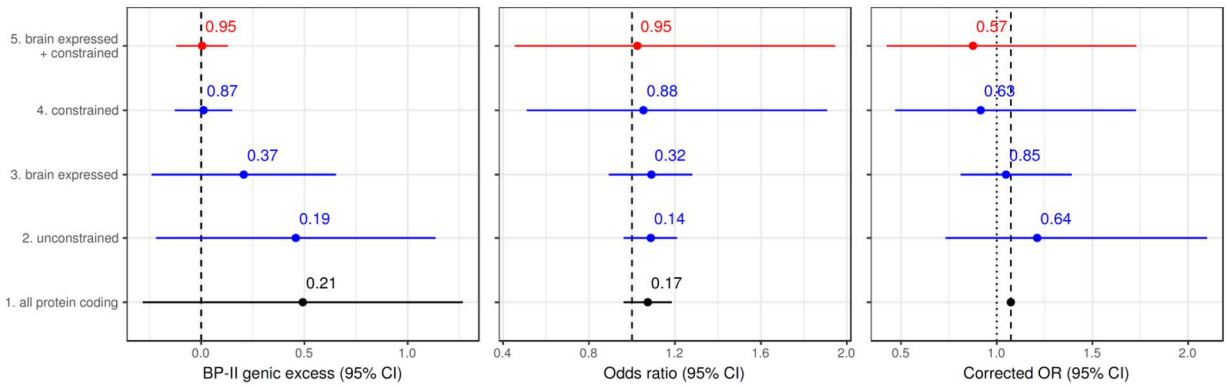


Figure C-2. Genic burden of rare CNVs among sets of protein coding genes for BP2.

Genic burden analysis comparing the number of genes affected by rare CNVs between all BP2 cases and controls for different sets of protein coding genes. Excess of affected genes (left) and odds ratio (middle) calculated using linear regression and logistic regression, respectively, controlling for subject sex, ancestry, and genome-wide rate of all rare CNVs. Corrected OR (right) shows odds ratios calculated with including the total number of affected protein coding genes as a covariate, and represents a conservative estimate of the enrichment of each subset of protein coding genes beyond the genome-wide excess that is observed in BP2.

References

Abelson, J.F., Kwan, K.Y., O'Roak, B.J., Baek, D.Y., Stillman, A.A., Morgan, T.M., Mathews, C.A., Pauls, D.L., Rasin, M.R., Gunel, M.*, et al.* (2005). Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science *310*, 317-320.

Abney, M. (2008). Identity-by-descent estimation and mapping of qualitative traits in large, complex pedigrees. Genetics *179*, 1577-1590.

Abrams, D.J., Rojas, D.C., and Arciniegas, D.B. (2008). Is schizoaffective disorder a distinct categorical diagnosis? A critical review of the literature. Neuropsychiatr Dis Treat *4*, 1089-1109.

Abreu, P.C., Hodge, S.E., and Greenberg, D.A. (2002). Quantification of type I error probabilities for heterogeneity LOD scores. Genet Epidemiol *22*, 156-169.

Adhikari, K., Mendoza-Revilla, J., Chacon-Duque, J.C., Fuentes-Guajardo, M., and Ruiz-Linares, A. (2016). Admixture in Latin America. Curr Opin Genet Dev *41*, 106-114.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods *7*, 248-249.

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res *19*, 1655-1664.

Alexander, G.E., DeLong, M.R., and Strick, P.L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci *9*, 357-381.

Ament, S.A., Szelinger, S., Glusman, G., Ashworth, J., Hou, L., Akula, N., Shekhtman, T., Badner, J.A., Brunkow, M.E., Mauldin, D.E.*, et al.* (2015). Rare variants in neuronal excitability genes influence risk for bipolar disorder. Proc Natl Acad Sci U S A *112*, 3576-3581.

Angelika, F., Howard, M., Kornberg, Z., and others (2012). Familial cortical myoclonus with a mutation in NOL3. Annals of.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet *12*, 745-755.

Barcia, J.J. (2007). The Giemsa stain: its history and applications. International journal of surgical pathology *15*, 292-296.

Barnett, J.H., and Smoller, J.W. (2009). The genetics of bipolar disorder. Neuroscience *164*, 331-343.

Bassett, A.S., and Scherer, S.W. (2017). Copy Number Variation in Tourette Syndrome. Neuron *94*, 1041-1043.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R.*, et al.* (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol *28*, 1045-1048.

Bilder, R.M., Sabb, F.W., Cannon, T.D., London, E.D., Jentsch, J.D., Parker, D.S., Poldrack, R.A., Evans, C., and Freimer, N.B. (2009). Phenomics: the systematic study of phenotypes on a genome-wide scale. Neuroscience *164*, 30-42.

Boehnke, M. (1991). Allele frequency estimation from data on relatives. Am J Hum Genet *48*, 22-25.

Browne, H.A., Hansen, S.N., Buxbaum, J.D., Gair, S.L., Nissen, J.B., Nikolajsen, K.H., Schendel, D.E., Reichenberg, A., Parner, E.T., and Grice, D.E. (2015). Familial clustering of tic disorders and obsessive-compulsive disorder. JAMA Psychiatry *72*, 359-366.

Burd, L., Li, Q., Kerbeshian, J., Klug, M.G., and Freeman, R.D. (2009). Tourette syndrome and comorbid pervasive developmental disorders. J Child Neurol *24*, 170-175.

Bureau, A., Younkin, S.G., Parker, M.M., Bailey-Wilson, J.E., Marazita, M.L., Murray, J.C., Mangold, E., Albacha-Hejazi, H., Beaty, T.H., and Ruczinski, I. (2014). Inferring rare disease risk variants based on exact probabilities of sharing by multiple affected relatives. Bioinformatics *30*, 2189-2196.

Casey, B.J., Galván, A., and Somerville, L.H. (2016). Beyond simple models of adolescence to an integrated circuit-based account: A commentary. Dev Cogn Neurosci *17*, 128-130.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Celedon, J.C*., et al.* (2016). Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. Am J Hum Genet *98*, 653-666.

Cheung, C.Y., Thompson, E.A., and Wijsman, E.M. (2013). GIGI: an approach to effective imputation of dense genotypes on large pedigrees. Am J Hum Genet *92*, 504-516.

Ching, M.S.L., Shen, Y., Tan, W.-H., Jeste, S.S., Morrow, E.M., Chen, X., Mukaddes, N.M., Yoo, S.-Y., Hanson, E., Hundley, R*., et al.* (2010). Deletions of NRXN1 (neurexin-1) predispose to a wide spectrum of developmental disorders. Am J Med Genet B Neuropsychiatr Genet *153B*, 937-947.

Colella, S., Yau, C., Taylor, J.M., Mirza, G., Butler, H., Clouston, P., Bassett, A.S., Seller, A., Holmes, C.C., and Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res *35*, 2013-2025.

Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V*., et al.* (2011). A copy number variation morbidity map of developmental delay. Nat Genet *43*, 838-846.

Craddock, N., O'Donovan, M.C., and Owen, M.J. (2009). Psychosis genetics: modeling the relationship between schizophrenia, bipolar disorder, and mixed (or "schizoaffective") psychoses. Schizophr Bull *35*, 482-490.

Cross-Disorder Group of the Psychiatric Genomics, C. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet *381*, 1371-1379.

Cross-Disorder Group of the Psychiatric Genomics, C., Lee, S.H., Ripke, S., Neale, B.M., Faraone, S.V., Purcell, S.M., Perlis, R.H., Mowry, B.J., Thapar, A., Goddard, M.E.*, et al.* (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet *45*, 984-994.

Cruceanu, C., Ambalavanan, A., Spiegelman, D., Gauthier, J., Lafreniere, R.G., Dion, P.A., Alda, M., Turecki, G., and Rouleau, G.A. (2013). Family-based exome-sequencing approach identifies rare susceptibility variants for lithium-responsive bipolar disorder. Genome *56*, 634-640.

Cui, X.-Y., Hu, Q.-D., Tekaya, M., Shimoda, Y., Ang, B.-T., Nie, D.-Y., Sun, L., Hu, W.-P., Karsak, M., Duka, T.*, et al.* (2004). NB-3/Notch1 pathway via Deltex1 promotes neural progenitor cell differentiation into oligodendrocytes. J Biol Chem *279*, 25858-25865.

Darrow, S.M., Illmann, C., Gauvin, C., Osiecki, L., Egan, C.A., Greenberg, E., Eckfield, M., Hirschtritt, M.E., Pauls, D.L., Batterson, J.R.*, et al.* (2015). Web-based phenotyping for Tourette Syndrome: Reliability of common co-morbid diagnoses. Psychiatry Res *228*, 816-825.

Davis, L.K., Yu, D., Keenan, C.L., Gamazon, E.R., Konkashbaev, A.I., Derks, E.M., Neale, B.M., Yang, J., Lee, S.H., Evans, P.*, et al.* (2013). Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. PLoS Genet *9*, e1003864.

de Wit, J., and Ghosh, A. (2016). Specification of synaptic connectivity by cell surface interactions. Nat Rev Neurosci *17*, 22-35.

Delorme, C., Salvador, A., Valabrègue, R., Roze, E., Palminteri, S., Vidailhet, M., de Wit, S., Robbins, T., Hartmann, A., and Worbe, Y. (2016). Enhanced habit formation in Gilles de la Tourette syndrome. Brain *139*, 605-615.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M.*, et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491-498.

Diskin, S.J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J.M., and Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Research *36*, e126-e126.

Draganski, B., Martino, D., Cavanna, A.E., Hutton, C., Orth, M., Robertson, M.M., Critchley, H.D., and Frackowiak, R.S. (2010). Multispectral brain morphometry in Tourette syndrome persisting into adulthood. Brain *133*, 3661-3675.

Draper, A., Stephenson, M.C., Jackson, G.M., Pépés, S., Morgan, P.S., Morris, P.G., and Jackson, S.R. (2014). Increased GABA contributes to enhanced control over motor excitability in Tourette syndrome. Curr Biol *24*, 2343-2347.

Esmaeeli-Nieh, S., Fenckova, M., Porter, I.M., Motazacker, M.M., Nijhof, B., Castells-Nobau, A., Asztalos, Z., Weißmann, R., Behjati, F., Tzschach, A.*, et al.* (2016). BOD1 Is Required for Cognitive Function in Humans and Drosophila. PLOS Genetics *12*, e1006022.

Etherton, M.R., Blaiss, C.A., Powell, C.M., and Südhof, T.C. (2009). Mouse neurexin-1alpha deletion causes correlated electrophysiological and behavioral changes consistent with cognitive impairments. Proc Natl Acad Sci U S A *106*, 17998-18003.

Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2015). PRSice: Polygenic Risk Score software. Bioinformatics *31*, 1466-1468.

Fears, S.C., Service, S.K., Kremeyer, B., Araya, C., Araya, X., Bejarano, J., Ramirez, M., Castrillon, G., Gomez-Franco, J., Lopez, M.C.*, et al.* (2014). Multisystem component phenotypes of bipolar disorder for genetic investigations of extended pedigrees. JAMA Psychiatry *71*, 375-387.

Fernandez, T.V., Sanders, S.J., Yurkiewicz, I.R., Ercan-Sencicek, A.G., Kim, Y.-S., Fishman, D.O., Raubeson, M.J., Song, Y., Yasuno, K., Ho, W.S.C.*, et al.* (2012). Rare copy number variants in tourette syndrome disrupt genes in histaminergic pathways and overlap with autism. Biol Psychiatry *71*, 392-402.

Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K.*, et al.* (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet *47*, 1228-1235.

Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192-195.

Flaherty, A.W., and Graybiel, A.M. (1991). Corticostriatal transformations in the primate somatosensory system. Projections from physiologically mapped body-part representations. J Neurophysiol *66*, 1249-1263.

Fromer, M., Moran, Jennifer L., Chambert, K., Banks, E., Bergen, Sarah E., Ruderfer, Douglas M., Handsaker, Robert E., McCarroll, Steven A., O'Donovan, Michael C., Owen, Michael J.*, et al.* (2012). Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth. American Journal of Human Genetics *91*, 597-607.

Fuccillo, M.V., Földy, C., Gökce, Ö., Rothwell, P.E., Sun, G.L., Malenka, R.C., and Südhof, T.C. (2015). Single-Cell mRNA Profiling Reveals Cell-Type-Specific Expression of Neurexin Isoforms. Neuron *87*, 326-340.

Garner, C., McInnes, L.A., Service, S.K., Spesny, M., Fournier, E., Leon, P., and Freimer, N.B. (2001). Linkage analysis of a complex pedigree with severe bipolar disorder, using a Markov chain Monte Carlo method. Am J Hum Genet *68*, 1061-1064.

Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A.*, et al.* (2015). A global reference for human genetic variation. Nature *526*, 68-74.

Georgi, B., Craig, D., Kember, R.L., Liu, W., Lindquist, I., Nasser, S., Brown, C., Egeland, J.A., Paul, S.M., and Bucan, M. (2014). Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate. PLoS Genet *10*, e1004229.

Gershon, E.S., Hamovit, J., Guroff, J.J., Dibble, E., Leckman, J.F., Sceery, W., Targum, S.D., Nurnberger, J.I., Jr., Goldin, L.R., and Bunney, W.E., Jr. (1982). A family study of schizoaffective, bipolar I, bipolar II, unipolar, and normal control probands. Arch Gen Psychiatry *39*, 1157-1167.

Gilbert, D.L., Bansal, A.S., Sethuraman, G., Sallee, F.R., Zhang, J., Lipps, T., and Wassermann, E.M. (2004). Association of cortical disinhibition with tic, ADHD, and OCD severity in Tourette syndrome. Mov Disord *19*, 416-425.

Gilissen, C., Hoischen, A., Brunner, H.G., and Veltman, J.A. (2011). Unlocking Mendelian disease using exome sequencing. Genome Biol *12*, 228.

Girirajan, S., Brkanac, Z., Coe, B.P., Baker, C., Vives, L., Vu, T.H., Shafer, N., Bernier, R., Ferrero, G.B., Silengo, M.*, et al.* (2011). Relative Burden of Large CNVs on a Range of Neurodevelopmental Phenotypes. PLoS Genet *7*, e1002334.

Glessner, J.T., Wang, K., Cai, G., Korvatska, O., Kim, C.E., Wood, S., Zhang, H., Estes, A., Brune, C.W., Bradfield, J.P.*, et al.* (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature *459*, 569-573.

Goes, F.S., Pirooznia, M., Parla, J.S., Kramer, M., Ghiban, E., Mavruk, S., Chen, Y.C., Monson, E.T., Willour, V.L., Karchin, R.*, et al.* (2016). Exome Sequencing of Familial Bipolar Disorder. JAMA Psychiatry *73*, 590-597.

Graybiel, A.M. (2008). Habits, rituals, and the evaluative brain. Annu Rev Neurosci *31*, 359-387.

Green, E.K., Grozeva, D., Jones, I., Jones, L., Kirov, G., Caesar, S., Gordon-Smith, K., Fraser, C., Forty, L., Russell, E.*, et al.* (2010). The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. Mol Psychiatry *15*, 1016-1022.

Green, E.K., Rees, E., Walters, J.T., Smith, K.G., Forty, L., Grozeva, D., Moran, J.L., Sklar, P., Ripke, S., Chambert, K.D.*, et al.* (2016). Copy number variation in bipolar disorder. Mol Psychiatry *21*, 89-93.

Greene, D.J., Williams, A.C.I., Koller, J.M., Schlaggar, B.L., and Black, K.J. (2016). Brain structure in pediatric Tourette syndrome. Mol Psychiatry.

Grozeva, D., Kirov, G., Conrad, D.F., Barnes, C.P., Hurles, M., Owen, M.J., O'Donovan, M.C., and Craddock, N. (2013). Reduced burden of very large and rare CNVs in bipolar affective disorder. Bipolar Disord *15*, 893-898.

Grozeva, D., Kirov, G., Ivanov, D., Jones, I.R., Jones, L., Green, E.K., St Clair, D.M., Young, A.H., Ferrier, N., Farmer, A.E.*, et al.* (2010). Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. Arch Gen Psychiatry *67*, 318-327.

Gu, Z., Kalambogias, J., Yoshioka, S., Han, W., Li, Z., Kawasawa, Y.I., Pochareddy, S., Li, Z., Liu, F., Xu, X.*, et al.* (2017). Control of species-dependent cortico-motoneuronal connections underlying manual dexterity. Science *357*, 400.

Gunduz, A., Morita, H., Rossi, P.J., Allen, W.L., Alterman, R.L., Bronte-Stewart, H., Butson, C.R., Charles, D., Deckers, S., de Hemptinne, C.*, et al.* (2015). Proceedings of the Second Annual Deep Brain Stimulation Think Tank: What's in the Pipeline. Int J Neurosci *125*, 475-485.

Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. Genome Res *22*, 1154-1162.

Haber, S.N., and Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. Neuropsychopharmacology *35*, 4-26.

Haldane, J.B.S. (1935). The rate of spontaneous mutation of a human gene. Journal of Genetics *31*, 317.

Han, J., Walters, J.T.R., Kirov, G., Pocklington, A., Escott-Price, V., Owen, M.J., Holmans, P., O'Donovan, M.C., and Rees, E. (2016). Gender differences in CNV burden do not confound schizophrenia CNV associations. Sci Rep *6*, 25986.

Handsaker, R.E., Korn, J.M., Nemesh, J., and McCarroll, S.A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet *43*, 269-276.

Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., Kashin, S., Boettger, L.M., and McCarroll, S.A. (2015). Large multiallelic copy number variations in humans. Nat Genet *47*, 296-303.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D.*, et al.* (2006). GENCODE: producing a reference annotation for ENCODE. Genome Biol *7 Suppl 1*, S4 1-9.

Hassler, R., and Dieckmann, G. (1970). [Stereotaxic treatment of tics and inarticulate cries or coprolalia considered as motor obsessional phenomena in Gilles de la Tourette's disease]. Rev Neurol *123*, 89-100.

Helbig, I., Mefford, H.C., Sharp, A.J., Guipponi, M., Fichera, M., Franke, A., Muhle, H., de Kovel, C., Baker, C., von Spiczak, S.*, et al.* (2009). 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat Genet *41*, 160-162.

Herzberg, I., Jasinska, A., Garcia, J., Jawaheer, D., Service, S., Kremeyer, B., Duque, C., Parra, M.V., Vega, J., Ortiz, D.*, et al.* (2006). Convergent linkage evidence from two Latin-American population isolates supports the presence of a susceptibility locus for bipolar disorder in 5q31-34. Hum Mol Genet *15*, 3146-3153.

Hirschtritt, M.E., Darrow, S.M., Illmann, C., Osiecki, L., Grados, M., Sandor, P., Dion, Y., King, R.A., Pauls, D.L., Budman, C.L.*, et al.* (2016). Social disinhibition is a heritable subphenotype of tics in Tourette syndrome. Neurology *87*, 497-504.

Hirschtritt, M.E., Lee, P.C., Pauls, D.L., Dion, Y., Grados, M.A., Illmann, C., King, R.A., Sandor, P., McMahon, W.M., Lyon, G.J.*, et al.* (2015). Lifetime prevalence, age of risk, and genetic relationships of comorbid psychiatric disorders in Tourette syndrome. JAMA Psychiatry *72*, 325-333.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934-947.

Hu, J., Liao, J., Sathanoori, M., Kochmar, S., Sebastian, J., Yatsenko, S.A., and Surti, U. (2015). CNTN6 copy number variations in 14 patients: a possible candidate gene for neurodevelopmental and neuropsychiatric disorders. J Neurodev Disord *7*, 26.

Huang, A.Y., Yu, D., Davis, L.K., Sul, J.H., Tsetsos, F., Ramensky, V., Zelaya, I., Ramos, E.M., Osiecki, L., Chen, J.A.*, et al.* (2017). Rare Copy Number Variants in NRXN1 and CNTN6 Increase Risk for Tourette Syndrome. Neuron *94*, 1101-1111.e1107.

Huang, Z., Yu, Y., Shimoda, Y., Watanabe, K., and Liu, Y. (2012). Loss of neural recognition molecule NB-3 delays the normal projection and terminal branching of developing corticospinal tract axons in the mouse. J Comp Neurol *520*, 1227-1245.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. Nat Genet *36*, 949-951.

International HapMap, C. (2003). The International HapMap Project. Nature *426*, 789-796.

International HapMap, C., Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P.*, et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

International Schizophrenia, C. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature *455*, 237-241.

International Schizophrenia, C., Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748-752.

Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet *48*, 214-220.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-H., Narzisi, G., Leotta, A.*, et al.* (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285-299.

Jacquemont, S., Coe, B.P., Hersch, M., Duyzend, M.H., Krumm, N., Bergmann, S., Beckmann, J.S., Rosenfeld, J.A., and Eichler, E.E. (2014). A Higher Mutational Burden in Females Supports a “Female Protective Model” in Neurodevelopmental Disorders. Am J Hum Genet *94*, 415.

Jahanshahi, M., Obeso, I., Rothwell, J.C., and Obeso, J.A. (2015). A fronto-striato-subthalamic-pallidal network for goal-directed and habitual inhibition. Nat Rev Neurosci *16*, 719-732.

Jalbrzikowski, M., Lazaro, M.T., Gao, F., Huang, A., Chow, C., Geschwind, D.H., Coppola, G., and Bearden, C.E. (2015). Transcriptome Profiling of Peripheral Blood in 22q11.2 Deletion Syndrome Reveals Functional Pathways Related to Psychosis and Autism Spectrum Disorder. PLoS One *10*, e0132542.

Jog, M.S., Kubota, Y., Connolly, C.I., Hillegaart, V., and Graybiel, A.M. (1999). Building neural representations of habits. Science *286*, 1745-1749.

Jongsma, H.E., Gayer-Anderson, C., Lasalvia, A., Quattrone, D., Mulè, A., Szöke, A., Selten, J.-P., Turner, C., Arango, C., Tarricone, I.*, et al.* (2018). Treated Incidence of Psychotic Disorders in the Multinational EU-GEI Study. JAMA Psychiatry *75*, 36-46.

Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science *258*, 818.

Kashevarova, A.A., Nazarenko, L.P., Schultz-Pedersen, S., Skryabin, N.A., Salyukova, O.A., Chechetkina, N.N., Tolmacheva, E.N., Rudko, A.A., Magini, P., Graziano, C.*, et al.* (2014). Single gene microdeletions and microduplication of 3p26.3 in three unrelated families: CNTN6 as a new candidate gene for intellectual disability. Mol Cytogenet *7*, 97.

Kataoka, Y., Kalanithi, P.S.A., Grantz, H., Schwartz, M.L., Saper, C., Leckman, J.F., and Vaccarino, F.M. (2010). Decreased number of parvalbumin and cholinergic interneurons in the striatum of individuals with Tourette syndrome. J Comp Neurol *518*, 277-291.

Kearney, H.M., Thorland, E.C., Brown, K.K., Quintero-Rivera, F., South, S.T., and Working Group of the American College of Medical Genetics Laboratory Quality Assurance, C. (2011). American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genet Med *13*, 680-685.

Kelly, B.J., Fitch, J.R., Hu, Y., Corsmeier, D.J., Zhong, H., Wetzel, A.N., Nordquist, R.D., Newsom, D.L., and White, P. (2015). Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. Genome Biol *16*, 6.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet *46*, 310-315.

Kirov, G. (2015). CNVs in neuropsychiatric disorders. Hum Mol Genet *24*, R45-49.

Kirov, G., Pocklington, A.J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., Moran, J., Chambert, K., Toncheva, D., Georgieva, L.*, et al.* (2012). De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. Mol Psychiatry *17*, 142-153.

Knowlton, B.J., Mangels, J.A., and Squire, L.R. (1996). A neostriatal habit learning system in humans. Science *273*, 1399-1402.

Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet *11*, 241-247.

Lange, K., Papp, J.C., Sinsheimer, J.S., Sripracha, R., Zhou, H., and Sobel, E.M. (2013). Mendel: the Swiss army knife of genetic analysis programs. Bioinformatics *29*, 1568-1570.

Lee, A.B., Luca, D., Klei, L., Devlin, B., and Roeder, K. (2010). Discovering genetic ancestry using spectral graph theory. Genet Epidemiol *34*, 51-59.

Lee, S., Takeda, Y., Kawano, H., Hosoya, H., Nomoto, M., Fujimoto, D., Takahashi, N., and Watanabe, K. (2000). Expression and regulation of a gene encoding neural recognition molecule NB-3 of the contactin/F3 subgroup in mouse brain. Gene *245*, 253-266.

Lejeune, J., Gautier, M., and Turpin, R. (1959). [Study of somatic chromosomes from 9 mongoloid children]. Comptes rendus hebdomadaires des seances de l'Academie des sciences *248*, 1721-1722.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B.*, et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285-291.

Leppa, Virpi M., Kravitz, Stephanie N., Martin, Christa L., Andrieux, J., Le Caignec, C., Martin-Coignard, D., DyBuncio, C., Sanders, Stephan J., Lowe, Jennifer K., Cantor, Rita M.*, et al.* (2016). Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families. American Journal of Human Genetics *99*, 540-554.

Levy, D., Ronemus, M., Yamrom, B., Lee, Y.-H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K.*, et al.* (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron *70*, 886-897.

Li, B., Chen, W., Zhan, X., Busonero, F., Sanna, S., Sidore, C., Cucca, F., Kang, H.M., and Abecasis, G.R. (2012). A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet *8*, e1002944.

Lin, H., Katsovich, L., Ghebremichael, M., Findley, D.B., Grantz, H., Lombroso, P.J., King, R.A., Zhang, H., and Leckman, J.F. (2007). Psychosocial stress predicts future symptom severities in children and adolescents with Tourette syndrome and/or obsessive-compulsive disorder. J Child Psychol Psychiatry *48*, 157-166.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B.*, et al.* (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823-828.

Maier, W., Lichtermann, D., Minges, J., Hallmayer, J., Heun, R., Benkert, O., and Levinson, D.F. (1993). Continuity and discontinuity of affective disorders and schizophrenia. Results of a controlled family study. Arch Gen Psychiatry *50*, 871-883.

Malhotra, D., McCarthy, S., Michaelson, J.J., Vacic, V., Burdick, K.E., Yoon, S., Cichon, S., Corvin, A., Gary, S., Gershon, E.S.*, et al.* (2011). High frequencies of de novo CNVs in bipolar disorder and schizophrenia. Neuron *72*, 951-963.

Malhotra, D., and Sebat, J. (2012). CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell *148*, 1223-1241.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A.*, et al.* (2009). Finding the missing heritability of complex diseases. Nature *461*, 747-753.

Marsh, R., Maia, T.V., and Peterson, B.S. (2009). Functional disturbances within frontostriatal circuits across multiple childhood psychopathologies. Am J Psychiatry *166*, 664-674.

Marshall, C.R., Howrigan, D.P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D.S., Antaki, D., Shetty, A., Holmans, P.A., Pinto, D.*, et al.* (2016). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nature Genetics *49*, 27.

Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y.*, et al.* (2008). Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet *82*, 477-488.

Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet *100*, 635-649.

Mataix-Cols, D., Isomura, K., Pérez-Vigil, A., Chang, Z., Rück, C., Larsson, K.J., Leckman, J.F., Serlachius, E., Larsson, H., and Lichtenstein, P. (2015). Familial Risks of Tourette Syndrome and Chronic Tic Disorders. A Population-Based Cohort Study. JAMA Psychiatry *72*, 787-793.

Matise, T.C., Chen, F., Chen, W., De La Vega, F.M., Hansen, M., He, C., Hyland, F.C., Kennedy, G.C., Kong, X., Murray, S.S.*, et al.* (2007). A second-generation combined linkage physical map of the human genome. Genome Res *17*, 1783-1786.

McCarthy, S.E., Makarov, V., Kirov, G., Addington, A.M., McClellan, J., Yoon, S., Perkins, D.O., Dickel, D.E., Kusenda, M., Krastoshevsky, O.*, et al.* (2009). Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet *41*, 1223-1227.

McGrath, L.M., Yu, D., Marshall, C., Davis, L.K., Thiruvahindrapuram, B., Li, B., Cappi, C., Gerber, G., Wolf, A., Schroeder, F.A.*, et al.* (2014). Copy number variation in obsessive-compulsive disorder and tourette syndrome: a cross-disorder study. J Am Acad Child Adolesc Psychiatry *53*, 910-919.

McInnes, L.A., Escamilla, M.A., Service, S.K., Reus, V.I., Leon, P., Silva, S., Rojas, E., Spesny, M., Baharloo, S., Blankenship, K.*, et al.* (1996). A complete genome screen for genes predisposing to severe bipolar disorder in two Costa Rican pedigrees. Proc Natl Acad Sci U S A *93*, 13060-13065.

McQuillin, A., Bass, N., Anjorin, A., Lawrence, J., Kandaswamy, R., Lydall, G., Moran, J., Sklar, P., Purcell, S., and Gurling, H. (2011). Analysis of genetic deletions and duplications in the University College London bipolar disorder case control sample. Eur J Hum Genet *19*, 588-592.

Mell, L.K., Davis, R.L., and Owens, D. (2005). Association between streptococcal infection and obsessive-compulsive disorder, Tourette's syndrome, and tic disorder. Pediatrics *116*, 56-60.

Mercati, O., Huguet, G., Danckaert, A., André-Leroux, G., Maruani, A., Bellinzoni, M., Rolland, T., Gouder, L., Mathieu, A., Buratti, J.*, et al.* (2016). CNTN6 mutations are risk factors for abnormal auditory sensory perception in autism spectrum disorders. Mol Psychiatry.

Merikangas, K.R., and Low, N.C. (2004). The epidemiology of mood disorders. Curr Psychiatry Rep *6*, 411-421.

Mink, J.W. (2006). Neurobiology of basal ganglia and Tourette syndrome: basal ganglia circuits and thalamocortical outputs. Adv Neurol *99*, 89-98.

Morris, D.W., Pearson, R.D., Cormican, P., Kenny, E.M., O'Dushlaine, C.T., Perreault, L.P., Giannoulatou, E., Tropea, D., Maher, B.S., Wormley, B.*, et al.* (2014). An inherited duplication at the gene p21 Protein-Activated Kinase 7 (PAK7) is a risk factor for psychosis. Hum Mol Genet *23*, 3316-3326.

Motlagh, M.G., Katsovich, L., Thompson, N., Lin, H., Kim, Y.-S., Scahill, L., Lombroso, P.J., King, R.A., Peterson, B.S., and Leckman, J.F. (2010). Severe psychosocial stress and heavy cigarette smoking during pregnancy: an examination of the pre- and perinatal risk factors associated with ADHD and Tourette syndrome. Eur Child Adolesc Psychiatry *19*, 755-764.

Nag, A., Bochukova, E.G., Kremeyer, B., Campbell, D.D., Muller, H., Valencia-Duarte, A.V., Cardona, J., Rivas, I.C., Mesa, S.C., Cuartas, M.*, et al.* (2013). CNV analysis in Tourette syndrome implicates large genomic rearrangements in COL8A1 and NRXN1. PLoS One *8*, e59061.

Nambu, A. (2011). Somatotopic organization of the primate Basal Ganglia. Front Neuroanat *5*, 26.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V.*, et al.* (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242-245.

Nordentoft, M., Thorup, A., Petersen, L., Ohlenschlaeger, J., Melau, M., Christensen, T.Ø., Krarup, G., Jørgensen, P., and Jeppesen, P. (2006). Transition rates from schizotypal disorder to psychotic disorder for first-contact patients included in the OPUS trial. A randomized clinical trial of integrated treatment and standard treatment. Schizophr Res *83*, 29-40.

O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D.*, et al.* (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246-250.

Ogawa, J., Kaneko, H., Masuda, T., Nagata, S., Hosoya, H., and Watanabe, K. (1996). Novel neural adhesion molecules in the Contactin/F3 subgroup of the immunoglobulin superfamily: isolation and characterization of cDNAs from rat brain. Neurosci Lett *218*, 173-176.

Pagani, L., St Clair, P.A., Teshiba, T.M., Service, S.K., Fears, S.C., Araya, C., Araya, X., Bejarano, J., Ramirez, M., Castrillon, G.*, et al.* (2016). Genetic contributions to circadian activity rhythm and sleep pattern phenotypes in pedigrees segregating for severe bipolar disorder. Proc Natl Acad Sci U S A *113*, E754-761.

Pak, C., Danko, T., Zhang, Y., Aoto, J., Anderson, G., Maxeiner, S., Yi, F., Wernig, M., and Südhof, T.C. (2015). Human Neuropsychiatric Disease Modeling using Conditional Deletion Reveals Synaptic Transmission Defects Caused by Heterozygous Mutations in NRXN1. Cell Stem Cell *17*, 316-328.

Palacio, C.A., Garcia, J., Arbelaez, M.P., Sanchez, R., Aguirre, B., Garces, I.C., Montoya, G.J., Gomez, J., Agudelo, A., Lopez, C.A.*, et al.* (2004). [Validation of the Diagnostic Interview for Genetic Studies (DIGS) in Colombia]. Biomedica *24*, 56-62.

Paschou, P. (2013). The genetic basis of Gilles de la Tourette Syndrome. Neurosci Biobehav Rev *37*, 1026-1039.

Paschou, P., Yu, D., Gerber, G., Evans, P., Tsetsos, F., Davis, L.K., Karagiannidis, I., Chaponis, J., Gamazon, E., Mueller-Vahl, K.*, et al.* (2014). Genetic association signal near NTN4 in Tourette syndrome. Ann Neurol *76*, 310-315.

Pato, M.T., Sobell, J.L., Medeiros, H., Abbott, C., Sklar, B.M., Buckley, P.F., Bromet, E.J., Escamilla, M.A., Fanous, A.H., Lehrer, D.S.*, et al.* (2013). The genomic psychiatry cohort: partners in discovery. Am J Med Genet B Neuropsychiatr Genet *162B*, 306-312.

Pauls, D.L., Fernandez, T.V., Mathews, C.A., State, M.W., and Scharf, J.M. (2014). The Inheritance of Tourette Disorder: A review. J Obsessive Compuls Relat Disord *3*, 380-385.

Petrovski, S., Gussow, A.B., Wang, Q., Halvorsen, M., Han, Y., Weir, W.H., Allen, A.S., and Goldstein, D.B. (2015). The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. PLoS Genet *11*, e1005492.

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A.C., Thiruvahindrapuram, B., Macdonald, J.R., Mills, R.*, et al.* (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol *29*, 512-520.

Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S.*, et al.* (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature *466*, 368-372.

Power, C., and Elliott, J. (2006). Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol *35*, 34-41.

Power, R.A., Kyaga, S., Uher, R., MacCabe, J.H., Långström, N., Landen, M., McGuffin, P., Lewis, C.M., Lichtenstein, P., and Svensson, A.C. (2013). Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. JAMA Psychiatry *70*, 22-30.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet *38*, 904-909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J.*, et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet *81*, 559-575.

Qiao, D., Lange, C., Laird, N.M., Won, S., Hersh, C.P., Morrow, J., Hobbs, B.D., Lutz, S.M., Ruczinski, I., Beaty, T.H.*, et al.* (2017). Gene-based segregation method for identifying rare variants in family-based sequencing studies. Genet Epidemiol *41*, 309-319.

Raj, A., Stephens, M., and Pritchard, J.K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics *197*, 573-589.

Raychaudhuri, S., Korn, J.M., McCarroll, S.A., International Schizophrenia, C., Altshuler, D., Sklar, P., Purcell, S., and Daly, M.J. (2010). Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. PLoS Genet *6*, e1001097.

Reich, D., Patterson, N., Campbell, D., Tandon, A., Mazieres, S., Ray, N., Parra, M.V., Rojas, W., Duque, C., Mesa, N.*, et al.* (2012). Reconstructing Native American population history. Nature *488*, 370-374.

Ripatti, P., Ramo, J.T., Soderlund, S., Surakka, I., Matikainen, N., Pirinen, M., Pajukanta, P., Sarin, A.P., Service, S.K., Laurila, P.P.*, et al.* (2016). The Contribution of GWAS Loci in Familial Dyslipidemias. PLoS Genet *12*, e1006078.

Ritter, B.P., Angelo, G.W., Durner, M., Rossy-Fullana, E., Carrion-Baralt, J., Silverman, J.M., and Bespalova, I.N. (2012). Mutation screening of PDZD2, GOLPH3, and MTMR12 genes in patients with schizophrenia. Psychiatr Genet *22*, 51-52.

Robertson, M.M., Eapen, V., Singer, H.S., Martino, D., Scharf, J.M., Paschou, P., Roessner, V., Woods, D.W., Hariz, M., Mathews, C.A.*, et al.* (2017). Gilles de la Tourette syndrome. Nat Rev Dis Primers *3*, 16097.

Romanelli, P., Esposito, V., Schaal, D.W., and Heit, G. (2005). Somatotopy in the basal ganglia: experimental and clinical evidence for segregated sensorimotor channels. Brain Res Brain Res Rev *48*, 112-128.

Ronemus, M., Iossifov, I., Levy, D., and Wigler, M. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. Nat Rev Genet *15*, 133-141.

Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M.*, et al.* (2015). The UCSC Genome Browser database: 2015 update. Nucleic Acids Res *43*, D670-681.

Ross, J., Gedvilaite, E., Badner, J.A., Erdman, C., Baird, L., Matsunami, N., Leppert, M., Xing, J., and Byerley, W. (2016). A Rare Variant in CACNA1D Segregates with 7 Bipolar I Disorder Cases in a Large Pedigree. Mol Neuropsychiatry *2*, 145-150.

Ruderfer, D.M., Hamamsy, T., Lek, M., Karczewski, K.J., Kavanagh, D., Samocha, K.E., Exome Aggregation, C., Daly, M.J., MacArthur, D.G., Fromer, M.*, et al.* (2016). Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. Nat Genet.

Rujescu, D., Ingason, A., Cichon, S., Pietiläinen, O.P.H., Barnes, M.R., Toulopoulou, T., Picchioni, M., Vassos, E., Ettinger, U., Bramon, E.*, et al.* (2009). Disruption of the neurexin 1 gene is associated with schizophrenia. Hum Mol Genet *18*, 988-996.

Sakurai, K., Toyoshima, M., Ueda, H., Matsubara, K., Takeda, Y., Karagogeos, D., Shimoda, Y., and Watanabe, K. (2009). Contribution of the neural cell recognition molecule NB-3 to synapse formation between parallel fibers and Purkinje cells in mouse. Dev Neurobiol *69*, 811-824.

Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A.*, et al.* (2014). A framework for the interpretation of de novo mutation in human disease. Nat Genet *46*, 944-950.

Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A.*, et al.* (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron *70*, 863-885.

Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S.*, et al.* (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron *87*, 1215-1233.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L.*, et al.* (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237-241.

Sanders, S.J., Neale, B.M., Huang, H., Werling, D.M., An, J.Y., Dong, S., Abecasis, G., Arguello, P.A., Blangero, J., Boehnke, M.*, et al.* (2017). Whole genome sequencing in psychiatric disorders: the WGSPD consortium. Nat Neurosci *20*, 1661-1668.

Scharf, J.M., Miller, L.L., Gauvin, C.A., Alabiso, J., Mathews, C.A., and Ben-Shlomo, Y. (2015). Population prevalence of Tourette syndrome: a systematic review and meta-analysis. Mov Disord *30*, 221-228.

Scharf, J.M., Yu, D., Mathews, C.A., Neale, B.M., Stewart, S.E., Fagerness, J.A., Evans, P., Gamazon, E., Edlund, C.K., Service, S.K.*, et al.* (2013). Genome-wide association study of Tourette's syndrome. Mol Psychiatry *18*, 721-728.

Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421-427.

Schmitz, L.J.M., Klaassen, R.V., Ruiperez-Alonso, M., Zamri, A.E., Stroeder, J., Rao-Ruiz, P., Lodder, J.C., van der Loo, R.J., Mansvelder, H.D., Smit, A.B.*, et al.* (2017). The AMPA receptor-associated protein Shisa7 regulates hippocampal synaptic function and contextual memory. eLife *6*.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J.*, et al.* (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445-449.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M.*, et al.* (2004). Large-scale copy number polymorphism in the human genome. Science *305*, 525-528.

Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., and Dunbar, G.C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry *59 Suppl 20*, 22-33;quiz 34-57.

Shirley, M.D., Baugher, J.D., Stevens, E.L., Tang, Z., Gerry, N., Beiswanger, C.M., Berlin, D.S., and Pevsner, J. (2012). Chromosomal variation in lymphoblastoid cell lines. Hum Mutat *33*, 1075-1086.

Singh, S.K., Stogsdill, J.A., Pulimood, N.S., Dingsdale, H., Kim, Y.H., Pilaz, L.-J., Kim, I.H., Manhaes, A.C., Rodrigues, W.S., Jr., Pamukcu, A.*, et al.* (2016). Astrocytes Assemble Thalamocortical Synapses by Bridging NRX1α and NL1 via Hevin. Cell *164*, 183-196.

Stahl, E., Forstner, A., McQuillin, A., Ripke, S., , Ophoff, R., Scott, L., Cichon, S., Andreassen, O.A., Sklar, P.*, et al.* (2017). Genomewide association study identifies 30 loci associated with bipolar disorder. bioRxiv.

Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E.*, et al.* (2008). Large recurrent microdeletions associated with schizophrenia. Nature *455*, 232-236.

Südhof, T.C. (2008). Neuroligins and neurexins link synaptic function to cognitive disease. Nature *455*, 903-911.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.*, et al.* (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75-81.

Sul, J.H., Cade, B.E., Cho, M.H., Qiao, D., Silverman, E.K., Redline, S., and Sunyaev, S. (2016). Increasing Generality and Power of Rare-Variant Tests by Utilizing Extended Pedigrees. Am J Hum Genet *99*, 846-859.

Sundaram, S.K., Huq, A.M., Wilson, B.J., and Chugani, H.T. (2010). Tourette syndrome is associated with recurrent exonic copy number variants. Neurology *74*, 1583-1590.

Takeda, Y., Akasaka, K., Lee, S., Kobayashi, S., Kawano, H., Murayama, S., Takahashi, N., Hashimoto, K., Kano, M., Asano, M.*, et al.* (2003). Impaired motor coordination in mice lacking neural recognition molecule NB-3 of the contactin/F3 subgroup. J Neurobiol *56*, 252-265.

Thompson, E. (2011). The structure of genetic linkage data: from LIPED to 1M SNPs. Hum Hered *71*, 86-96.

Tjio, J.H., and Levan, A. (2010). THE CHROMOSOME NUMBER OF MAN. Hereditas *42*, 1-6.

Trost, B., Walker, S., Wang, Z., Thiruvahindrapuram, B., MacDonald, J.R., Sung, W.W.L., Pereira, S.L., Whitney, J., Chan, A.J.S., Pellecchia, G.*, et al.* (2018). A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data. Am J Hum Genet *102*, 142-155.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet *45*, 124-130.

Ullrich, B., Ushkaryov, Y.A., and Südhof, T.C. (1995). Cartography of neurexins: more than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. Neuron *14*, 497-507.

Ushkaryov, Y.A., Petrenko, A.G., Geppert, M., and Südhof, T.C. (1992). Neurexins: synaptic cell surface proteins related to the alpha-latrotoxin receptor and laminin. Science *257*, 50-56.

Vacic, V., McCarthy, S., Malhotra, D., Murray, F., Chou, H.-H., Peoples, A., Makarov, V., Yoon, S., Bhandari, A., Corominas, R.*, et al.* (2011). Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. Nature *471*, 499-503.

Verkerk, A.J.M.H., Mathews, C.A., Joosse, M., Eussen, B.H.J., Heutink, P., Oostra, B.A., and Tourette Syndrome Association International Consortium for, G. (2003). CNTNAP2 is disrupted in a family with Gilles de la Tourette syndrome and obsessive compulsive disorder. Genomics *82*, 1-9.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and Bucan, M. (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.

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J.T., Abrahams, B.S., Salyakina, D., Imielinski, M., Bradfield, J.P., Sleiman, P.M.*, et al.* (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature *459*, 528-533.

Wang, Z., Maia, T.V., Marsh, R., Colibazzi, T., Gerber, A., and Peterson, B.S. (2011). The neural circuits that generate tics in Tourette's syndrome. Am J Psychiatry *168*, 1326-1337.

Westfall, P.H., and Troendle, J.F. (2008). Multiple testing with minimal assumptions. Biom J *50*, 745-755.

Willsey, A.J., Fernandez, T.V., Yu, D., King, R.A., Dietrich, A., Xing, J., Sanders, S.J., Mandell, J.D., Huang, A.Y., Richer, P.*, et al.* (2017). De Novo Coding Variants Are Strongly Associated with Tourette Disorder. Neuron *94*, 486-499.e489.

Worbe, Y., Marrakchi-Kacem, L., Lecomte, S., Valabregue, R., Poupon, F., Guevara, P., Tucholka, A., Mangin, J.-F., Vidailhet, M., Lehericy, S.*, et al.* (2015). Altered structural connectivity of cortico-striato-pallido-thalamic networks in Gilles de la Tourette syndrome. Brain *138*, 472-482.

Ye, H., Tan, Y.L.J., Ponniah, S., Takeda, Y., Wang, S.-Q., Schachner, M., Watanabe, K., Pallen, C.J., and Xiao, Z.-C. (2008). Neural recognition molecules CHL1 and NB-3 regulate apical dendrite orientation in the neocortex via PTP alpha. EMBO J *27*, 188-200.

Zarrei, M., MacDonald, J.R., Merico, D., and Scherer, S.W. (2015). A copy number variation map of the human genome. Nat Rev Genet *16*, 172-183.

Zhang, D., Cheng, L., Qian, Y., Alliey-Rodriguez, N., Kelsoe, J.R., Greenwood, T., Nievergelt, C., Barrett, T.B., McKinney, R., Schork, N.*, et al.* (2009). Singleton deletions throughout the genome increase risk of bipolar disorder. Mol Psychiatry *14*, 376-380.