GENOMICS AND PHENOMICS OF OBSESSIVE-COMPULSIVE AND RELATED DISORDERS

By

FRANJO IVANKOVIC

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2022

To my parents, Andrijana and Zlatko Ivanković for teaching me that where there is a will, there is a way

ACKNOWLEDGMENTS

First and foremost, I would like to thank my parents, Andrijana and Zlatko Ivanković for encouraging me to always be curious and persistent. Their hard work, self-sacrifice, and continued selfless support that sometimes came at a very high price are the main reasons I am writing this dissertation today. I would like to thank my sister, Ivana Ivanković, for being my very first frenemy and a partner in many life lessons (most of which resulted from us being troublemakers), and my two not-so-little brothers, Ivan and Filip Ivanković, whose childhood I have missed entirely due to my pursuit of education, but who have nonetheless been sources of a lot of support and joy.

Next, I would like to thank my husband - Jared Ivanković-MacGibbon, without whom I also would not be writing this dissertation. His ability to ground me during the darkest moments of my depression and anxiety, and somehow get me excited about the road ahead puzzles me to this day. Everyone should have a Jared. I would like to thank my parents in-law, Brenda MacGibbon and Todd Dixon, as well as brother in-law, Jacob MacGibbon, for their support, care, and being as far away from the in-law stereotype as one can be.

A big thank you to my American family, Ivan and Patrick Miličević, Kim LeBlanc, and Taylor Dossett for being a family far away from home – especially for the holidays, and great many lessons about adjusting to the society in the USA.

I am also very grateful to Milica Marković, my life-long friend and partner in crime, for being a source of many joyous and memorable moments throughout my life and around the world. I would like to acknowledge my amazing friends Elizabeth Delgado, Antonija Kolobarić, Marian Segura, Addelynn Sagadevan, and Caroline Jamieson, for

the abundance of love, support, and memorable experiences during my great American academic journey and hopefully beyond.

I would like to thank my Ph.D. mentor, Dr. Carol Mathews, for being an embodiment of an amazing mentor and scientist. Dr. Mathews made academia a safe space for learning and development, which is not a very easy thing to do (ask any other graduate student). Dr. Mathews has been a great inspiration during this journey, and I am forever grateful for the opportunity to be her trainee. Should my life take me to a similar position, I only hope I could be to my students what Dr. Mathews has been to me.

I would like to thank my clinical coordinator Robyn Nelson for her tremendous help from the moment I indicated interest in joining this lab to this very day. Robyn is one of the most kind and resourceful people I have ever had the pleasure of meeting. Next, I would like to thank Dr. Luis Sordo Vieira and Dr. Jason Cory Brunson for their mathematical bits of wisdom; and my fellow graduate students in the Mathews lab: Jessica Zakrzewski, Binh Nguyen, Sara Nutlet, and Elizabeth Chapman for making the lab an amazing learning environment. I would like to thank my undergraduate trainees James Shen, Sharon Johnson, and Daisy Valle for being self-motivated, resourceful, and overall amazing people to work with. Ultimately, I would like to thank Yi-Chieh Chang, Alexa Valko, Marie Jean Gilles, and Joelle Dorsett, for being amazing co-workers.

I would like to acknowledge my collaborators, Dongmei Yu, Dr. Laura Domenech Sanglaro, Dr. Maria Niarchou, Dr. Lingyu Zhan, Lisa Osiecki, Dr. Roel Ophoff, Dr. Lea Davis, and Dr. Jeremiah Scharf for their tremendous help during the research that has

eventually led to this dissertation, their insightful feedback on methodology and scientific reasoning, as well as their overall help and resources in all things academic.

I would also like to thank Dr. Ruan Oliveira, Dr. Łukasz Sznajder, and Dr. Michael Lewis for countless advice and opportunities, emotional support, friendship, and resourcefulness. I would also like to acknowledge Dr. Maurice Swanson, Dr. James Resnick, and Myrna Stenberg for their mentorship and support during the initial years of my doctoral education.

I am extremely grateful to my committee members: Dr. Yan Gong, Dr. Matthew Farrer, and Dr. Joseph Antonelli, for their amazing feedback, time and resources, and valuable career advice. I would like to thank my T32 grant principal investigators: Dr. Dawn Bowers, Dr. David Vaillancourt, and Dr. Zhigang Li, for their resourcefulness, feedback, and learning opportunities. I would like to thank Dr. Lauren McIntyre for reminding me just how extensive my love of statistics is and her unmatched teaching skills.

I would like to thank the Department of Psychiatry, Department of Clinical and Health Psychology, the Genetics Institute, the McKnight Brain Institute, and the Center for OCD, Anxiety, and Related Disorders for facilitating my education and graduate training. I am also grateful to the Genetics and Genomics graduate program, specifically Dr. Brittany Hollister, who has provided so much help and encouragement, particularly when things got tough, in addition to resolving many administrative hurdles that came along the way. Additionally, I would like to thank Cornelia Frazier, Cindy Heesacker, Dr. Samantha Brooks, Dr. Connie Mulligan, and Hope Parmeter for their commitment to

improving the Genetics and Genomics graduate program year after year, and their outstanding resourcefulness and care for students.

As it is obvious by now, I have a lot of people to acknowledge because each and every one of them played a unique and instrumental part in my academic career so far, however I cannot finalize this without first recognizing people who were instrumental in getting me to this point to begin with, starting with Dr. Cliff Hayes who has done so much to help me navigate American education system since my very first semester here. Special thanks Bill Kolb, Regan Garner, and Debra Anderson for being such an amazing and supportive team to myself and the rest of the international students. I can never repay them for all they have done for me and my undergraduate education at the University of Florida. I would also like to thank Shelby Davis and Davis United World College Scholars program for providing financial means to fulfill my dreams through education at the University of Florida. Special thanks to Dr. Nicole Gerlach and Dr. Bernard Hauser for teaching me so much about evolution and genetics and helping me along my way to graduate school.

I would also like to thank the United World College in Mostar for providing such a rich multicultural high-school learning experience. I will always be proud of being a part of the United World College movement. I would like to especially thank my teachers Ilvana Čišić, Merima Homarac, Ivona Sušac, and Tanja Čvoro, for their valuable mentorship and commitment to education, as well as houseparents Maja Pandža and Alisa Kazić for their friendship and care.

Special thanks to Ruđer Bošković Electrical Engineering and Computer Science High School, especially Marija Grgić, Nina Kovačević, and Željka Bevanda, for playing a

crucial role in my academic career and helping me with taking important steps which ultimately lead me here.

Which leads me to my baby academic steps, where I would like to acknowledge people who have played an important role during my formative years. Thanks to Mladenka Korać, Ružica Ćorić, Anica Martinac, and Silvana Smoljan for planting the seeds of curiosity which eventually turned into love of science and nature.

Lastly, I would like to acknowledge and thank everyone else not mentioned above. Every teacher, friend, and family member who has shaped me into a person I am today. This dissertation and its accompanying degree are my proudest accomplishments so far, and I truly believe all of you played an important role in getting me here. I will be forever grateful.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1kGPp3	The 1000 Genomes Project, phase 3
ABCD	The Adolescent Brain Cognitive Development (Study)
ADHD	Attention deficit / hyperactivity disorder
AN	Anorexia nervosa
ASEBA	Achenbach system of empirically based assessments
ASD	Autism spectrum disorder
AUD	Alcohol use disorder
BAF	B allele frequency
BED	Binge eating disorder
BHR	The Brain Health Registry
bOCD	Broad OCD
bnOCD	Broad and narrow OCD
BPD	Bipolar disorder
CBCL	Child Behavioral Checklist
ccGWAS	Case-control GWAS
CD	Conduct disorder
CGH	Comparative genome hybridization
CNV	Copy-number variant
CMVT	Chronic motor or vocal tic disorder, synonymous to PMVT
CT	Chronic tics, synonymous to PMVT
DNA	Deoxyribonucleic acid
DOCS	Dimensional Obsessive-Compulsive Scale
DSM-5	*Diagnostic and Statistical Manual of Mental Disorders*, 5th edition
DSM-IV	*Diagnostic and Statistical Manual of Mental Disorders*, 4th edition

EFD	Eating or feeding disorder
EM	Expectation maximization
ER	Endoplasmic reticulum
qQTL	Expression quantitative trait loci
FDR	False discovery rate
FISH	Fluorescent *in situ* hybridization
GAD	Generalized anxiety disorder
gnomAD	Genome Aggregation Database
GO	Gene ontology
GRM	Genomic relationship matrix
GSAv1	Global Screening Array-24 BeadChip version 1
GSAv3	Global Screening Array-24 BeadChip version 3
GTEx	Genotype-Tissue Expression (database)
GWA	Genome-wide association
GWAS	GWA study
HD	Hoarding disorder
HMM	Hidden markov model
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
IBS	Identity by state
IOCDF-GC	The International OCD Foundation Genetics Collaborative
IRB	Internal review board
IRR	Incidence rate ratio
ISH	*In situ* hybridization
kb	Kilobase(s): 1,000 base pairs

KSADS-5	Kiddie schedule for affective disorders and schizophrenia for DSM-5
LASSO	Least absolute shrinkage and selection operator
LD	Linkage disequilibrium
LDSC	LD score regression
LMM	Linear mixed model
LRR	Log R ratio
MAF	Minor allele frequency
MAGMA	Multi-marker analysis of genomic annotation
Mb	Megabase(s): 1,000,000 bases
MCMC	Markov chain Monte Carlo
MDD	Major depressive disorder
ML	Machine learning
MLE	Maximum likelihood estimation
mQTL	Methylation quantitative trait loci
MR	Mendelian randomization
NDA	NIMH Data Archives
NDD	Neurodevelopmental disorder(s)
NHGRI	Natinal Human Genome Research Institute
NIDA	National Institute for Drugs and Addiction
NIMH	National Institute for Mental Health
nOCD	Narrow OCD
OCD	Obsessive-compulsive disorder
OCGAS	OCD Collaborative Genetics Association Study
OCI-R	Obsessive-Compulsive Inventory, Revised
OCP	Obsessive-compulsive problems

OCRD	Obsessive-compulsive and related disorders
OCS	Obsessive-compulsive symptoms
ODD	Oppositional defiant disorder
OR	Odds ratio
PBWT	Positional Burrows-Wheeler transform
PC	Principal component
PCA	Principal component analysis
PD	Panic disorder
PFB	Population frequency of B allele
PGC	Psychiatric Genomics Consortium
PMVT	Persistent motor or vocal tic (disorder)
PMVTD	Persistent motor or vocal tic disorder
PRS	Polygenic risk score
PTD	Provisional tic disorder
PTSD	Post-traumatic stress disorder
QC	Quality control
qGWAS	Quantitative GWAS
qPCR	Quantitative polymerase chain reaction
QQ	Quantile-quantile (plot)
QTL	Quantitative trait loci
REML	Restricted maximum likelihood
RNA	Ribonucleic acid
RRR	Relative recurrence risk
SCZ	Schizophrenia
SeAD	Separation anxiety disorder

SFARI	Simons Foundation Autism Research Initiative
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SoAD	Social anxiety disorder
SP	Specific phobia
SPARK	Simons Foundation Powering Autism Research and Knowledge
SPD	Sleeping problems and disorders
SSC	Simons Simplex Cohort
TAAICG	Tourette Association of America International Consortium for Genetics
TADA	Transmission and *de novo* association
TD	Tic disorder
TOCD	Tourettic OCD
TOCS	Toronto Obsessive-Compulsive Scale
TS	Tourette syndrome
WES	Whole exome sequencing
WGS	Whole genome sequencing
WTCCC	Wellcome Trust Case Control Consortium
Y-BOCS	Yale-Brown Obsessive Compulsive Scale

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

GENOMICS AND PHENOMICS OF OBSESSIVE-COMPULSIVE AND RELATED DISORDERS

By

Franjo Ivankovic

August 2022

Chair: Carol A. Mathews
Major: Genetics and Genomics

Tourette syndrome (TS) and obsessive-compulsive disorder (OCD) are neuropsychiatric disorders with onset in childhood affecting 0.6% and 2.3% of people, respectively. TS and OCD are also highly comorbid with 50-60% of TS patients endorsing OCD, and 10% of OCD patients endorsing TS. Both TS and OCD are highly heritable, with heritability estimates ranging 30% to 60% in family and twin studies. Despite substantial heritability estimates, little is known about underlying genetic mechanisms of OCD and related disorders (OCRD).

In this dissertation, I explore OCRDs from both phenomic and genomic aspects. I use rich phenotypes from ABCD Study to investigate OCRD comorbidity and relationships with symptom-level data from the child behavioral checklist (CBCL). I also leverage genome-wide association to explore genetic architecture of OCD and related phenotypes, including polygenic risk score (PRS) analysis with tic disorders within ABCD Study and 12 disorders from the Psychiatric Genomics Consortium. I additionally explore copy-number variation (CNV) among neurodevelopmental disorders, specifically focusing on neurodevelopmental disorders including TS and autism spectrum disorder (ASD).

Phenomic analysis of psychopathology in ABCD Study has shown hyperinflated rates of psychiatric disorders in the ABCD Study, likely due to self-endorsement bias. To circumvent that, I define a narrow diagnosis construct that utilizes longitudinal data to refine psychiatric diagnoses. Narrow OCD (nOCD) better reflected childhood OCD prevalence rates and comorbidity patters, and a stronger relationship with symptom-level data from CBCL. Genomic assessment of nOCD has also shown stronger PRS relationship with OCD symptoms compared to broad OCD. Similar effects were also observed in PRS analysis with 12 PGC disorders. CNV analysis of TS has resulted in successful replication of TS-risk contribution by *NRXN1* deletions and *CNTN6* duplications, as well as identification of 39 additional genes that could potentially contribute to TS pathology. However, genome-wide burdens of CNV numbers or sizes were not replicated.

Deconvoluting genetic and phenomic relationships and underpinnings of OCRDs is a complicated task confounded primarily by low sample sizes and suboptimal methodologies. Thus, increased recruitment efforts and improvements to statistical and computational methodologies to analyze these data will likely be the main drivers of discoveries in the OCRD genomics space.

CHAPTER 1
INTRODUCTION

Statistical Genomics

Twenty-first century advances in deoxyribonucleic acid (DNA) genotyping and sequencing, specifically with the development of high-density microarrays and next generation sequencing, have resulted in a large amount of high-dimensional data. This circumstance has enabled scientists to explore complex genomic architectures of non-Mendelian disorders. However, our ability to generate high-dimensional genomic data is rapidly outpacing our ability to analyze and make sense of them. Some estimates place the cumulative number of human DNA sequences (genome and exome) in 2025 at about 500-1000 million (Stephens et al., 2015). Presently, about 1 million whole genomes were sequenced, and the Genome Aggregation Database (gnomAD) has collected 91,864 of them (Karczewski et al., 2020). For this reason, there is a high need for development and application of advanced statistical methodologies and utilization of complex computational workflows which can handle such large datasets.

Statistical genomics is an interdisciplinary area of science broadly defined by its focus on applications of statistical and computational methodologies on large genomic datasets. In most applications, generalizations of existing statistical methods coupled with modern computational technology is sufficient to explore this data, however the necessity of innovative approaches is ever-increasing.

In this dissertation, I discuss several well-established methods within this field, as well as novel modifications thereof, to analyze high-dimensional, high-throughput data on several neurodevelopmental disorders. These methods can be broadly classified as polymorphism analyses and structural variant analyses. Furthermore, in this

dissertation, all data were generated by microarray genotyping, a highly affordable technique capable of assessing known markers across the human genome.

Microarray Genotyping

Microarray genotyping has its roots in mid-19th century, as a hypothesized application of Southern blotting on large numbers of probes of interest, however, the first such system was published in the 1990s, to analyze differential expression of 45 genes in Arabidopsis thaliana (Schena et al., 1995). Since then, millions of people have had their genome genotyped on one of such arrays. Various companies have developed human microarrays, which have been commercialized to deliver ancestry, health, and other trait data to their consumers. For example, 23andMe alone has genotyped over 12 million individuals world-wide, revolutionizing direct-to-consumer genetic testing (23andMe, n.d.). Due to low price, well-defined protocols, and ability to handle large numbers of samples quickly and in a reproducible manner, microarray genotyping has been a research tool of choice for almost three decades. The resolution of microarrays has also changed substantially, from 45 probes on the Schena array in 1995, to Illumina's Infinium Omni5-4 microarray which has more than 4.3 million genome-wide probes and can genotype 4 samples simultaneously (Illlumina, n.d.).

Microarray technology works on the principle of nucleic acid hybridization: an array is prepared with pre-determined immobilized nucleic acid sequences bound to quenched fluorophores (targets), followed by hybridization of one or more labeled samples, and detection of released fluorophores with microscopic cameras (Amersham Biosciences, 2002). The source of target sequences can be a cDNA library (for RNA detection), amplified regions of genomic DNA, and in silico synthesized oligonucleotides from gene databases. Although initially developed to measure differential expression of

RNA, microarrays have soon found applications in common variant (polymorphism) and structural variant applications (Iafrate et al., 2004). Polymorphism analysis leverages microarray data to call single nucleotide polymorphisms (SNPs) for the purposes of genome-wide association (GWA) analyses, whereas structural variant analysis leverages microarray data to estimate copy-number variation (CNV) across the genome.

Polymorphism Analysis

SNPs are the simplest yet most common of all polymorphic variation in the human genome and are characterized as loci that usually have only two alleles, corresponding to two different nucleotides at that specific genomic locus (Nussbaum et al., 2016). Genome-wide association study (GWAS) leverages genotypes at hundreds of thousands of loci for hundreds or thousands of individuals to identify genetic correlations with the trait of interest. This approach has been found to be particularly useful for polygenic and multifactorial traits, where the transmission of numerous genetic variants in complex inheritance patterns underlies the phenotype (NHGRI, 2020). Additionally, many SNPs that are found to be associated with a trait might not be directly responsible for the phenotypic presentation, their proximity to causal genetic variants results in linkage-mediated association with the trait. Thus, GWAS approaches can be used to find either causative genetic variants or help localize regions of the genome harboring such variants for the traits of interest.

Statistically, GWAS is a method encompassing SNP-wise independent tests of association usually involving Chi-square, Fisher's exact, or logistic regression tests where the trait is categorical (e.g., presence or absence of a disorder) and linear regression and its derivatives (such as Poisson, negative binomial, etc.) where the trait

is quantitative (e.g., systolic blood pressure). Due to a large number of probes, and consequently a large number of tests, in GWAS, multiple-correction testing is necessary to curb the high number of spurious false positive associations. Assuming α represents α level or significance threshold, and k represents the number of SNPs (i.e., number of independent tests), then the probability of finding one positive SNP association can be expressed as:

$$P_{SNP} = 1 - (1 - \alpha)^k, \tag{1-1}$$

where the probability of finding a positive association P_{SNP} is asymptotic to 1 (virtually guaranteed) for a typical number of SNPs examined in the GWAS analyses (Streiner & Norman, 2011). In essence, at α = 0.05 threshold, and k = 250,000 SNPs tested, the probability of finding at least one positive SNP association $P_{SNP} = 1$, even in the absence of a true meaningful association between that SNP and the trait. To address this issue, a strict cut-off of $\alpha = 5 \times 10^{-8}$ was introduced to account for the linkage and multiple tests (Risch & Merikangas, 1996), it was derived as a Bonferroni correction for 1,000,000 SNPs at nominal α = 0.05. At this α, the probability of one positive association for k = 250,000 SNPs tested $P_{SNP} = 0.0124$. Although this approach has been used for several decades now to correct for multiple testing in GWA analyses, recent studies point at the possibility of overcorrection and excessive stringency (Chen et al., 2021).

To further increase the power of GWAS analyses, genetic imputation of the phased genotypes was introduced to allow allele estimation of millions of loci using the well genotyped and sequenced references (Li et al., 2009). Genetic phasing is a statistical process by which haplotype blocks are estimated on the genotyped data.

Unlike sequencing, in genotyping data it is not immediately apparent which allele on each independent locus is physically linked to which allele on the next independent locus. To get around that, clustering-based algorithms in conjunction with well-sequenced reference data are used to statistically estimate which SNPs are inherited together (Roach et al., 2011). Phased genotypes can further be used in conjunction with hidden Markov models (HMM) to estimate the values for missing genotypes or even non-genotyped loci, which is known as genetic imputation (Howie, Donnelly & Marchini, 2009).

Such drastic increase in testable SNPs expands the ability of GWAS to probe genetic associations at a much higher resolution and allow fine mapping of associated loci. However, it is prudent to mention that the quality of genotype imputation is directly related to the quality of reference panels used and the need for more comprehensive reference panels (Shi et al., 2019), particularly among underrepresented ancestries (admixed and Native Americans, Africans, and Asians). It is important to mention that presently, imputations are used to form datasets of about 5,000,000 loci to be used in downstream GWASes. Based on the Equation 1-1, the probability of observing at least one positive hit at cutoff $\alpha = 5 \times 10^{-8}$ is $P_{SNP} = 0.2212$. For this reason, a closer examination of associating loci with respect to local recombination rates is usually conducted to determine the likelihood of a real association. Furthermore, due to linkage, it is expected that neighboring loci will also stand out in their strength of association - hence the reason why towering number of SNPs is usually preferred to a result of a sole singular association on the Manhattan plot.

It has been 15 years since the first high-resolution high-throughput GWAS was published examining 7 common disorders (~2,000 cases for each) and 3,000 shared controls totaling 17,000 samples on a 500,000 SNP Affymetrix GeneChip microarray (WTCCC, 2007). Since then, numerous aggregation databases have been established to track reported GWAS results. One such database, the GWAS Catalog, has documented results from 5,690 publications and 372,752 associations as of April 7th, 2022 (Buniello et al., 2019). Based on thousands of GWAS publication for variety of traits, we have learned that complex traits are highly polygenic, pleiotropy is pervasive, methodological improvements drive new discoveries, common variants tag a substantial portion of additive genetic variance, and GWAS discoveries enable functional and therapeutic research (Visscher et al., 2017). Of course, improvement to GWAS methodology have also been introduced over this period, including improvements to quality control (QC) procedures, microarray genotyping protocols, diversification of the samples as well as aggregation of the large number of samples in GWAS studies, improved statistical testing for association, improved computational pipelines for phasing and imputation, etc. Sample sizes have increased exponentially since the early 2010s, especially given the emergence of large national and international collections, as well as commercial collections, such as UK Biobank with 500,000 participants, BioBank Japan with 260,000 participants, China Kadoorie Biobank with 510,000 participants, H3Africa with 100,000 participants, BioMe with 50,000 participants, TOPMed with 145,000 participants, Million Veteran Programme with 840,000 participants, All of Us with 132,000 individuals, and 23andme with several million participants (Uffelmann et al., 2021). Another contributing factor to increased sample sizes and power has been

collaborative aggregations of samples or GWAS summary statistics into large meta-analyses facilitated by consortia such as the Psychiatric Genomics Consortium (PGC).

Concomitantly to GWAS methodology development, numerous secondary-level analyses on genotyping data have been described that allow exploration of genetic underpinnings of complex hereditary disorders. The first obvious approach would be to estimate heritability for a given trait using actual genetic data instead of pedigrees – specifically referred to as SNP heritability (and annotated as h_{SNP}^2). SNP heritability has been traditionally defined as a proportion in phenotypic variance explained by genotyped SNPs (Yang et al., 2017). However, further evaluations of SNP heritability methods have shown that accurate estimations of h_{SNP}^2 requires accounting for minor allele frequency (MAF) and LD, as well as large sample sizes (Evans et al., 2018; Visscher et al., 2014). Extension to SNP heritability estimation, by considering aggregating nonlinear effects and interactions, LD score regression has been developed as tool that has the added benefit of potentially distinguishing between polygenic effects and confounding factors such as population structure (Bulik-Sullivan, Loh, et al., 2015). When it comes to analyzing relationships between multiple complex traits, genetic correlations have emerged as methods of choice, initial forms thereof essentially being an extrapolation of LD score regression to cross-trait analysis (Bulik-Sullivan, Finucane, et al., 2015). Causal analyses methods are still a very active area of research, with one notable example being Mendelian randomization (MR), a method that can be used to infer causality of an exposure for a complex disease outcome (Verbanck et al., 2018; Davey Smith & Ebrahim, 2003).

In addition to heritability and genetic correlation analysis, polygenic risk score (PRS) based phenotype prediction methodology has also been developed. In their simplest forms, PRS are the sum of products of risk alleles at numerous loci with the weights derived from GWAS summary statistics reports at those loci (Choi et al., 2020). This can be expressed as a simple equation to calculate PRS, $\hat{S}$.

$$\hat{S} = \sum_{i=1}^{n} X_i \beta_i, \tag{1-2}$$

where $\hat{S}$ represents PRS, *n* represents number of SNPs examined, X_i represents number of risk alleles at the i^{th} locus, and β_i represents the weight of risk based on GWAS summary statistics. This procedure relies on a large-enough discovery sample where GWAS is performed to derive weights, and an independent target sample where the associations can be examined. Some considerations for valid PRS estimations include powered discovery and target samples, proper accounting for ancestry, and analysis of LD-independent loci (Euesden et al., 2014; Coombes et al., 2020; Wand et al., 2021). There are numerous specific methods for PRS analysis, those utilized in this dissertation will be discussed further in later sections.

Due to these methodological advancements and collaborative science, GWAS analyses are still powerful tools for better understanding of genetic underpinnings and genetic-based relationships between complex phenotypes. However, adequate sample sizes and diversity of samples recruited into these studies remain a major challenge to the field.

Structural Variant Analysis

CNVs are more complex common variants in the human genome, characterized as large deletions or duplications of the genomic segments of size range 10kb to over

1Mb (Nussbaum et al., 2016). These deletions can occur on one or both homologous chromosomes, leading to complete loss of the genetic material. The duplications can also occur on one or both homologous chromosomes, however duplications can also occur once, twice, or even more times, leading to substantial amplification of the genetic material. CNV analyses from microarrays trace back to the early 2000s, with the use of specialized microarrays and analysis tools such as representational oligonucleotide microarray analysis (Sebat et al., 2004). The statistical approach to CNV analysis, however, is a bit more involved and limited. Unlike GWAS where signals are interpreted as binary alleles, CNV analysis relies on direct measurements of hybridization signal intensity. This means that CNV analysis is more vulnerable to stochastic and technical fluctuations in such signals, and neighboring loci cannot be imputed. Furthermore, CNV calls rely on an additional layer of computations using hidden Markov models (HMM) or likelihood-based approaches modeled on various copy-number scenarios (Seiser & Innocenti, 2014; Illumina, 2014). While SNP arrays can be re-appropriated for the purposes of CNV analysis with valid and reliable output, a substantial limitation to this approach is the inability to detect sequence-neutral alterations such as translocations and inversions (Coughlin, II et al., 2012). Additionally, in the extant literature, multiple algorithms for CNV estimation are usually combined to maintain low rates of false positive CNV calls and, in some cases, additional *in silico* or *in vitro* validations are also performed (Marshall et al., 2017; Huang et al., 2017; Wang et al., 2018).

Much like SNPs, CNVs can also be polymorphic and do not necessarily have a clinical consequence (Coughlin, II et al., 2012). About 4.8-9.5% of the human genome is subject to common CNV occurrence, and about 100 genes can be deleted completely

without apparent phenotypic consequences (Zarrei et al., 2015). Conversely, *de novo* and rare CNVs have been implicated in various neuropsychiatric phenotypes, including autism, schizophrenia, and bipolar disorder (Malhotra & Sebat, 2012).

Like GWAS, methodological advancements and collaboration enable insight into genetic risk conferred by these mutations that can in some instances affect multiple genes simultaneously. Both global effects of CNVs as well as locus specific CNVs can be explored as potential effectors in the psychiatric pathogenesis.

Psychiatric Genomics

Developments in genomic methodology and technology have ushered the field of psychiatric research into the big data era. The Psychiatric Genomics Consortium (PGC), formed in 2007, operates under a central idea of leveraging global collaboration to advance genetic discovery of biologically, clinically, and therapeutically meaningful insights within the realm of psychiatry (PGC, n.d.). After over a decade of efforts, PGC now counts over 800 investigators from over 150 institutions and 40 countries and has insofar produced more than 300 publications. Central to PGC efforts are high-throughput, high-resolution methods like GWAS, CNV analysis, and whole genome and exome sequencing with emphasis on elucidating the genetic portions of these complex traits, and further informing research that will yield fundamental understanding of underlying biology, inform clinical practice, and deliver therapeutic targets (Sullivan er al., 2018).

While a more detailed overview of genomic literature of psychiatric disorders will follow in subsequent chapters, there are several key lessons learned in the recent history of psychiatric genomics. Multifactorial polygenic inheritance plays a key role in transmission of risk of psychiatric disorder development, especially through often

neglected common variants which individually confer effects with OR < 1.1 (WTCCC, 2007). Pleiotropic effects, i.e., the ability of a single variant to influence multiple traits, have also been found to be ubiquitous across virtually all psychiatric disorders (O'Donovan, 2015). Such findings have led to a conclusion that the heterogeneity of psychiatric traits might be traced to variability of genetic risk profiles that underlie them.

This dissertation aims to further explore pleiotropy and polygenicity, focusing on a cluster of neurodevelopmental disorders: obsessive-compulsive and related disorders (OCRD), specifically obsessive-compulsive disorder (OCD) and chronic tic disorders (TD). OCRDs generally include disorders with similar phenomenological features to OCD, are highly comorbid with OCD, or can be considered unusual presentations of OCD; including tic disorders (TD), body dysmorphic disorder, hypochondriasis, autism spectrum disorder (ASD), and eating disorders (Murphy et al., 2010). It is important to note that in DSM-5, OCD is clustered in a group of disorders similarly called OCRD which include OCD, body dysmorphic disorder, hoarding disorder (HD), trichotillomania, excoriation, etc (APA, 2013). In this dissertation, however, OCRD will be defined as a set of neurodevelopmental disorders with shared genetic and phenotypic characteristics and focus specifically on OCD, TD, and ASD.

In the next two chapters I provide an overview of the genomic literature of the OCRDs and the complex phenotypic and genetic relationship between them. In the subsequent chapter, I explore the relationship between CNVs and OCRDs in a large sample of parent-child trios. In the final two chapters, I conduct an association study between common variants and OCD, as well as a deeper dive into how best to construct the OCD phenotype.

CHAPTER 2
PROGRESS IN GENOMICS OF NEURODEVELOPMENTAL DISORDERS

Overview of the Traits

Neurodevelopmental disorders (NDD) are psychiatric disorders with an onset in the developmental period. They usually manifest early in childhood, and their severity ranges from transient, mild impairments with minimal effect on everyday life, to severe disorders that drastically reduce quality of life and persist well into adulthood.

Tic Disorders

Tourette syndrome (TS) was first described by Georges Gilles de la Tourette in a 1885 collection of case histories of a nervous disorder characterized by involuntary movements, repetition of speech, and use of obscene language (Yorston & Hindley, 1998).

The Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM5) defines tic disorders as neurodevelopmental motor disorders characterized by the presence of sudden, rapid, recurrent, nonrhythmic, stereotyped motor movements or vocalizations - known as motor and vocal tics (APA, 2013). There are four specific diagnostic categories within this cluster TS, persistent motor or vocal tic disorder (PMVTD), provisional tic disorder (PTD), and other (un)specified tic disorders.

Obsessive-Compulsive Disorder

The history of OCD is a bit more extensive than that of TS. One of the earliest potential cases of OCD involves a 7th century record by John Climacus (a Christian monk) of a young monk plagued by constant and overwhelming temptations to blasphemy (Osborn, 1999), what in modern days would be considered obsessions of unwanted taboo (sexual, religious or aggressive) thoughts. A more empirical view of

OCD came about in the 20th century, particularly due to the rise of psychoanalytic theory and Sigmund Freud, whose approaches to OCD treatment remained dominant until the 1980s (Zohar, 1987). DSM-5 defines OCD as the presence of recurrent and persistent, intrusive, and unwanted, thoughts, urges, and images - obsessions; and/or repetitive behaviors or mental acts that one feels driven to perform in response to an obsession or according to rules that must be followed rigidly - compulsions (APA, 2013). OCD can be further stratified by the insight of the patient, and the presence or absence of tics.

Genomics of Neurodevelopmental Disorders

A substantial amount of research has been done to elucidate the genetic underpinnings of neurodevelopmental disorders. Initial studies focused on analyses of family transmission and family history; however, the field of psychiatric genomics has kept pace with developing genomic technologies and methodologies. Modern approaches rely on previously discussed high-throughput, computationally demanding methods like GWAS or CNV analyses, as well as whole genome and whole exome sequencing, to detangle complex relationships between the human genome and heterogeneous psychiatric phenotypes.

Tic Disorders

Family history and pedigree-based studies have been historically important tools in mapping disorders afflicting humans and making inferences about their heritability. Despite initial suggestions that TS was heritable, made by Gilles de la Tourette himself, the very first published familial instance was that of two sisters and the son of one of the sisters in Connecticut in 1973 (Friel, 1973).

The initial evidence for the significant genetic component to TS etiology came from twin studies published nearly 10 years later. Price et al. reported concordance

rates for TS as 53% in 30 pairs of monozygotic and 8% in 13 pairs of dizygotic twins; expanding the criteria to include any tic disorder resulted in concordance rates of 77% and 23% in monozygotic and dizygotic twins, respectively (Price et al., 1985). In another study of monozygotic twins, Hyde et al. reported a concordance rate of 56% for TS and 94% for any tic disorder in 16 pairs of monozygotic twins (Hyde et al., 1992).

These findings suggest both genetics as a primary driver of TS pathology, and notable but incomplete penetrance for the risk variants. This is further corroborated by Pauls et al. (1991) in the family study of 86 TS probands and their 338 biological relatives, where the rates of TS (8.3%), PMVT (16.3%), and OCD (9.5%) were substantially higher among the relatives of TS probands compared to control probands (0%, 2.3%, 2.3% for TS, PMVT, and OCD, respectively). This paper is also one of the first to demonstrate a possibility of shared genetic risk between TS and OCD. Mataix-Cols et al. (2015) published a large family study of 4,826 individuals with tic disorders (both TS and PMVT) which found the risk of tic disorders in relatives to be proportional to genetic relatedness, with first-degree relative (OR = 18.7), second-degree relative (OR = 4.6), and third-degree relative (OR = 3.1).

Subsequent studies have attempted to clarify the pattern(s) of inheritance and identify specific genetic variants causing TS. Initial work using segregation and linkage studies suggested that TS might be a highly penetrant, sex-influenced, autosomal dominant trait, however subsequent studies have suggested that inheritance was not dominant and argued for mixed model of inheritance, instead (Pauls & Leckman, 1986; Eapen et al., 1993; Hasstedt et al., 1995; Walkup et al., 1996). Despite substantial

efforts to identify specific susceptibility genes using linkage studies, no reproducible variants have been identified with this method (Singer, 2000; Pauls, 2003).

Genetic linkage studies take advantage of linkage disequilibrium to identify regions of the genome associating with a targeted phenotype. Thus far, various loci across 9 different chromosomes have been associated with TS, however none have resulted in the discovery of specific causal mutations (Paschou et al., 2004; Ercan-Sencicek et al., 2010; Qi et al., 2019; Abelson et al., 2005; Díaz-Anzaldúa et al., 2005; Simonic et al., 1998; Mérette et al., 2000; Díaz-Anzaldúa et al., 2004). Two potential candidates: *HDC* and *SLITRK1* have stood out (Ercan-Sencicek et al., 2010; Abelson et al., 2005), however these are likely accounting for a very small proportion of TS cases or are family specific. Additional obstacles to validation of these variants might be their low penetrance and low frequency. *SLITRK1* has been associated with TS on multiple occasions, however these associations have not been validated in more recent, large-scale studies.

Additional early attempts to elucidate the genetic underpinnings of TS also included candidate gene studies. This type of study relies on examining the association of specified genes, preselected based on the potential biological relevance, to determine allelic contribution to phenotypic manifestation. Unlike linkage studies which probe for low-resolution disequilibrium blocks/cytobands, candidate gene studies have an advantage of testing for association between a specific gene against the targeted phenotype using one or more SNPs in the candidate locus (Qi et al., 2019). Moreover, carrying out these tests is relatively simple and inexpensive, which is what made them a

popular tool to test specific genes found in associated loci or other genes known to have a neurobiological function (Qi et al., 2017).

However, these experiments are vulnerable to high false-positive rates, and the results are usually difficult to validate for several reasons, but principally due to disregard for population structure, variable allele frequencies, and inability to accurately reproduce/measure phenotypes in animal models. Numerous candidate genes have been tested, yet for all of them, validation has proven to be difficult (Qi et al., 2017; Brett et al. 1995; Chou et al., 2007; Thompson et al., 1998; Gelernter et al., 1993; Brett et al., 1995; He et al., 2015; Abdulkadir et al., 2017; Hebebrand et al., 1993; Barr et al., 1997; Cavallini et al., 2000; Tarnok et al., 2007; Barr et al., 1999).

Studies probing structural genomic variation have historically relied on traditional molecular biology techniques to assess structural changes in the genomes of the patients. These variants have been usually classified as chromosomal insertions, deletions, duplications, inversions, translocations, and copy-number variations. The inexhaustive list of methods to detect these variants include G-banding via Giemsa stain, DNA (fluorescent) in-situ hybridization or (F)ISH, comparative genome hybridization (CGH), chromosome microarray analysis, quantitative polymerase chain reaction (qPCR), and (Sanger) sequence analysis. Numerous loci, most notably *SLITRK1*, have been found to associated with TS (Karagiannidis et al., 2013; Bertelsen et al., 2014; Melchior et al., 2013; Hooper et al. 2012; Patel et al., 2011; Lawson-Yuen et al. 2008; Pies 2008; Jankovic & Deng, 2007; Shelley et al., 2007; Belloso et al., 2007; Robertson et al., 2006; Abelson et al., 2005; Cuker et al., 2004; Crawford et al., 2003; Verkerk et al., 2003; State et al., 2003; Kerbeshian et al., 2000; Petek et al., 2001;

Matsumoto et al., 2000). It is important to note, however, that most of these findings identify rare structural variants that are unique to individuals or individual families, and thus, while they may be useful for identifying genes and biological pathways of interest, as individual risk variants, their generalizability is limited.

While there have been multiple genetic studies of TS, only in the past decade has progress in understanding the genetic architecture of this disorder accelerated. In part this is due to an increased understanding of the complex and often polygenic effect of diverse genetic variants on TS risk. While informative, most of the previous studies have a limited reach due to small sample sizes and limited resolution of genome testing. These limitations have been greatly ameliorated by the introduction of large-scale, population stratified, genome-wide techniques such as genome-wide microarrays, whole exome, and whole genome sequencing.

Members of the PGC workgroup for TS and OCD have so far published three different TS GWAS projects (Scharf et al., 2013; Yu et al., 2019; Tsetsos et al., 2021). While only a single locus was identified at a genome-wide significance threshold in the second and largest of these studies, with the first identifying no loci that met this criterion, they have nevertheless provided valuable insights into genetic architecture of TS. The third study took an alternative approach of analyzing gene sets.

Scharf et al. (2013) was the first reported GWAS of TS, looking at 1,285 TS cases and 4,964 ancestry matched controls. This study looked at individuals of European ancestry from 20 sites across USA, Canada, Netherlands, and Israel. While no markers were associated with TS at a genome-wide significance threshold, the top signal found was rs7868992 marker on chromosome 9q32 within *COL27A1* gene (p =

1.85×10^{-6}). The second phase of the study looked at additional samples from two closely related Latin American population isolates from the Central Valley of Costa Rica (124 cases) and Antioquia, Colombia (87 cases), for a total of 1,496 cases. The same rs7868992 marker emerged as the top signal ($p = 3.60 \times 10^{-7}$) in this expanded sample. Subsequent enrichment analysis of expression and methylation quantitative trait loci (eQTL and mQTL, respectively) found that the top SNPs from the primary analysis were nominally enriched for eQTL in frontal cortex ($p = 0.045$), borderline enriched for eQTL in cerebellum ($p = 0.077$), and nominally enriched for mQTLs in cerebellum ($p = 0.011$). The authors of this study also examined the associations of 2,135 SNPs that fell within 50kb of 24 previously reported TS candidate genes within this GWAS. None of these SNPs met the threshold for statistically significant association with TS.

In the second published GWAS, Yu et al. (2019) performed genome-wide association studies using 4 different TS datasets, resulting in the final GWAS meta-analysis which consisted of 8,265,319 SNPs and 4,819 TS cases (1,285 of which were from the first GWAS by Scharf et al., 2013), and resulted in a genome-wide significant hit of rs2504235 locus on chromosome 13p12.2 ($p = 2.1 \times 10^{-8}$), with an odds ratio of OR = 1.16. This marker lies within an intron of *FLT3*, a tyrosine kinase gene. However, this SNP was not replicated in the independent replication sample from Iceland. SNP-based heritability estimates of TS yielded $h_{SNP}^2 = 0.56$ in the Scharf et al. (2013) sample and $h_{SNP}^2 = 0.29$ in the large web- and clinic-based sample.

Ancestry-adjusted PRS was calculated from the GWAS meta-analysis and found to be larger, on average, in case subjects from multiplex versus simplex families. Additionally, higher TS PRS was significantly correlated with increased worst-ever tic

severity (β = 0.93; SE = 0.42; p = 0.026). Subsequently, this PRS was compared in the replication sample from Iceland. TS PRS was significantly higher in Icelandic TS cases (OR = 1.33; $p = 5.3 \times 10^{-9}$) and PMVT/unspecified tic disorder cases (OR = 1.20; $p = 5.2 \times 10^{-4}$). These variants also explained 0.78% and 0.42% of phenotypic variance, respectively. The direct comparison of these two groups confirmed higher PRS burden in TS (OR = 1.14; p = 0.05), representing an excess of 0.37% of the phenotypic variance. This evidence is compatible with the idea of tic disorders being a single continuous neurodevelopmental tic spectrum disorder, with TS being a more severe manifestation thereof. Gene-based enrichment and association analysis performed using meta-analysis summary statistics in MAGMA and gene expression data in GTEx identified *FLT3* as significant after correcting for 18,079 gene tests. No gene sets were significantly associated with TS after multiple testing correction, and the only adult human tissue associating with TS after multiple testing correction was the dorsolateral prefrontal cortex.

Tsetsos et al. (2021) examined microarray genotype data from the former GWAS studies, looking at 3,581 TS cases and 7,682 ancestry-matched controls, investigating associations of TS with sets of genes expressed in neurons, glia, and neuronal-related cells. The analysis resulted in identification of four sets: cell adhesion and trans-synaptic signaling (identified twice by two different methods), the ligand-gated ion channel signaling, and the lymphocytic set. Genes within these sets have previously been associated with cognitive performance, depression, anxiety, Asperger's syndrome, eating disorders, and bipolar disorder with schizophrenia, and several other phenotypes known to be either co-occurring with or disrupted in TS.

The first genome wide CNV study in TS was reported in 2010 (Sundaram et al., 2010). Total of 307 CNVs were identified in 111 TS cases, and 216 CNVs were identified in 73 controls. There was no difference in overall CNV burden, or by CNV type (deletions vs. duplications). CNVs of all sizes were examined, however filtering criteria in CNV calling involved coverage of at least 10 markers which limits the ability to look at small CNVs - such filtering was necessary to reduce rates of false positive CNV calls. A noteworthy limitation of this study is the lack of diagnoses of commonly comorbid disorders, making it difficult to distinguish between TS and pleiotropic CNVs. In addition to small sample size complicating inferences about burden of observed CNVs, the case-control design limits the ability to infer if CNVs are inherited (and transmitted with TS) or *de novo*.

A subsequent study of 460 TS cases (148 trios) and 1,131 controls (436 trios) reported on rare CNVs (Fernandez et al, 2012). There was no difference in overall CNV burden in cases vs. controls, or by CNV etiology (inherited vs. *de novo*). However, cases were marked by larger and more gene rich CNVs, albeit the differences were statistically not significant. A total of 745 rare, high confidence CNVs in cases and 1,910 in controls were found. One major limitation of this study is the small sample size, which was further exacerbated by the removal of 185 probands during QC and ancestry matching.

Nag et al. (2013) published a study on Latin population including 210 TS cases and 285 controls found a significant excess of large CNV calls (> 500kb; $p = 0.006$). Among those large CNVs were also *NRXN1* and *COL8A1*. In addition to CNV changes in those regions, authors also found chromosomal rearrangements using multiplex

ligation-dependent probe amplification (the same technique was used to validate the CNVs in those two genes). Subsequent analyses of parents found that *COL8A1* was inherited whereas *NRXN1* was a *de novo* variant. However, this conclusion is limited since only a small subset of parents was available for testing. Additionally, parents were only tested for *COL8A1* and *NRXN1* using a low throughput method.

Bertelsen et al. (2016) looked specifically at *AADAC* deletion in a large European cohort of 1,181 TS cases and 118,730 controls and found an increased *AADAC* association with TS in the final meta-analysis (OR = 1.9; $p = 4.4 \times 10^{-4}$).

The very first large-scale, high-throughput case-control study of CNVs in TS involved 2,434 cases and 4,093 controls (Huang et al., 2017). CNVs were called using PennCNV and QuantiSNP, resulting in 9,375 rare CNV calls. Genic CNVs (n = 4,604) showed a significant but modest increase in burden for CNV count (OR = 1.05; p = 0.027), CNV gene count (OR = 1.09; p = 0.019), and CNV length (OR = 1.15; $p = 1.9 \times 10^{-4}$) in TS cases vs. controls. The highest burden was attributable to large CNVs, > 1Mb (OR = 1.26; $p = 5.3 \times 10^{-3}$), and singleton CNVs (OR = 1.13; $p = 2.9 \times 10^{-3}$). Further stratifying CNVs by their pathogenicity (according to the American College of Medical Genetics guidelines, as published by Riggs et al., 2020) showed a higher burden of pathogenic CNVs (OR = 3.03; $p = 1.5 \times 10^{-5}$), particularly when it comes to CNV deletions (OR = 3.94; $p = 6.3 \times 10^{-4}$). Further analysis revealed a high burden of *NRXN1* deletions (OR = 20.3, $p = 8.5 \times 10^{-4}$) and *CNTN6* duplications (OR = 10.1, $p = 8.3 \times 10^{-3}$). Pair-matching cases with their closest ancestry matched controls revealed that these results were not due to inter-European population stratification. There are several limitations to this study, including the fact it was a case-control design which

limits discerning between *de novo* and inherited CNVs, a sample size that is still too low to detect rare CNVs or those with moderate effect sizes, and limited data on comorbidities (only information on attention deficit/hyperactivity disorder, ADHD, and OCD were available).

Whole exome sequencing (WES) analysis of 789 TS trios and 1,136 quartets from the Simon Simplex Cohort (SSC) resulted in identification of 27 *de novo* CNVs (Wang et al., 2018). Incidence rate ratios (IRRs) were increased in two independent sets of TS samples and in the combined dataset: $IRR_1 = IRR_2 = IRR_{12} = 2.2$ ($p_1 = 0.004$; $p_2 = 0.024$; $p_{12} = 0.0025$). Further analysis found that both *de novo* deletions (IRR = 2.13; p = 0.04) and duplications (IRR = 2.25; p = 0.015) are risk factors for TS. There was also an increased rate of *de novo* CNVs in ASD probands in the SSC sample, and the rate of *de novo* CNVs did not differ between ASD and TS (rate ratio of *de novo* CNV burden among TS probands to SSC probands was 1.10, but not significant at p = 0.83). A separate analysis of 412 TS trios and 763 SSC quartets using microarray data replicated the finding of an increased CNV burden (but not the specific CNVs) from the WES experiment. There was an increased burden of *de novo* CNVs in TS samples (IRR = 2.8; p = 0.024), specifically in *de novo* CNV deletions (IRR = 3.8; p = 0.02). Similarly, the rates were increased in ASD, but no difference in rates between ASD and TS were observed (rate ratio of *de novo* CNV burden among TS probands to SSC probands was 0.89, but not significant at p = 0.63). Further analysis revealed that 46.3% of the *de novo* CNVs identified in this study were associated with TS and that 1.5% of TS cases carried such variants. Cross-disorder comparison has revealed that *de novo* CNVs observed in TS have also been observed in other disorders like ASD, SCZ, and

epilepsy. While this study could resolve the origin of the CNVs (*de novo* vs. inherited) and had a decent sample size, it was still underpowered to delineate CNVs with moderate effects.

The last CNV study to be briefly discussed is a discordant twin-pair WES study which only involved a single family where the father and proband (but not their monozygotic twin) had TS (Vadgama et al., 2019). This study found a single CNV duplication spanning *TOP3B* and *NLGN1* genes of the father and the discordant affected monozygotic twin.

So far, there have been 8 different studies that examined genetic variation in TS using high throughput sequencing approaches, specifically looking at WES and WGS. One of the first studies to do that looked at the exome of a 10-member, 3-generation pedigree where 7 members had diagnosed TS/PMVT (Sundaram et al., 2011). The authors report 3 novel, nonsynonymous single nucleotide variants (SNVs) in *MRPL3*, *DNAJC13*, and *OFCC1* genes segregating with chronic tic (CT) phenotype, but not present in controls and either dbSNP or 1000 Genomes databases (Cukier et al., 2014). Willsey et al. (2017) conducted a WES on two independent TS cohorts and found *de novo* likely gene disrupting variants are present in 5% cases, and 11.6% of the cases carried a *de novo* damaging (although not necessarily gene disrupting) variant contributing to TS risk. Maximum likelihood estimation (MLE) predicted that about 420 genes were contributing to TS risk. Four genes were significantly enriched for probably damaging missense and likely gene disrupting variants, one of which was classified as a high-confidence TS gene (*WWC1*) and three as probable TS risk genes (*CELSR3*,

NIPBL, and *FN1*). Additional MLE analysis further illustrated the importance of sample sizes in genetic studies of complex phenotypes (visualized on Figure 2-1).

A small-scale study of a 3-generation, 9-member multiplex family recruited through TIC Genetics cohort reported a rate heterozygous nonsense mutation in *PKND* co-segregating with TS phenotype (Sun et al., 2017). A previously discussed study by Wang et al. (2018) has also reported *de novo* sequence variants in TS, identifying 2 high-confidence TS genes (validating *WWC1*, in addition to *CELSR3*) and 4 probable TS risk genes (validating *NIPBL* and *FN1*, in addition to *OPA1* and *FBN2*). A smaller trio study looking at 97 TS trios and a replication sample of 524 TS cases has suggested *ASH1L* as a susceptibility gene in TS (Liu et al., 2019). A small WES study of 15 TS trios has found 25 coding *de novo* variants (Zhao et al., 2020). One of the affected genes in one TS proband was *CELSR3*, a high-confidence TS gene identified in previous studies. A small-scale WES study of Chinese Han families with TS has resulted in identification of *CLCN2* as a potentially important, but ultimately not statistically significant, gene in TS (Yuan et al., 2020).

Although there is a WGS study planned by the PGC-TSOCD workgroup, there have been no large-scale whole genome sequencing studies on TS published so far. This represents a significant gap in knowledge, particularly around the importance of intergenic variation on risk of TS. It is furthermore important to note that there is a substantial overlap in samples in a lot of studies discussed above, which limits the potential for meta-analysis of the results. Nonetheless, Table 2-1 summarizes the most important findings from genomic probes into TS genetics, with OR estimates reported where available.

A brief gene ontology (GO) analysis of the genes identified across the genomic studies of TS (as listed in Table 2-1), has shown enrichment among genes involved in embryonic and anatomical structure morphogenesis and developmental biological processes; extracellular matrix function, organization, and structure; integrin signaling pathways. Table 2-2 shows a complete list of GO terms, as well as fold-enrichment, direction, and false discovery rate (FDR) corrections for p-values. Method for GO analysis is described in Chapter 5 section Methods, subsection Gene Ontology Analysis.

While a substantial amount of effort has been put into genetic analysis of TS, most of the studies are still limited in statistical power due to small sample sizes. Recurrent observation of extracellular matrix and cell structure genes highlights the developmental nature of TS. Continued recruitments and expanded collaborative networks will play a substantially important role in driving further discoveries of novel variants associating with the TS.

Obsessive-Compulsive Disorder

As is the case with TS, the earliest genetic studies of OCD were also primarily twin and family studies. One of the earliest studies that examined multiple twin pairs in 1936 reported three cases of monozygotic twin-pairs presenting with OCD-like symptomatology. Two of those pairs had similar severity of OCD-like symptoms, with one of those pairs being twins raised apart. In the third pair, one twin had more severe and chronic OCD-like symptoms, whereas the other twin only had an acute episode of contamination symptoms (Lewis, 1936). A subsequent, larger study of twin pairs has found OCD to be present in both twins in 10/11 instances (Tienari, 1963). Subsequent studies have found an 80% concordance rate among monozygotic twins and a 20%

concordance rate among dizygotic twins (Inouye, 1965); and an 87% concordance among monozygotic, and 47% concordance among dizygotic twins (Carey & Gottesman, 1981). A large twin study examining incidence and cross-disorder instances of OCD, tics, and anxiety in 854 pairs has reported tetrachoric correlations. For OCD, the cross-twin tetrachoric correlations were 0.57 and 0.22, for monozygotic and dizygotic twins, respectively. For tics, the cross-twin tetrachoric correlations were 0.64 and 0.33 for monozygotic and dizygotic twins, respectively. Ultimately, for coinciding OCD and tics, the cross-twin cross-trait tetrachoric correlations were 0.25, 0.19, and 0.28 for the whole sample, monozygotic pairs, and dizygotic pairs, respectively (Bolton et al., 2007). More recent large-sample, cross-cultural twin studies of over 4,200 twin pairs have estimated OCD and OC behaviors heritability to be at 0.55 and 0.65, respectively (Eley et al., 2003; Hudziak et al., 2004).

In a family study of 145 first-degree relatives (ascertained on 46 probands with OCD), 30% of probands had at least one first-degree relative with OCD: 25% of fathers and 9% of mothers received this diagnosis, and 13% of both mothers and fathers had subclinical obsessive-compulsive behaviors (Lenane et al., 1990). A case-control family history analysis found a marked increase of lifetime OCD prevalence among first-degree relatives in cases vs. controls, 11.7% vs. 2.7%, respectively (Nestadt et al., 2000). In a small family study ascertained on 7 OCD probands and 65 relatives, 49.2% of the relatives had a diagnosis of OCD, 42.9% of the probands had co-occurring TS, and 78.1% of the relatives with OCD had a history of tics compared to none of the relatives without OCD (Hanna et al., 2002). A subsequent, slightly larger study of 106 probands with OCD, 44 control probands, and their 465 first-degree relatives has found

high rates of comorbid TS (33%) and PMVT (13.2%). Case relatives had a higher risk of OCD (OR = 32.5) and PMVT (OR = 7.9). The same study found that childhood onset OCD was indicative of higher OCD prevalence among first-degree relatives, and potentially higher genetic loading (do Rosario-Campos et al., 2005). Another study family study of 144 OCD probands has found 44% comorbidity with tic disorders among the probands, 17% rate of OCD and 12% rate of tic disorders among the relatives, with 32.6% of the probands having a relative with OCD (Chabane et al., 2005). The occurrence of tics among OCD probands and relatives is fairly common, another study of 100 OCD probands has reported the first-degree relative rates of 10.3% of OCD, 7.9% of subclinical OCD, and 4.6% of tic disorders. Among the probands who had tic disorders in addition to OCD, the reported first-degree relative rates are 18.2% for OCD and 4.6% for tic disorders (Pauls et al., 1995). This sampling of numerous family aggregation and segregation, and twin studies of OCD highlights an important genetic component to this disorder.

Several genome-wide linkage studies have been conducted to analyze segregation patterns in OCD families resulting in several susceptibility loci, but no specific genes have been identified using this methodology (Brett et al., 1995; Weissbecker et al., 1989; Pauls, 2010; Nestadt et al., 2000; Hanna et al., 2005; Shugart et al., 2006; Mathews et al., 2012; Nestadt et al., 2011). However, over 80 candidate gene studies have been examined, mostly focusing on serotonin transporters and receptors, tryptophan hydroxylase (involved in serotonin synthesis), dopamine transporters and receptors, and a few more neurobiologically relevant genes. As was the case with TS, out of numerous attempted studies on candidate genes, no instances

were ever successfully validated, particularly in large-scale genome-wide studies (Brett et al., 1995; Weissbecker et al., 1989; Pauls, 2010; Nestadt et al., 2000; Hanna et al., 2005; Shugart et al., 2006; Mathews et al., 2012; Nestadt et al., 2011; Comings et al., 1993; Billett et al., 1997; Nicolini et al., 1996; Billett et al., 1998). This lack of replicable results further illustrates the need for larger sample sizes and genome-wide approaches, as well as attention to comorbidities and endophenotypes (Altemus et al., 1996).

Despite generating an extensive amount of research on the genetic etiology of OCD, the risk-conferring variants remain elusive. As is the case with many other complex disorders, detangling complex genetic underpinnings will require large sample sizes and population stratified, genome-wide resolution approaches. The limitations of early approaches reliant on traditional molecular biology techniques have been somewhat resolved with modern population-stratification-conscious statistical genomic approaches, which are less vulnerable to false positive rates, such as GWAS studies.

There have been several GWAS studies reported for OCD. The very first collaborative effort by the International OCD Foundation Genetics Collaborative (IOCDF-GC) was published in 2013 (Stewart et al., 2013). After QC, 1,465 cases, 5,557 ancestry-matched controls, and 400 complete trios were analyzed on 469,410 autosomal and 9,657 X-chromosome SNPs. Albeit no genome-wide significant hits were found in the case-control sample, the two lowest p-value hits were located within a single gene, *DLGAP*. Analysis specific to the trio sample found one genome-wide significant SNP located 90kb upstream of *BTBD3* ($p = 3.84 \times 10^{-8}$). Meta-analysis of trio and case-control data resulted in no significant associations. Attempts to validate

putative OCD linkage regions and 22 candidate genes resulted in no significant hits either, thus failing to validate regions and genes in this sample. Significant enrichment was found in frontal eQTL ($p = 0.001$), cerebellar eQTLs ($p = 0.033$), and parietal eQTLs ($p = 0.003$). Additionally, significant enrichment was found in cerebellar mQTLs ($p < 0.001$). Additional miRNA and pathway analyses were conducted; however, no evidence of enrichment was found after multiple testing corrections.

The second GWAS study was published shortly thereafter (Mattheisen et al., 2014). A total of 2,895 samples from OCGAS have passed QC pipeline, 1,406 of which were OCD cases. Additional 192 cases without any genotyped family members were included, as well as 1,984 unrelated controls from a previously reported study on Parkinson's disease (Hamza et al., 2010). Overall, 549,123 autosomal SNPs were included in the analysis after QC. No markers were significantly associated with OCD at genome-wide significance level. The marker with the lowest p-value was located on chromosome 9, 1.3Mb downstream of the *PTPRD* ($p = 4.13 \times 10^{-7}$). Gene-based analysis for 21,567 protein-coding genes and miRNAs in two significant hits: *C16orf88* ($p = 1.94 \times 10^{-7}$) and *IQCK* ($p = 1.94 \times 10^{7}$). A query of interactome revealed 16 interacting genes for *DLGAP1* and 14 interacting genes for *GRIK2*, notably *GRIK2* was identified as an interactor for *DLGAP1*. The authors also attempted to conduct additional miRNA enrichment analysis; however, no significant results were reported.

The two GWAS studies, independently published by the two independent OCD consortia: IOCDF-GC and OCGAS, have been meta-analyzed together (IOCDF-GC & OCGAS, 2018). The resulting sample of 2,688 cases and 7,037 controls were analyzed across 8,693,187 markers. No SNPs were significant at genome-wide level, but 29 LD-

independent SNPs were observed with $p < 10^{-5}$. PRS scores were derived for OCD in each of the samples, and subsequently tested in the other sample. IOCDF-GC samples reasonably predicted case-control status in OCGAS sample, at p = 0.003 and explaining 0.9% of the phenotypic variance. Similarly, OCGAS samples reasonably predicted case-control status in the European IOCDF-GC sample, at p = 0.0009 and explaining 0.9% of the phenotypic variance. Genome-wide complex trait analysis (GCTA) of the samples revealed heritability estimates of $h_{SNP}^2 = 0.32$, $h_{SNP}^2 = 0.25$, and $h_{SNP}^2 = 0.25$ for IOCDF-GC, OCGAS, and combined sample, respectively. Linkage disequilibrium score regression on the combined sample resulted in the heritability estimate of $h_{SNP}^2 = 0.28$. Additionally, analysis of genetic correlation between the two samples resulted in $r_g = 0.83$; SE = 0.28; p = 0.003. Ultimately, the largest proportion of heritability was attributed to the highest allele frequency bins (MAF > 0.4), further substantiating the polygenic nature of OCD and significant contribution of common variants to its expression.

Another GWAS study looked at the sex-specific differences in the genetic architecture of OCD, utilizing genotyping data from the previous two GWAS studies (Khramtsova et al., 2019). The final sample included 1,249 male cases, 2,789 male controls, 2,774 female cases, and 7,096 female controls. While no individual SNPs were significant at the genome-wide significant level, there were two significant genes at the genome-wide significance level as computed by MAGMA, these genes were female-specific. They include *GRID2* ($p_F = 1.07 \times 10^{-7}$; $p_M = 7.23 \times 10^{-1}$) and *GPR135 (*$p_F = 1.55 \times 10^{-6}$; $p_M = 7.04 \times 10^{-1}$). The difference in LDSC (linkage disequilibrium score regression) heritability estimates between males and females was not significant, and

the genetic correlation between sexes was substantial ($h_{SNP,\,M}^2$ = 0.131; SE_M = 0.097; $h_{SNP,\,F}^2$ = 0.296; SE_F = 0.079; h_g^2 = 1.043; SE_g = 0.509; p_g = 0.041). Restricted maximum likelihood (REML) method yielded similar results for heritability estimates and genetic correlation. Furthermore, REML estimation of X chromosome-specific heritability resulted in h_X^2 = 0.01 (SE = 0.005; p = 0.006; 3.8% of total OCD heritability). X chromosome heritability in females was not significantly different from that in males. Sex differentiated effects have shown significant enrichment in for eQTLs from CD4+ T cells, the combination of immune tissues, and combined brain tissues excluding the functionally distinct cerebellum. No significant enrichments in eQTLs were found for SNPs. There was little overlap and no enrichment for anthropometric sex differentiated effects.

OCD is not a homogenous disorder and there is a substantial variability in OCD severity among the probands. A recent study published in 2020 looked at quantitative GWAS of OCD severity (Alemany-Navarro, Cruz, Real, Segalàs, Bertolín, Baenas et al., 2020). The Yale-Brown Obsessive Compulsive Scale (Y-BOCS) was used to measure OCD severity. Total of 376 European ancestry Spanish OCD cases and 258,937 SNPs passed QC and were included in the study. No SNPs were significant at the genome-wide level or have passed a multiple testing correction. There were 6 SNPs associated with OCD severity at a less stringent level, $p < 10^{-5}$. No genes were significantly associating after multiple testing correction, but the lowest p-value hits included *SLC8A1*, *MAP4K4*, *WTAP*, and *SLC22A10*. Functional annotation resulted in significant enrichment of porin activity, transmembrane, and *MAPK* signaling categories.

The same lab published another GWAS study simultaneously, looking at genetic determinants of dimensional variability in OCD (Alemany-Navarro, Cruz, Real, Segalàs, Bertolín, Rabionet et al., 2020). A total of 376 European ancestry Spanish OCD cases and 258,937 SNPs passed QC. Data were collected using the Dimensional Y-BOCS, including the following dimensions: aggression, contamination, order, hoarding, and sexual/religious. No SNPs were significant at the genome-wide level; however, some significant hits did occur at the lower level of stringency ($p < 10^{-5}$) for the aggressive, contamination, order, and hoarding dimensions. Gene-based associations revealed a genome-significant gene associating with the hoarding dimension (*SETD3*, $p = 1.9 \times 10^{-9}$), and a gene associating with the aggressive dimension (*CPE*, $p = 4.4 \times 10^{-6}$).

In addition to the genetic studies of OCD, it is important to make note of dimensional GWASes of OCD and obsessive-compulsive symptoms. Burton et al. (2021) reported the first validated genome-wide significant variant for OC traits (*PTPRD*), based on obsessive-compulsive symptom scores derived from TOCS in the Spit for Science sample. They also report a high, albeit not significant, association between TOCS obsessive-compulsive symptom score and OCD ($r_g = 0.71$, $p = 0.062$). CBCL obsessive-compulsive symptom scores analyzed in the same study resulted in no genome-wide significant associations, although the top-scoring variant was in the same direction and of similar effect size to that identified in the TOCS analysis.

Only two studies to date have looked specifically at CNVs in OCD (additional cross-disorder studies will be discussed in the next chapter). The first study, published in 2016, looked at 307 cases of unrelated idiopathic OCD, including 174 from parent-child trios, and 3,861 ancestry matched controls via WES (Gazzellone et al. 2016).

Among trios, 4 probands were found to carry a *de novo* CNV. While authors reported cases with CNVs in genes that have been associated with other psychiatric disorders like schizophrenia, ASD, and fragile X syndrome, burden analyses revealed no statistically significant associations.

The second study looked at 121 pediatric OCD cases and 124 random controls to identify rare CNVs (Grünblatt et al., 2017). They did not find a significant burden associated with frequency or size of rare CNVs. However, further stratifying the data by expression patterns revealed significantly higher frequency of rare CNVs affecting brain related genes (OR = 1.98, p = 0.0398), especially deletions (OR = 3.61, p = 0.021), in patients. Furthermore, enrichment analysis of gene content confirmed clustering of genes involved in synaptic/brain related pathways in patients, but not in controls. Two of the patients carried CNVs previously associated with different neurodevelopmental disorders, including *NRXN1*.

There have been 3 studies of WES in OCD, one of which focused primarily on CNVs and was thus discussed in the previous subsection. The second WES study examined 17 OCD simplex trios (Cappi et al., 2016). Out of the genes carrying single nucleotide variants (SNVs) identified by the WES, 3 have been previously identified in CNV studies in TS (*NDE1*, *SLC35G5*, and *WWP1*) and OCD (*VCX2* and *NDE1*) cases. Functional network enrichment implicated embryonic development, cell-to-cell signaling, cell death and survival, and cellular function and maintenance. No significant overlap was found with other neurodevelopmental disorders.

Recently, a more comprehensive WES study by the same lab provided more insight into SNV burden in OCD (Cappi et al., 2020). In total, 184 OCD trios and 777

unaffected trios from the SSC cohort were used to compare contributions of inherited and *de novo* SNVs in cases and controls. Overall, 22% of cases carry a damaging *de novo* mutation contributing to OCD risk. Recurrent SNVs have been found in *SCUBE1* and *CHD8* which were subsequently classified as high-confidence OCD risk genes, based on transmission and de novo association (TADA) algorithm (He et al., 2013). MLE analysis estimated that about 335 genes contributed to OCD risk. Power calculations have yielded projected discovery of high-confidence and probable risk OCD genes with varying numbers of trios (visualized on Figure 2-1), indicating at least 500 trios necessary for discovery of probable risk genes.

Just as is the case with TS, no large-scale WGS studies of OCD were reported insofar, leaving a large gap in knowledge that needs to be addressed.

Table 2-1. Summary of results from genomic studies of TS.

Study	Type	Gene	OR	P
Genome-Wide Association Studies				
Scharf et al. (2013)	SNP	COL27A1	1.29	1.9×10^{-6}
Yu et al. (2019)	SNP	FLT3	1.16	2.1×10^{-8}
Copy-Number Variant Analyses				
Nag et al. (2013)	Deletion	NRXN1	-	3.0×10^{-2}
	Duplication	COL8AA1	-	4.0×10^{-3}
Huang et al. (2017)	Deletion	NRXN1	20.30	8.5×10^{-4}
	Duplication	CNTN6	10.10	8.3×10^{-3}
Whole Exome Sequencing Analysis				
Willsey et al. (2017)	SNV	WWC1	-	7.3×10^{-6}
	SNV	CELSR3	-	1.7×10^{-5}
	SNV	NIPBL	-	5.4×10^{-5}
	SNV	FN1	-	5.9×10^{-5}
Wang et al. (2018)	SNV	WWC1	-	1.9×10^{-5}
	SNV	CELSR3	-	2.2×10^{-5}
	SNV	OPA1	-	6.7×10^{-5}
	SNV	NIPBL	-	1.1×10^{-4}
	SNV	FN1	-	1.2×10^{-4}
	SNV	FBN2	-	1.3×10^{-4}
Liu et al. (2019)	SNV	ASH1L	-	1.4×10^{-3}

Note: The summary only includes genes identified from genome-wide approaches.

Table 2-2. GO enrichment analysis of previously identified important TS genes.

GO Term	Fold-enrichment	+/-	P_{FDR}
Biological Process			
Anatomical structure morphogenesis	7.07	+	4.4×10^{-3}
System development	4.61	+	1.3×10^{-2}
Uterus morphogenesis	> 100.00	+	2.2×10^{-2}
Multicellular organism development	4.16	+	2.4×10^{-2}
Embryonic morphogenesis	14.64	+	3.9×10^{-2}
Molecular Function			
Extracellular matrix molecular constituent	39.67	+	1.2×10^{-2}
Cellular Component			
Endoplasmic reticulum lumen	2.00	+	4.8×10^{-2}
Pathways			
Integrin signalling pathways	25.74	+	3.0×10^{-2}
Protein Class			
Extracellular matrix structural protein	71.49	+	9.0×10^{-4}
Extracellular matrix protein	48.33	+	2.1×10^{-4}
Reactome Pathways			
Extracellular matrix organization	28.69	+	1.2×10^{-3}

Note: Table shows only the statistically significant enriched terms after Benjamini Hotchberg correction for multiple testing. Methods described in Chapter 5.

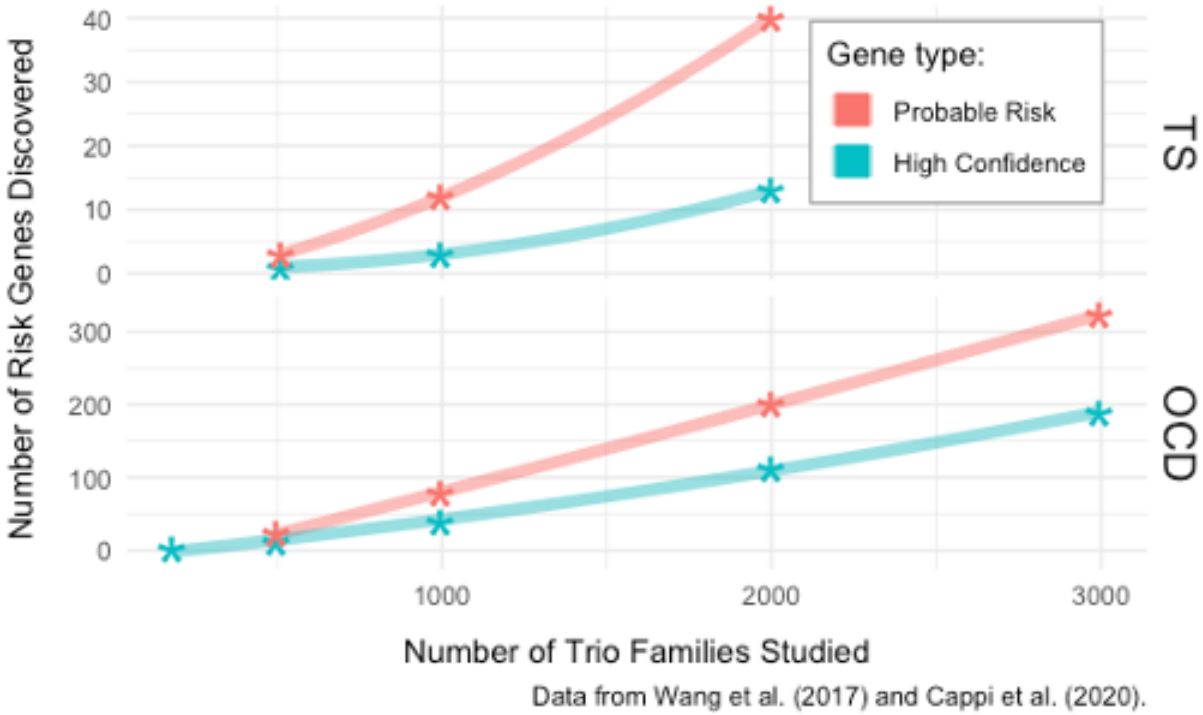

Figure 2-1. Theoretical discovery of high-confidence and probable TS (upper) and OCD (lower) risk genes for a given number of trio families. Data obtained from Wang et al. (2017) and Cappi et al. (2017) MLE simulations.

CHAPTER 3
COMPLEX RELATIONSHIP BETWEEN NEURODEVELOPMENTAL DISORDERS

Clinical Perspective

TS and OCD are both psychiatric disorders with a substantial developmental component. Phenotypically, TS and OCD share some characteristics, for example, premonitory urges to tic could be seen as a form of obsession (Reese et al., 2014). Furthermore, the complex phenotypic relationship between TS and OCD may be explained, in part, by highly heritable cross-disorder symptom clusters: symmetry and disinhibition (Darrow et al., 2017; Hirschtritt et al., 2016). Symmetry, generally understood as an OCD-related set of symptoms, includes symptoms like evening up, checking obsessions, ordering, and arranging, counting, etc., which also occur among individuals with TS (Rosario-Campos et al., 2006). Conversely, disinhibition is typically understood as TS-related set of symptoms and it includes obsessive urges to offend, mutilate, or be destructive, coprolalia (involuntary and repetitive use of obscene language), copropraxia (involuntary and repetitive use of obscene gestures), etc., which can also occur among individuals with OCD (Darrow et al., 2017). Darrow and colleagues (2017) have estimated heritability of these shared symptom clusters to be 0.38 and 0.35 for symmetry and disinhibition, respectively, which demonstrates that these symptom clusters are not only shared between disorders, but also have a significant genetic component indicating an underlying shared biological etiology.

Beyond shared phenotypic characteristics, TS and OCD are highly comorbid and usually exhibit evidence of familial clustering. For example, Hirschtritt et al. (2015) have found 66.4% of patients with TS to have comorbid OCD. OCD comorbidity was higher among females than males (70.6% vs. 64.4%, $p = 0.03$). Claudio-Campos et al. (2021)

found similar rates of comorbidity. Interestingly, rates of comorbid OCD were higher among TS cases than PMVT cases, 65.6% vs. 42.6% at $p < 0.001$, indicating OCD comorbidity may be a function of TD severity. A meta-analysis of OCD comorbidity among PMVT vs. TS cases in the same study found OR = 0.37, further solidifying evidence of OCD symptomatology and comorbidity being a function of the severe end of the TD spectrum.

A population-based study of Danish national health registries of 1,741,271 individuals born between 1980 and 2007 has found relative recurrence risk (RRR) of TD-OCD cross-disorder to be similar to OCD-OCD among the siblings (Browne et al., 2015). RRR for TD among individuals given older siblings with TD was 18.63 ($p < 0.001$), whereas RRR for OCD among the same individuals was 3.98 ($p < 0.001$). RRR for TD among individuals given older siblings with OCD was 4.88 ($p < 0.001$), RRR for OCD among same individuals was 4.89 ($p < 0.001$). RRR for TD among individuals with siblings who have both TD and OCD was 26.37 ($p < 0.001$), whereas RRR for OCD among same individuals was 10.90 ($p < 0.001$). Similar patterns of RRR were observed when examining parent-proband relationships as opposed to older sibling-proband relationships.

A family study of 222 TS-affected sib-pair families has found significant genetic correlations between TS-OCD and OCD-ADHD, but not between TS-ADHD (Mathews & Grados, 2011). Heritability analysis of OCRDs from this study has found heritabilites to be: $h_{TS}^2 = 0.32$, $h_{TD}^2 = 0.29$, $h_{OCD}^2 = 0.56$, $h_{ADHD}^2 = 0.68$, $h_{OCD+ADHD}^2 = 0.62$. Analysis of genetic and environmental correlations among these disorders has revealed genetic correlations to be stronger between TS-OCD and ADHD-OCD, whereas the

environmental correlation was stronger between TS-ADHD. Interestingly, the correlation between TD broadly defined and obsessive-compulsive symptoms (OCS) was 1.0, indicating a perfect correlation, but this relationship was not statistically significant. However, genetic correlations for TS-OCD were 0.92 ($p = 2 \times 10^{-5}$), OCD-ADHD was 0.63 ($p = 1 \times 10^{-4}$), and TS-OCS was 0.90 ($p = 7 \times 10^{-4}$). Significant environmental correlations were noticed between TS-ADHD at 0.66 ($p = 2 \times 10^{-3}$), and TS-OCS at 0.35 ($p = 1 \times 10^{-2}$). While a larger sample size would be beneficial, this paper clearly shows that family members of individuals with TS would be at an increased risk of TS, OCD, or some combined form of these two disorders.

The clinical overlap between these two disorders is so significant that some clinicians distinguish clinical presentations of comorbid TD and OCD as Tourettic OCD (TOCD). Specifically, individuals who are usually diagnosed with TOCD are those who present with OCD and some of the following symptomatology: (1) pronounced touching, tapping, and repeating behaviors that serve an identifiable function of relieving somatic discomfort or vague psychological distress; (2) a preoccupation with unrelenting discomfort for nonperformance of the repetitive behaviors; and (3) the presence of unelaborated obsessional themes (Mansueto & Keuler, 2007).

While conceptualization of TOCD might be clinically relevant and beneficial for the treatment purposes, further exploration of such an idea is necessary to determine whether TOCD symptomatology is simply a result of TD and OCD comorbidity. Additionally, genetic studies of such a construct would be helpful in determining whether TOCD and OCD are genetically divergent, and thus potentially a consequence of different underlying pathophysiological processes. Nonetheless, there is clear evidence

of co-occurrence of symptoms traditionally conceptualized as TD and OCD. Furthermore, there is clear evidence of clinical overlap in symptomatology and familial aggregation which indicates that such shared symptomatology might result from shared genetic factors. Recently, several genetic studies have examined such relationships, and will be discussed in the next section.

Genetic Perspective

Motivated by clinical findings and reports, genetic studies of psychiatric disorders often analyze genetic correlations between such related disorders. Examples of such paired psychopathology include schizophrenia and bipolar disorder, ADHD and ASD, depression and anxiety, and TD and OCD.

While OCD and TD do share some of the underlying genomic architecture, they also have their distinct components (Yu et al., 2015). Additionally, comorbid TD-OCD may have distinct underlying genetic susceptibility compared with OCD on its own. The study examined 1,310 OCD cases, 834 TS cases, 579 TS-OCD cases, 290 OCD trios, and 5,667 controls. No genome-wide significant loci were found in the GWAS analysis. PRS analysis of this sample has revealed that pure OCD (OCD without TD) was better at predicting OCD cases (measured by Nagelkerke's pseudo R^2), at $R_N^2 = 0.032$, compared to all OCD cases at $R_N^2 = 0.021$. When the discovery sample was a combination of OCD and TD cases, this association was even smaller at $R_N^2 = 0.017$. TS cases alone were not as good at predicting OCD case status, at $R_N^2 = 0.0004$ Conversely, pure OCD was not very good at predicting TS case status, at $R_N^2 = -0.012$. Combined OCD and TD and TS alone performed similarly, at $R_N^2 = 0.002$ and $R_N^2 = 0.006$, respectively. All associations except pure OCD → OCD, OCD → OCD, combined TD and OCD → OCD were not statistically significant. Potential explanation

for poor performance of TS predictions could be imbalance between OCD and TS cases in the sample: 1,600 pure OCD cases vs. 834 TS cases. Notwithstanding, this paper shows OCD with and without chronic tics may have different genetic architectures.

The largest cross-disorder GWAS to date was a 2019 report from the Cross-Disorder Group of the PGC (Lee et al., 2019). This collaborative effort examined 232,964 cases, 494,162 controls, and 6,786,993 SNPs from genome-wide studies of anorexia nervosa (AN), ADHD, ASD, bipolar disorder (BPD), major depression (MDD), OCD, schizophrenia (SCZ), and TS. The notable genetic correlations include SCZ-BPD (r_g = 0.70), OCD-AN (r_g = 0.50), MDD-ASD (r_g = 0.45), MDD-ADHD (r_g = 0.44), and OCD-TS (r_g = 0.41). Network analysis of the correlations reveals three disorder clusters: SCZ-BPD, TS-OCD-AN, and MDD-ADH-ASD. Cross-disorder meta-analysis revealed 136 LD-independent loci at genome-wide significance level, 101 overlap with previously reported loci, while 35 of them represent novel genome-wide significant associations. Further analysis concluded that the number of cross-disorder associations was correlated to the statistical power as given by the following formula:

$$B_{A-B} = \sqrt{N_A \times N_B \times h_A^2 \times h_B^2 \times r_{g,A-B}^2}\,, \tag{3-1}$$

where $B_{A\text{-}B}$ stands for power of cross-disorder association test between disorders A and B, N_A and N_B stand for sample size in GWASes of disorders A and B, h_A^2 and h_B^2 stand for narrow-sense heritability of disorders A and B, and $r_{g,\,A\text{-}B}^2$ stands for genetic correlation between the two traits. Psychiatric disorder-associated loci have shown a significant enrichment in genes expressed in pituitary and all brain tissues in the GTEx. Genes mapped to the 146 risk loci show higher expression values in neurons and oligodendrocytes, with much higher neuronal specificity for pleiotropic loci. GO analysis

(described in Chapter 5 methods section) suggests involvement of pleiotropic loci in neurodevelopmental processes, with enrichment in genes involved in neurogenesis, regulation of nervous- system development, neuron differentiation, and specific neurotransmitter-related pathways including glutamate receptor signaling and voltage-gated calcium channel complex. Pleiotropic risk genes were also enriched in cortical glutamatergic neurons, providing further support for the involvement of glutamate receptor signaling pathways in etiopathogenesis of various neuropsychiatric disorders. More than 41% of the genes in these pleiotropic loci were found to be intolerant to the loss-of-function mutation (exceeding chance occurrence with Fisher's exact test $p = 4.9 \times 10^{-8}$). No significant differences in spatiotemporal expressions were found. Pleiotropic risk loci were also enriched among genes associated with neuroticism, cognitive ability, night sleep phenotypes, and BMI (supporting previous suggestions of association between psychiatric disorders and obesity).

Another large study examined shared heritability between 25 common brain disorders using data from 265,218 patients and 784,643 control participants (Anttila et al., 2018). Cross-disorder genetic associations revealed a high degree of relationship between TS with ADHD, MDD, and OCD, and between OCD with AN, BPD, MDD, SCZ, and TS. Comparison of psychiatric disorders and neurological phenotypes revealed similarities between OCD and TS, especially relating to (focal) epilepsy and migraines (with aura). However, TS and OCD had opposite correlations with educational attainment and cognitive performance (TS and OCD were negatively and positively correlated, respectively). Both disorders were characterized by neuroticism, depressive symptoms, and diminished subjective well-being.

A more recent analysis of OCRDs including OCD, TS, and ASD, as well as ADHD has found the highest pairwise genetic correlation to exist between TS and OCD (Yang et al., 2021). LDSC analysis of PGC GWAS summary statistics reports yielded the following genetic correlations: $r_{g,\ TS\text{-}OCD} = 0.38$ ($p = 2.00 \times 10^{-4}$), $r_{g,\ ADHD\text{-}ASD} = 0.35$ ($p = 1.33 \times 10^{-11}$), $r_{g,\ TS\text{-}ADHD} = 0.26$ ($p = 2.05 \times 10^{-5}$), $r_{g,\ TS\text{-}ASD} = 0.18$ ($p = 5.50 \times 10^{-3}$), $r_{g,\ ASD\text{-}OCD} = 0.12$ ($p = 1.50 \times 10^{-1}$), $r_{g,\ ADHD\text{-}OCD} = -0.17$ ($p = 2.00 \times 10^{-2}$). These findings highlight the importance of considering comorbid disorders and their relationships in genomic analyses, as discrete diagnostic boxes fail to capture the shared genetic risk.

While genetic relationships between OCD and TS are consistently substantial, both OCD and TS are found to share genetic risk factors with other disorders. A combined GWAS between OCD and ADHD with a final combined dataset consisting of 2,998 OCD samples and 5,415 ADHD samples yielded no genome-wide significant associations, however 0.08% overlap between genetic data in ADHD and OCD was noted (Ritter et al., 2017).

Another study looked at combined GWAS between OCD and ASD, with the combined OCD-ASD dataset consisting of 9,896 individuals (2,998 OCD cases, 6,898 ASD cases) also found no genome-wide significant marker associations (Guo et al., 2017). However, cross-disorder PRS approximated 0.11% of shared genetic variance between ASD and OCD.

A recent study reported a GWAS between OCD and AN, including 6,183 cases and 18,031 controls (Yilmaz et al., 2020). No variants were found to be significant at the genome-wide level. Heritability estimates were: $h_{AN}^2 = 0.18$, $h_{OCD}^2 = 0.29$, $h^{AN\text{-}OCD2} = 0.21$. The genetic correlation between the two disorders was reported at $r_g = 0.49$ (SE =

0.1, $p = 9.10 \times 10^{-7}$). Nineteen additional phenotypes have been significantly associated with AN-OCD shared genetic risk, notably: BPD, neuroticism, SCZ, and years of schooling. No significant enrichment was found in GTEx tissue types. No significant pathways were found. A single gene, KIT, was found to be significantly associated with AN-OCD.

So far, there has only been a single study focusing on cross-disorder CNV analysis between TS and OCD (McGrath et al., 2014). Sample consisted of 2,699 cases (1,613 OCD, 1,086 TS) and 1,789 controls. Parental data allowed a de novo analysis in 348 OCD trios. No global CNV burden was found in cross-disorder analysis or disease-specific analysis. There was a 3.3-fold increased burden of large deletions previously associated with other neurodevelopmental disorders. Regional analysis revealed 16p13.11 as an important locus in neurodevelopmental disorders and accounts for many called CNVs.

No high-throughput cross-disorder WES or WGS studies have been conducted to evaluate the shared and distinct genetic risks associated with above mentioned psychiatric disorders.

Cross-disorder genetic studies highlight the importance of phenotype definitions. In addition to considering comorbid disorders and genetic relationships between them, novel approaches that consider symptom dimensions in addition to diagnostic categories might yield to increased power of genetic studies of neurodevelopmental psychiatric disorders. These efforts, however, depend highly on the number of available large samples which have been thoroughly phenotyped.

CHAPTER 4
EXPLORATION OF OCRD PHENOTYPES IN ABCD STUDY

Background

OCD is highly heritable with a substantial genetic component; however, the genetic architecture of OCD is very complex, and little is known about exact causative genetic factors and pathways (Strom et al., 2021). Large-scale collaborative efforts and advancements in statistical methodology have driven progress in the field of psychiatric genomics, slowly unearthing numerous loci associating with psychiatric disorders and ushering the field into the era of polygenic and pleiotropic effects, and a paradigm shift towards complex genetic architecture as a probabilistic rather than deterministic factor in disorder etiology (Reynolds et al., 2021).

One of the main drivers of the novel discoveries, and power-driving factors in genome-wide association studies in general, is sample size. Collaborative approaches have led to aggregation of large datasets, or summary statistics of smaller datasets, which have been combined by meta-analyses to increase power and drive discoveries. Additionally, large, coordinated initiatives with wide phenotyping range such as the adolescent brain cognitive development (ABCD) study contribute data usable for psychiatric genomics, despite it not being the primary aim thereof (Bjork et al., 2017). However, simulation studies have shown that phenotype misclassification may present serious challenges to GWAS power and be as important if not more important than sample size (Manchia et al, 2013). When it comes to psychiatric disorders like OCD, where phenotype classification agreements between clinicians can be challenging, self- or parent-report-based data as those in the ABCD study can have extensive rates of

phenotype misclassification and negatively impact GWAS power (Freedman et al., 2013).

My initial epidemiological exploration of the ABCD study cohort revealed overrepresentation of OCD in the sample, 13.4% prevalence compared to the expected (reported) 2.3% prevalence in early adolescence (Zohar, 1999). Such inflation of OCD prevalence in a sample which was not ascertained for psychiatric disorders or OCD risk factors raises concerns of potentially high phenotype misclassification rates with high potential impact to secondary analyses, like GWAS. In such analyses, high rates of misclassification can be detrimental to test power and p-value inflation. However, the rich longitudinal data made available in the ABCD Study cohort provides an opportunity to explore various diagnostic refinement techniques to improve the accuracy of the psychiatric diagnoses. Because the ultimate goal in this dissertation is to conduct genetic analyses of OCD and tic disorders in ABCD Study, proper classification of those phenotypes is of utmost importance.

The ABCD Study utilizes a self-administered modified form of the computerized Kiddie Schedule for Affective Disorders and Schizophrenia for DSM-5 (KSADS-5) semi-structured interview to collect psychiatric symptom data and establish a diagnosis (Kaufman et al., 2021). The KSADS-5 is a tool developed primarily to be administered by clinicians, thus self(parent)-administration of the KSADS-5 in the ABCD study might explain such an unexpectedly high prevalence of OCD in the cohort. Table 4-1 summarizes disorder modules from the KSADS-5 by individuals who took it (parent or child) and the time-point of the study at which they have taken it. Figure 4-1 shows the

progress of ABCD study in terms of KSADS-5 data releases, given that actual data collections are ahead of the corresponding data release by about 1 year.

There are a few studies examining the validity of KSADS-5 when self-administered. Namely, Townsend et al. (2019), found substantial discrepancies between clinician-, parent-, and self-administered KSADS-5. Specifically, when it came to parents and youth, only in 63% of instances were they in agreement with respect to OCD diagnosis. This discrepancy was predominantly driven by a low positive agreement of 29%, whereas the negative agreement was relatively high at 85%. Youth were found to report OCD at a higher rate than parents, with endorsement frequency of OCD being 49.1% and 26.4%, respectively. These discrepancies were more severe when looking at clinicians and youth, where the rate of agreement was slightly higher at 66%, but it was driven mostly by a good negative agreement at 95% yet marked by a remarkably lower positive agreement of 18%. Unlike the youth, who endorsed OCD in 49.1% of instances, clinicians only endorsed OCD in 9.4% of instances. Ultimately, clinicians and parent discrepancies were lower, given agreement rate of 82%, which is mainly characterized by high negative agreement at 94% and low positive agreement at 26%. While parents had a closer rate of endorsement to clinicians than youth, they still endorsed OCD at 2.8-times higher rate. Youth endorsed OCD at a 5.2-times higher rate compared to clinicians. Since OCD diagnostic information in the ABCD study comes predominantly from parent-rated KSADS-5 data, the particular statistic of concern is low clinician-parent positive agreement at 26%. Due to such low agreement, it follows that using KSADS-5 data as reported in the ABCD data might be suboptimal in genetic analyses. Reliability of the KSADS-5 as it was originally designed to be used when it

comes to OCD diagnosis is quite high, with Cohen's kappa score of 0.74 (Young, 2010). However, when it comes to OCD specifically, KSADS-5 is not an extremely reliable tool when self-administered according to Townsend et al. (2020), with child-clinician Cohen's kappa of 0.15 (with youth overreporting), parent-clinician Cohen's kappa of 0.25 (parent overreporting), and parent-youth Cohen's kappa of 0.14 (youth overreporting). Self-report performance was much better when it comes to mood and anxiety disorders.

There are several inventories that measure OCD symptomatology and its dimensional psychopathology which could be used to further delineate which cases represent true OCD diagnosis and which are false-positive. A non-exhaustive list of them includes the Yale Brown Obsessive Compulsive Scale (YBOCS), the obsessive-compulsive inventory-revised (OCI-R), the dimensional obsessive-compulsive scale (DOCS), the Toronto obsessive-compulsive scale (TOCS), and the child behavioral checklist (CBCL). Notably, CBCL is not specifically an OCD scale, alebit OCD-specific questions are included in its global assessment of the childhood psychiatric symptoms.

The YBOCS (Goodman, 1989) is currently the most widely utilized inventory to capture severity of OCD without being influenced by the specific content of obsessions or compulsions. It consists of a symptom checklist, with between 30 and 90 items, depending on the version, and a clinician-administered 10-item symptom severity scale, with each item being rated on a 5-point scale. Five items assess obsession symptom severity, and five items assess compulsion symptom severity. Inter-rater reliability of the YBOCS severity portion on 40 patients by 4 raters was 0.97 for obsessions items, 0.96 for compulsions items, and 0.98 for overall inventory (Goodman, 1989).

Woody et al. (1995) conducted a study further characterizing the validity and reliability of the YBOCS inventory. Looking at the construct validity, total YBOCS score was not significantly influenced by participants' age, sex, or socioeconomic status. On subscale level, compulsions were positively associated with age (but not sex or socioeconomic status). Additionally, the YBOCS severity portion scores were unaffected by comorbid depression, further solidifying its ability to measure OCD-specific effects. However, obsession items were correlated with anxiety - indicating potential confounding. It is important to note that, traditionally and at the time of developing this inventory, OCD was considered an anxiety disorder - so by the sheer nature of the disorder definitions, as anxiety is a symptom of OCD, some overlap in phenomenology of the two would be expected. The DOCS is a 20-item measure that assesses the four dimensions of OC symptoms developed for use in both clinical and research settings (Abramowitz et al., 2010). Essentially a derivative of YBOCS, DOCS assesses severity of individual OCD symptoms.

The OCI-R is a revised form of the OCI inventory that was revised to eliminate redundant items, simplify the scoring across subscales, and reduce overlap across subscales (Foa et al., 2002; Foa et al., 1998). The TOCS is a 21-item multidimensional measure of OC traits in children and adolescents with sound psychometric properties, and a normal distribution in an unselected sample of children (Park et al., 2016). All of these inventories are suitable for OCD symptom assessment (and in case of DOCS, OCD diagnosis); however, they are not as widely utilized as the YBOCS. OCD is still underrepresented when it comes to the number, quality, and availability of inventories, batteries, or questionnaires, compared to other disorders. This has a substantial impact

on OCD diagnosis and screening in epidemiological studies, especially when it comes to large population screens where clinician-assisted testing is not a viable option. Furthermore, little is known about racial, ethnic, and socio-cultural impacts on validity and reliability of these measures - which may raise concerns about their performance in racially or ethnically diverse populations such as the ABCD Study. Regardless, none of these measures were administered to either parents or youth in the ABCD Study at the time of analysis, so they cannot be used to refine the OCD diagnosis in this sample. Instead, the ABCD Study used a broader dimensional tool, the child behavioral checklist (CBCL).

The CBCL is probably the most used tool of all of the aforementioned tools and is a caregiver report form aiming to identify problem behavior in children. It is a part of the Achenbach system of empirically based assessments (ASEBA), which primarily focuses on detecting behavioral and emotional problems in children and adolescents. Other tools in the ASEBA include a teacher report form and a youth self-report. Despite not being a tool designed to formally assess OCD, there are two questions in the CBCL that assess severity of obsessions and compulsions, specifically. The CBCL is designed to be completed by the parents, which eliminates the need for a trained mental health professional or clinical researcher, allowing for unhindered use in self-report-based research projects such as the ABCD study. Various OCD subscales have been derived from the CBCL to assess the predictive power of the CBCL as OCD diagnostic tools, yielding some success but not meeting the same level of success as dedicated OCD scales or semi-structured diagnostic interviews. For example, a 6-item CBCL subscale that included the two OCD and four additional anxiety related questions was found to

accurately identify 77% of OCD cases (Storch et al., 2006) – in this dissertation, I will refer to this subscale as the obsessive-compulsive problems (OCP) subscale.

Questionable reliability of self-report diagnoses, or diagnoses based on self-report data in large initiatives is not a new issue, and several approaches have been developed to address these problems. For example, studies in the brain health registry (BHR), which features predominantly self-report data for most participants and clinically determined diagnoses of various DSM-5 psychiatric disorders in a subset of participants who underwent direct assessments has shown that consistent self-reports of psychiatric illnesses are useful markers for clinician-diagnosed psychiatric illness (Sordo Vieira et al., 2022). In this study, incorporating the temporal consistency of self-report of a given psychiatric disorder diagnosis resulted in improved diagnostic precision (PPM_{MDD} = 0.85, PPV_{HD} = 1.00) and specificity (TNR_{MDD} = 0.92, TNR_{HD} = 1.00). Thus, I take this approach using the KSADS-derived OCD diagnoses in the ABCD sample to distinguish between individuals with high-confidence OCD diagnoses, to be classified as narrow OCD (nOCD), and individuals with low-confidence OCD diagnoses, to be classified as broad OCD (bOCD).

The ABCD study is a large scale, longitudinal brain development and child health study (n = 11,924), with rich phenotype data. The participants are recruited at the age of 9-10 and followed every 3-6 months until the age of 19-20 (Figure 4-1). There are numerous data collected for this cohort via in person and phone interviews, paper tests, iPad tasks, brain imaging, and biosamples. These data include neurocognitive and mental health assessments, as mentioned before. Data are housed by the National Institute of Mental Health (NIMH) Data Archive (NDA), where they are openly shared

with researchers upon request with approval of an institutional review board (IRB). Overall, the ABCD cohort consists of 11,924 samples (48% female). A K-SADS report is available for 11,877 samples (48% female).

The overarching goal for this chapter is to explore OCD and related phenotypes in the ABCD study to refine the diagnoses for subsequent genetic studies. I hypothesize that, due to lower rate of false-positive diagnoses in nOCD compared to bOCD, nOCD will be marked by higher rates of psychiatric comorbidity and more severe scores on CBCL individual item and subscale compared to bOCD.

Methods

Diagnoses

KSADS-5 reports diagnosis as a categorical variable with values 0 (absent) or 1 (present) for each disorder for both past and present. KSADS-5 reports based on parent self-administration (about their child) and youth self-administration (about themselves) were made available via NDA. Lifetime disorder diagnosis (LDx) is a Boolean or operation between past (Dx_{PAST}), present ($Dx_{PRESENT}$), and remission ($Dx_{REMISSION}$) diagnosis variables where available and can be mathematically represented according to Equation 4-1.

$$LDx = Dx_{PAST} \bigvee Dx_{PRESENT} \bigvee Dx_{REMISSION} . \quad (4\text{-}1)$$

LDx variable is equivalent to combined broad and narrow diagnoses, to be denoted with a bn prefix (e.g. bnOCD). To separate broad diagnoses from narrow, I first find a cumulative sum of lifetime diagnoses (CDx) for each participant across all time points where a module for a given disorder was administered (see Table 4-1 and Figure 4-1). This operation is mathematically represented as:

$$CDx = \sum_{t=1}^{T} LDx_t \,, \tag{4-2}$$

where t represents a time point iteration from the total number of timepoints with available disorder diagnoses T. Subsequently, if a given KSADS-5 module was administered to both youth and parents, I calculate the sum of lifetime diagnoses (TDx). If the module was administered only to parents or only to youth, then TDx is equivalent to parent or youth CDx, respectively. TDx can also be mathematically represented as:

$$TDx = CDx_{YOUTH} + CDx_{PARENTS} \,. \tag{4-3}$$

Ultimately, we can assign a final categorical diagnostic variable as diagnosis negative (coded as 0) if TDx = 0, broad diagnosis (coded as 1) if 0 < TDx ≤ T/2, and narrow diagnosis (coded as 2) if T/2 < TDx, where T/2 represents half of the total number of follow-ups in which the given disorder was assessed. This can also be mathematically represented as:

$$FDx = \begin{cases} 0, & TDx = 0 \\ 1, & 0 < TDx \leq \dfrac{T}{2} \\ 2, & \dfrac{T}{2} < TDx \end{cases} . \tag{4-4}$$

It is worth noting that this process is a modified form of that from Sordo Vieira et al. (2022), specifically their form of consistent longitudinal diagnosis endorsements termed “most.” For example, OCD module of KSADS-5 was only administered to parent at the baseline and year 2 follow-up (Table 4-1). Additionally, for OCD, only available diagnoses were OCD present and OCD past. Thus, according to the process outlined above, children whose parents endorsed OCD in either present or past in both baseline

and year 2 follow-up were labeled as nOCD cases. Conversely, children whose parents endorsed OCD in either present or past in either baseline or year 2 follow-up, but not both, were labeled as bOCD cases.

Diagnosis values for tic disorder (TD) were a bit different as TD diagnoses were not reported in the KSADS-5 report. The TD module of KSADS-5 was assessed only in year 3 follow-up and data are currently only available for about half of the entire ABCD cohort. Additionally, the TD module of KSADS-5 in the ABCD Study only reports diagnoses for other unspecified tic disorder, but not TS or PMVT. Thus, I utilize symptom-level (which corresponds to DSM-5 item-level) data to derive TD diagnosis values. Figure 4-2 shows a decision diagram for determining TD diagnosis. Briefly, lifetime symptom tics, both phonic/vocal and motor, are calculated in a similar fashion as LDx in Equation 4-1 and labeled as LSx, where Sx stands for symptoms. Subsequently, Dx for narrow TD is estimated by finding individuals who have both phonic/vocal and motor tics, as shown in Equation 4-5. Broad TD can be estimated according to Equation 4-6, identifying individuals who have ither phonic/vocal or motor tics, but not both. It is important to note that, unlike other disorders, TD was only assessed once so far, at year 3 follow-up (corresponding to the age of about 12), so temporal consistency of the symptoms was not assessable.

$$FDx_{TD} = LSx_{PHONIC\ TICS} \bigwedge LSx_{MOTOR\ TICS} \ . \tag{4-5}$$

$$FDx_{TD} = LSx_{PHONIC\ TICS} \bigvee LSx_{MOTOR\ TICS} \ . \tag{4-6}$$

Prevalence Rates

Prevalence rates are calculated as a ratio of the number of individuals with a disorder to the number of all individuals who were assessed for that disorder (Sullivan,

2017; Friis & Sellers, 2020). The following set of equations defines how prevalence rates for broad + narrow (p_{BN}) and narrow (p_N) diagnoses were calculated, together with their standard errors and 95% confidence intervals (using the formula for sample proportions confidence intervals).

$$\hat{p}_{BN} = \frac{\sum_{i=1}^{n} I_{BN}(FDx)}{n} \tag{4-7}$$

$$I_{BN}(FDx) = [FDx > 0] \tag{4-8}$$

$$SE[\hat{p}_{BN}] = \sqrt{\frac{\hat{p}_{BN}(1-\hat{p}_{BN})}{n}} \tag{4-9}$$

$$95\%\ CI[\hat{p}_{BN}] = \hat{p}_{BN} \pm 1.96 \times SE[\hat{p}_{BN}] \tag{4-10}$$

$$\hat{p}_{N} = \frac{\sum_{i=1}^{n} I_{N}(FDx)}{n} \tag{4-11}$$

$$I_{N}(FDx) = [FDx > 1] \tag{4-12}$$

$$SE[\hat{p}_{N}] = \sqrt{\frac{\hat{p}_{N}(1-\hat{p}_{N})}{n}} \tag{4-13}$$

$$95\%\ CI[\hat{p}_{BN}] = \hat{p}_{BN} \pm 1.96 \times SE[\hat{p}_{BN}] \tag{4-14}$$

Reference prevalence rates were obtained from the literature. If possible, US based prevalence rates of psychiatric disorders among children and adolescents were used. Reference prevalence rates for bipolar disorder (BPD), panic disorder (PD), separation anxiety disorder (SeAD), social anxiety disorder (SoAD), specific phobia (SP), generalized anxiety disorder (GAD), eating or feeding disorder (EFD), oppositional defiant disorder (ODD), conduct disorder (CD), post-traumatic stress disorder (PTSD), and alcohol use disorder (AUD) are based on the Merikangas et al. (2010) study. The reference prevalence rate for OCD was based on the Zohar (1999) study. The TD

reference prevalence was based on the Scharf et al. (2012) study. The BED reference prevalence was based on Kjeldbjerg & Clausen (2021). The SPD reference prevalence was based on Lewien et al. (2021).

Note that diagnoses were made using KSADS-5 which was developed on the DSM-5 diagnostic manual, whereas the reference prevalence rates are based on the DSM-IV diagnostic manual. Nonetheless, the changes for listed disorders between the two editions of DSM are not substantial enough to warrant concerns over prevalence rate comparisons (APA, 2013; APA, 2000).

Comorbidities were calculated by first subsetting individuals with a primary disorder of interest, then calculating the prevalence rate of the secondary disorder in the subset sample. Equation 4-15 demonstrates an example mathematical formula for calculation of comorbid narrow TD (secondary) among individuals with OCD (primary), $C_{TD|OCD}$.

$$C_{TD|OCD} = P\left(FDx_{TD} = 2 \bigcap FDx_{OCD} = 2\right) \tag{4-15}$$

Deciding which disorder is set as primary plays an important role when it comes to epidemiological assessments of psychiatric disorders because comorbidities are not symmetric. For example, tic disorders occur in about 15.3% of adults and 11.9% of children with OCD (Sharma et al., 2021). Conversely OCD occurs in about 66.1% of children with TS or 38.0% children with a TD (Hirschtritt et al., 2015; Huisman-van Dijk et al., 2019).

CBCL Variables

To calculate the scores for the obsessions and compulsions, I take the longitudinal sum of the obsessions and compulsions questions from the CBCL over the

first 3 timepoints, that is, the sum of $CBCL_9$ and $CBCL_{66}$, respectively. Any item CBCL score, CBCL_ITEM, can be calculated in a similar way. The sum of the obsessions and compulsions items was used to derive OCS score. The mathematical operation is shown in Equations 4-16 and 4-17.

$$CBCL_{ITEM} = \sum_{i=1}^{3} CBCL_{ITEM,\ i} \tag{4-16}$$

$$OCS = CBCL_9 + CBCL_{66} \tag{4-17}$$

Similarly, a 6-item OCP score as described by Storch et al. (2006) can be found by summing the following CBCL items: $CBCL_9$ (obsessions), $CBCL_{31}$ (fears of thinking/doing something bad), $CBCL_{52}$ (feelings of guilt), $CBCL_{66}$ (compulsions), $CBCL_{85}$ (strange ideas), and $CBCL_{112}$ (worries). Thus, the OCP score can be mathematically represented according to the Equation 4-18.

$$OCP = CBCL_9 + CBCL_{31} + CBCL_{52} + CBCL_{66} + CBCL_{85} + CBCL_{112} \tag{4-18}$$

Analysis of CBCL and Diagnoses

Odds ratio with 95% confidence intervals were extracted from logistic regressions of OCS, OCP, compulsions, and obsessions scores to narrow and broad diagnosis values covarying for sex (Tsuang et al., 2011; R Core Team, 2013). Histogram plots and OR plots are included to visually represent distributional patterns of these quantitative constructs.

I analyze these constructs with respect to OCD and TD. In case of OCD, my goal is to see if parent reports of symptom severity associate with their endorsement of OCD. I also want to identify which symptoms are primary drivers of OCD endorsement. In case of TD, I want to see if parent reports of symptom severity associate with TD as TD usually presents with some obsessive-compulsive symptoms and is highly comorbid.

Computational Resources

All computations were performed in RStudio version 1.4.1106, using R version 4.0.5 (2021-03-31), on x86_64-apple-darwin17.0 (64-bit) platform running under macOS Big Sur 10.16. Logistic regressions were fit using generalized linear models with binomial link in stats package (R Core Team, 2013). Data wrangling, transformations, and visualizations were done using the tidyverse set of packages (Wickham et al., 2019).

Results

There is an Over-endorsement of Psychiatric Disorders in the ABCD Study

Analysis of the prevalence of psychiatric disorders as depicted in Figure 4-3a and Table 4-2 shows large rates of over-endorsement in the ABCD Study. All disorders except for CD, EFD, PD, PTSD, and SoAD were substantially more prevalent in the sample than expected based on reported population prevalences, with the largest over-endorsements being in TD (8.24 times the reference rate) and OCD (5.80 times the reference rate). The prevalence of broad OCD was slightly higher among males vs. females (13.25% vs. 11.50%).

Narrow Definitions Reflect Reference Prevalence Rates Better

After modifying definitions for diagnosis and looking at the narrow definitions specifically, the prevalences of psychiatric disorders decreased across the board and more closely resembled the reported reference prevalence rates (Figure 4-3b, Table 4-2). Nonetheless, some disorders were still overrepresented, albeit to a less extreme extent, such as TD (1.89 times the reference prevalence rate) and OCD (1.13 times the reference prevalence rate). Conversely, using this definition, some disorders were underrepresented relative to population prevalences, like PD (0.03 times the reference

prevalence rate) and EFD (0.07 times the reference prevalence rate). Notably, as a group, EFD disorders excludes AN which was pulled from the dataset due to programming concerns, thus the EFD category is lacking one of the more relevant contributing disorders.

The prevalence of narrow OCD using the modified Sordo Viera method was slightly higher among males vs. females (3.69% vs. 2.28%). This finding is in line with literature reports of slightly higher OCD rates among males in this age group.

Comorbidities Show Variability Between nOCD and bnOCD

Comorbidity analysis of OCD indicated that comorbidity profiles differed between nOCD and bnOCD (Figure 4-4). Looking specifically at TD and OCD, 46.81% of individuals in the narrow category who had TD also had OCD, whereas 5.55% of individuals in the narrow category who had OCD also had TD. Conversely, 35.04% of individuals in the broad or narrow category with TD also had OCD, yet 18.16% of individuals in the broad or narrow category with OCD also had TD. These comorbidity rates approximate those reported in the literature: 11.9% of children with OCD having comorbid TD (Sharma et al., 2021) and 38.0% of children with a TD having comorbid OCD (Huisman-van Dijk et al., 2019). However, due to a low prevalence of OCD in the sample and only a fraction of the sample being assessed for TD, this number could be an underestimate.

OCS is a Better Predictor of nOCD than OCP

Based on previous literature, I expect individuals with OCD to have higher CBCL scores for obsessions and compulsions, as well as OCS and OCP derivative scales. Distribution analysis indicated that both the OCS and OCP constructs had OCD-dependent stratification, especially with respect to broad vs. narrow grouping of OCD

(Figure 4-5). The bOCD group shows a slight right shift in both OCS and OCP scores, but overall, still a negative binomial-like distribution. Conversely, individuals with nOCD show a substantial right shift and bell-like distribution for both OCS and OCP, but more dramatically so for OCS. Logistic regression of OCS to OCD diagnosis status (either bnOCD or nOCD) shows higher odds of OCD given nOCD vs. bnOCD ($OR_{bnOCD} = 1.71$, $p_{bnOCD} \ll 0.001$, $OR_{nOCD} = 1.81$, $p_{nOCD} \ll 0.001$). Furthermore, OCS had a higher predicted OR compared to OCP ($OR_{bnOCD} = 1.29$, $p_{bnOCD} \ll 0.001$, $OR_{nOCD} = 1.32$, $p_{nOCD} \ll 0.001$). Visual representation of these associations is shown in Figure 4-6 and Table 4-3. All associations were statistically significant at $p \ll 0.001$.

Association Between CBCL and OCD is Primarily Driven by Compulsions

Further analysis of the CBCL OCD constructs indicates that compulsions are the main driver of association between OCS and OCD (Figure 4-6, Figure 4-7, Table 4-3). However, compulsions did not differentiate as well between nOCD and bnOCD as obsessions did (Figure 4-6, Figure 4-7, Table 4-3). Association between compulsions and nOCD vs. bnOCD was high ($OR_{bnOCD} = 3.12$, $p_{bnOCD} \ll 0.001$, $OR_{nOCD} = 3.16$, $p_{nOCD} \ll 0.001$), but not very different and had overlapping 95% confidence intervals. Conversely, the association between obsessions and nOCD vs. bnOCD was smaller ($OR_{bnOCD} = 1.85$, $p_{bnOCD} \ll 0.001$, $OR_{nOCD} = 2.14$, $p_{nOCD} \ll 0.001$), but the estimates were different and had non-overlapping 95% confidence interval.

TD in the ABCD Study Follow Expected Prevalence Patterns

The broad and narrow TD prevalence was estimated at 6.59%, whereas the narrow prevalence of TD was estimated at 1.51%. Epidemiological studies put TD prevalence rate for TD at about 0.80% (Scharf et al., 2012). Sex ratios were also more

pronounced in TD, with bnTD being overrepresented in males at 2:1 ratio (7.20% vs. 3.28% for males and females, respectively) and nTD being overrepresented in males at about 3:1 ratio (2.37% vs. 0.82% for males and females, respectively). This sex ratio also mimics that reported in literature (2:1 to 4:1, APA, 2013).

CBCL Constructs Show TD-Dependent Stratification

Visual inspection of CBCL constructs with respect to TD has also shown stratification. This is expected for couple of reasons: high rates of comorbid OCD among individuals with TD and potential confusion of tics for compulsions in parents. Both OCS and OCP vary with respect to broad vs. narrow TD (Figure 4-8). This association seems to be primarily driven by compulsions score (Figure 4-9). Logistic regression has shown a substantial magnitude and strength between all CBCL constructs and TD, regardless of bnTD vs. nTD (Table 4-4, Figure 4-10).

Discussion

In this chapter, I outline the phenotypic exploration of OCD and related phenotypes, including comorbidity analysis and exploration of self-report symptom data available in the ABCD Study. The results indicate that when the data are taken at face value, the ABCD cohort shows a pervasive overdiagnosis of psychiatric disorders, which was expected given the study design. As previously mentioned, Townsend et al. (2019) show low clinician-parent positive agreement at 26%, with parents being prone to over-endorsing OCD. This is reflected in the prevalence of OCD being vastly more prevalent in ABCD Study than the epidemiologically established rates ($p_{bnOCD} = 0.1335$, $p_{refOCD} = 0.0230$), or about 5.8 times more prevalent (Table 4-2, Figure 4-3). The same was true of TD in the study which about 8.2 time more prevalent ($p_{bnTD} = 0.0659$, $p_{refTD} =$

0.0080); note that Townsend et al. (2019) did not assess tic disorders in their study. Thus, a closer inspection of psychiatric phenotypes in the ABCD Study was necessary.

Based on the similar principle reported by Sordo Vieira et al. (2022), longitudinal consistency was deemed a crucial component to improving confidence of diagnosis. I defined narrow OCD diagnosis by only classifying those individuals who endorsed OCD on most of the available assessment timepoints and used a similar approach for other psychiatric diagnoses as a test of this method in this sample. As a result, the prevalence of otherwise over-endorsed disorders was reduced, and for the two primary disorders of interest, the reduction resulted in prevalence rates that were much closer to the previously reported rates ($p_{nOCD} = 0.0260$, $p_{nTD} = 0.0151$). This was true across the board, except for a few disorders which had lower-than-expected prevalences, which can be attributable to age of onset for these disorders being later in adolescence.

Comorbidity analysis has shown that bnOCD and nOCD have substantially different comorbidity profiles (Figure 4-4). Specifically, the rate of bnOCD among individuals with bnTD was 35.04%, and the rate of nOCD among individuals with nTD was 46.81%, indicating higher risk of OCD among individuals with TD than the reverse, which is in line with the previous literature, such as the 62.78% rate of comorbid OCD among individuals with TD reported by Claudio-Campos et al. (2021). Curiously, constraining diagnosis to narrow when OCD is primary reduced the rates of comorbid TD, with rates of bnTD among bnOCD cases being 18.16% and rates of nTD among nOCD cases being 5.55%. One reason this might be the case is likely that only 52.63% of participants have reached the stage of third annual follow-up at which TD was assessed (Figure 4-1). Another explanation is that broad and narrow diagnoses have

shown globally higher rates of comorbidities. Ultimately, nTD might be too restrictive diagnosis construct as it requires the presence of both motor and phonic tics, thus acting as a proxy for TS more so than TD.

Further exploration of psychiatric symptoms reported in the CBCL questionnaire found that more severe symptoms related to OCD stratified by the OCD diagnosis construct. Specifically, distributions of OCS and OCP were substantially more right shifted among nOCD cases than bOCD cases (Figure 4-5). Further analysis has shown this effect to be primarily driven by compulsions, with obsessions being a differentiating factor (Figure 4-6, Figure 4-7, Table 4-3), i.e., compulsions drive any diagnosis of OCD whereas obsessions drive narrow diagnosis of OCD specifically. One reason why compulsions might be driving the differentiation between bOCD or nOCD from OCD negative individuals could be that compulsive symptoms are usually behaviors and are thus more easily noticed and remembered. Conversely, one reason why obsessions might be contributing to differentiation between bOCD and nOCD individuals could be the fact that nOCD individuals might experience more severe obsessive symptoms that would be more easily noticed by the parents.

Analysis of TD in the ABCD study also indicated that comorbidities and other characteristics follow previously reported clinical TD patterns. In addition to TD-OCD comorbidity patterns replicating those reported in the literature, sex ratio analysis has shown TD to follow similar patterns to those reported in the literature, with males being predominantly affected, especially in nTD, with male-to-female ratio of 3:1. Additional analysis of TD in this sample has shown, similar to OCD, TD-dependent stratification of psychiatric symptoms (Figure 4-8, Figure 4-9). Logistic regression of CBCL constructs

to TD diagnosis status has confirmed these observations (Table 4-4, Figure 4-10). In other words, broad TD cases have diluted overall relationship between TD status and the four examined CBCL constructs.

In conclusion, the high rate of OCD in the ABCD Study is likely due to phenotype misclassification. Revising the phenotype definition to a more conservative, longitudinal constructs reduces the observed diagnosis rates of OCD to much closer to those reported in the literature. This is likely due to the removal of some individuals who would be considered to have obsessive compulsive symptoms, but not meet diagnostic criteria if assessed by clinicians. Thus, bnOCD construct could be considered a proxy for either obsessive-compulsive symptoms across a variety of diagnoses or subclinical symptom states or of general neurodevelopmental psychopathology, whereas nOCD can be considered a proxy for clinical OCD diagnosis. These will be further explored in genetic studies in Chapter 5.

Table 4-1. KSADS-5 module administration schedule.

	KSADS-5 Modules	P0	P1	P2	P3	Y0	Y1	Y2	Y3
M1.	Depressive disorders [a]	X		X	X	X		X	
M2.	Bipolar disorders	X		X		X		X	
M3.	Disruptive mood regulation disorders [a]	X				X			
M4.	Psychosis	X	X	X	X				
M5.	Panic disorders	X		X					
M6.	Agoraphobia [a]	X		X					
M7.	Separation anxiety disorder	X		X					
M8.	Social anxiety disorder	X		X		X		X	
M9.	Specific phobia	X		X					
M10.	Generalized anxiety disorder	X		X		X		X	
M11.	Obsessive-compulsive disorder	X		X					
M12.	Enuresis and encopresis								
M13.	Eating disorders [a]	X	X	X	X	X		X	
M14.	Attention deficit / hyperactivity disorder [a]	X		X	X				
M15.	Oppositional defiant disorder	X	X	X					
M16.	Conduct disorder	X	X	X	X	X		X	
M17.	Tic disorders				X				
M18.	Autism spectrum disorder [a]	X		X					
M19.	Alcohol use disorder	X		X	X				X
M20.	Drug use disorder	X		X	X				X
M21.	Post-traumatic stress disorder	X		X					
M22.	Sleep problems	X		X		X		X	
M23.	Suicidaliity	X		X		X	X	X	X
M24.	Homicidality	X		X		X	X	X	
M25.	Selective mutism	X		X					

P: parents, Y: youth, 0: baseline, 1-3: follow-ups 1-3. [a] Data have been partially or completely removed from the dataset by ABCD study administrators due to a programming error in KSADS-5 data processing.

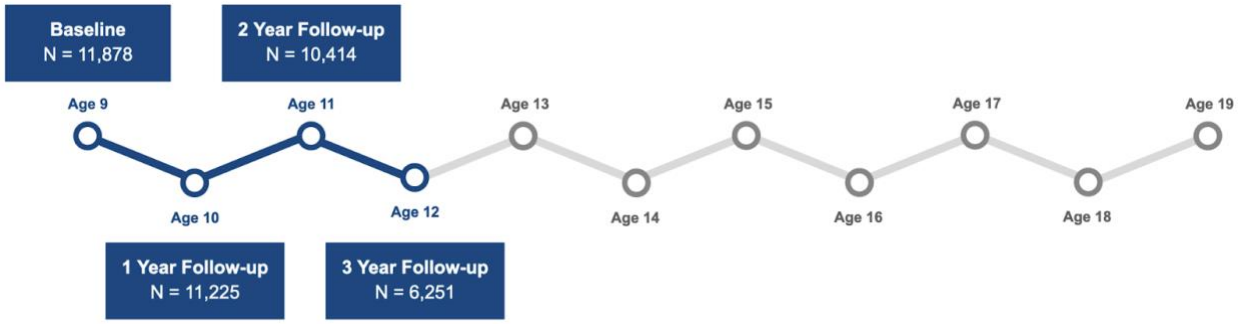

Figure 4-1. ABCD study timeline with number of KSADS-5 data points. So far, data are available for the first 3 assessment points for most participants and for the 4th assessment point for about half of participants, with 7-8 time points remaining in the study.

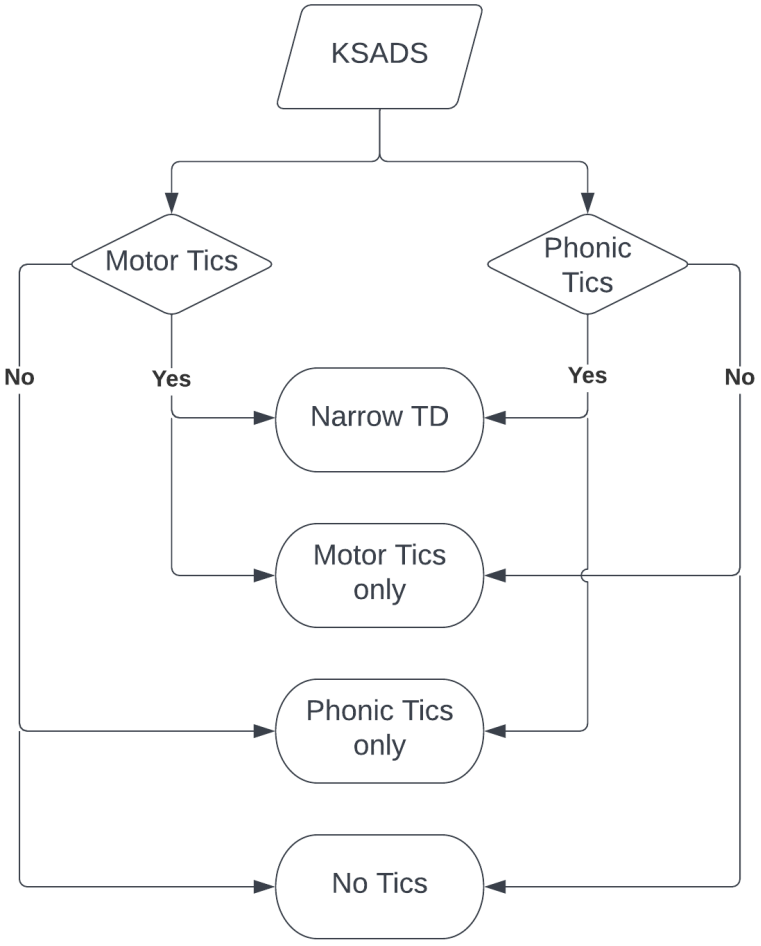

Figure 4-2. Decision diagram for TD diagnosis. Only individuals who have a history of both motor and vocal tics are categorized as having TD.

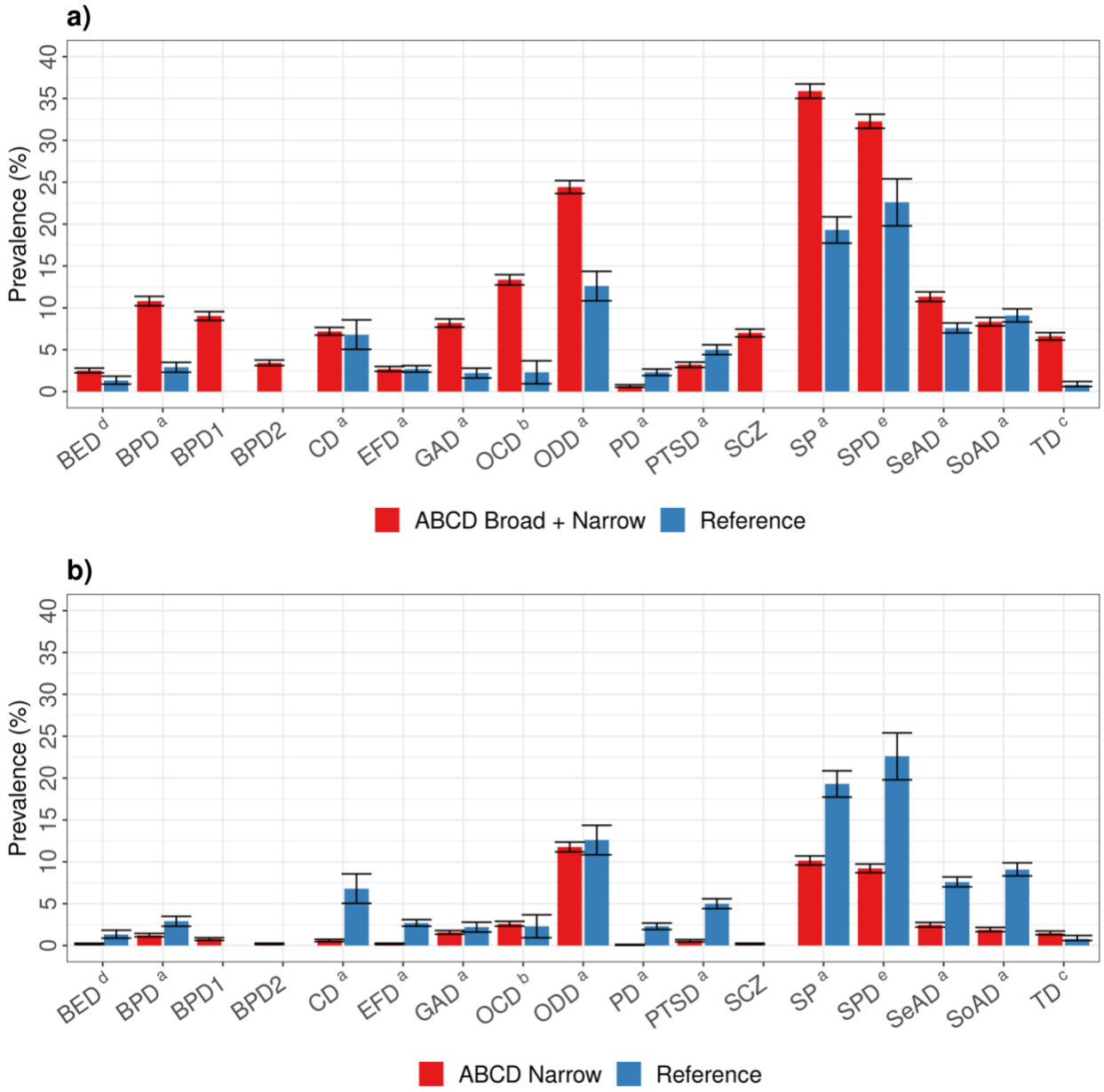

Figure 4-3. Prevalence rates of psychiatric disorders in the ABCD Study. a) if broad and narrow and b) if narrow only. Reference prevalence rates: [a] Merikangas et al. (2010), [b] Zohar (1999), [c] Scharf et al. (2012), [d] Kjeldbjerg & Clausen (2021), and [e] Lewien et al. (2021).

Table 4-2. Tabulated prevalence rates of psychiatric disorders in the ABCD Study.

Disorder	Broad + Narrow	Narrow	Reference
BED	2.52	0.19	[a] 1.32
BPD	10.81	1.25	[a] 2.90
BPD1	9.02	0.77	-
BPD2	3.42	0.20	-
CD	7.20	0.60	[a] 6.80
EFD	2.70	0.20	[a] 2.70
GAD	8.18	1.57	[a] 2.20
OCD	13.35	2.60	[b] 2.30
ODD	24.43	11.77	[a] 12.60
PD	0.66	0.08	[a] 2.30
PTSD	3.20	0.57	[a] 5.00
SCZ	6.99	0.20	-
SP	35.88	10.16	[a] 19.30
SPD	32.28	9.21	[e] 22.60
SeAD	11.33	2.48	[a] 7.60
SoAD	8.35	1.91	[a] 9.10
TD	6.59	1.51	[c] 0.80

[a] Merikangas et al. (2010)
[b] Zohar (1999)
[c] Scharf et al. (2012)
[d] Kjeldbjerg & Clausen (2021)
[e] Lewien et al. (2021)

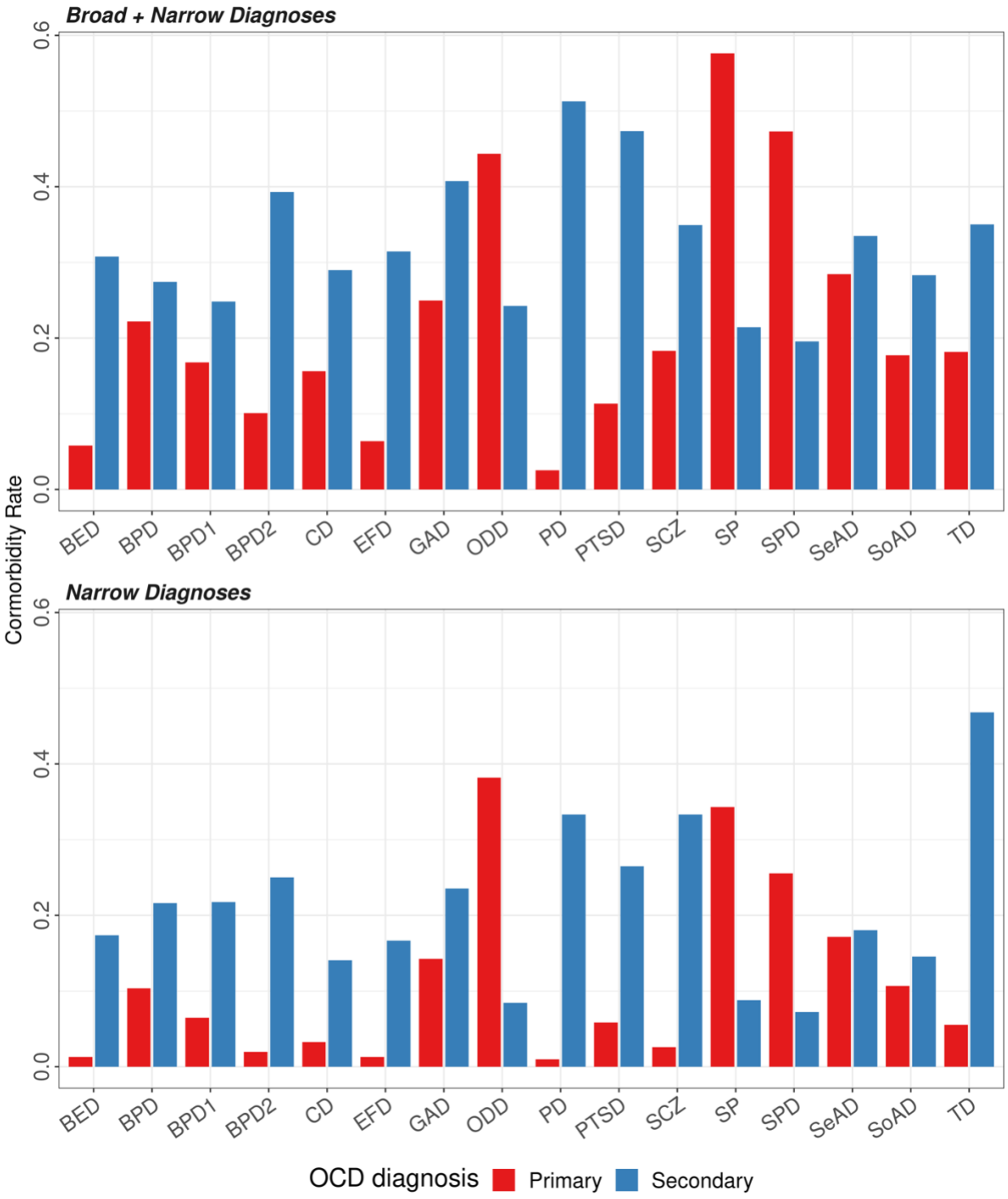

Figure 4-4. Comorbidity analysis of OCD in ABCD Study. Broad and narrow definitions (top) and narrow definitions only (bottom). Red bars represent comorbidities when OCD is a primary diagnosis, blue bars represent comorbidities when OCD is a secondary diagnosis.

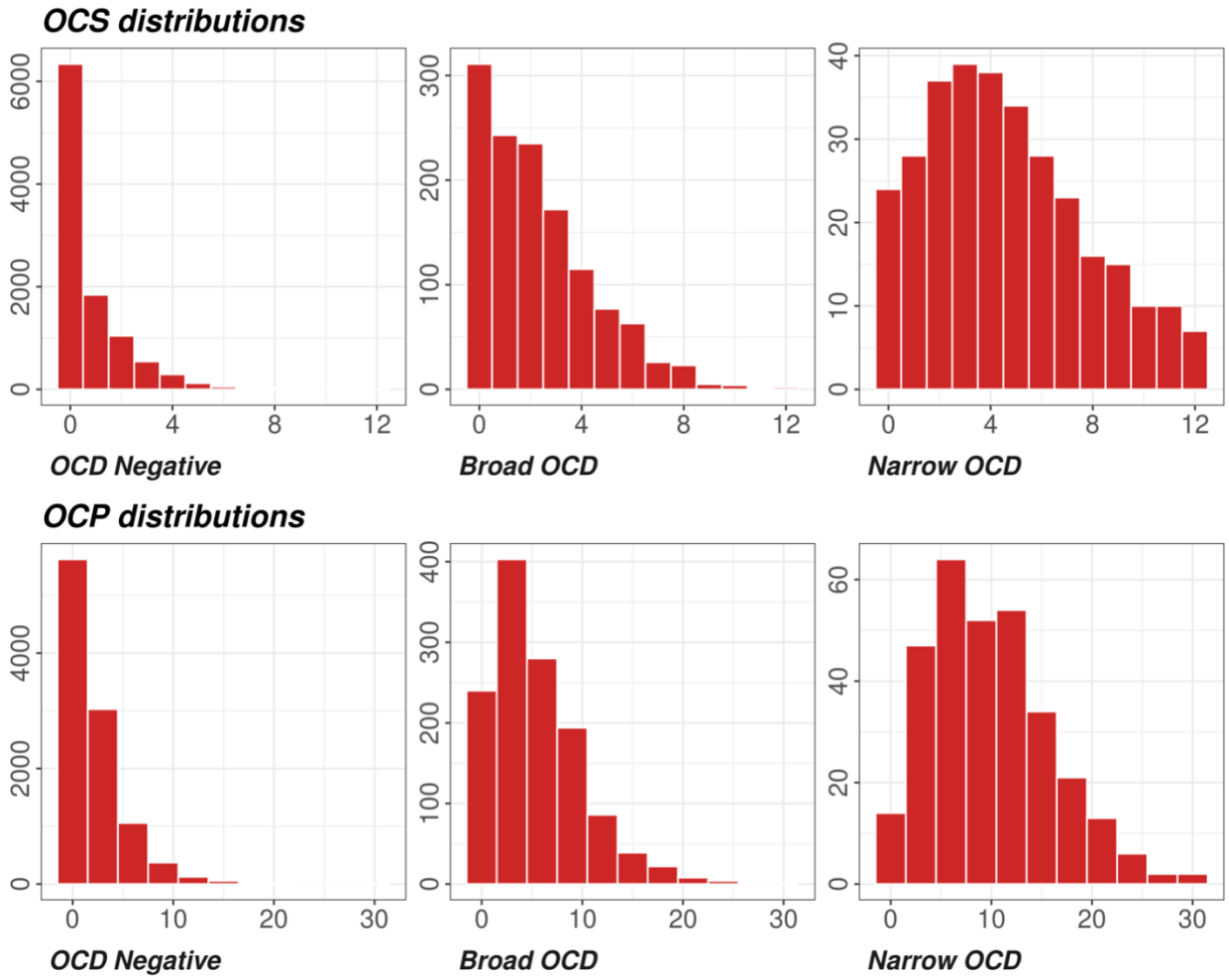

Figure 4-5. Histograms of OCS and OCP values stratified by OCD diagnosis. OCS (top) and OCP (bottom) both show OCD diagnosis related skew.

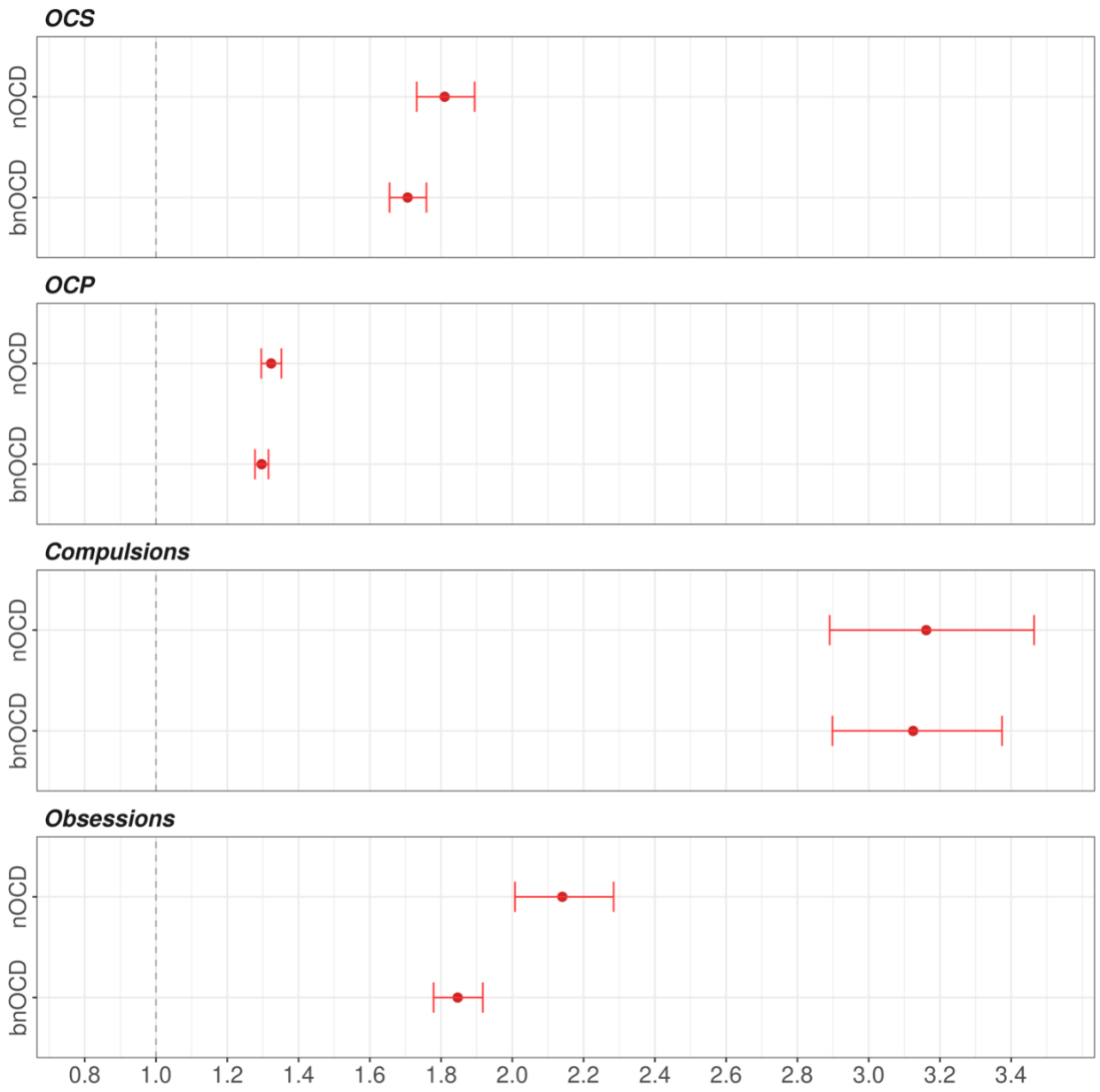

Figure 4-6. Graphical representation of logistic regressions of CBCL constructs on OCD diagnoses. Dashed line indicates no risk (OR = 1). All associations were statistically significant at $p \ll 0.001$.

Table 4-3. Summary statistics of logistic regressions of CBCL constructs on OCD diagnoses.

OCD Diagnosis	CBCL Construct	OR	Lower 95% CI	Upper 95% CI	P
Broad + narrow	OCS	1.71	1.66	1.76	$\ll 0.001$
Narrow	OCS	1.81	1.73	1.89	$\ll 0.001$
Broad + narrow	OCP	1.30	1.28	1.32	$\ll 0.001$
Narrow	OCP	1.32	1.30	1.35	$\ll 0.001$
Broad + narrow	Compulsions	3.12	2.90	3.37	$\ll 0.001$
Narrow	Compulsions	3.16	2.90	3.46	$\ll 0.001$
Broad + narrow	Obsessions	1.85	1.78	1.92	$\ll 0.001$
Narrow	Obsessions	2.14	2.01	2.28	$\ll 0.001$

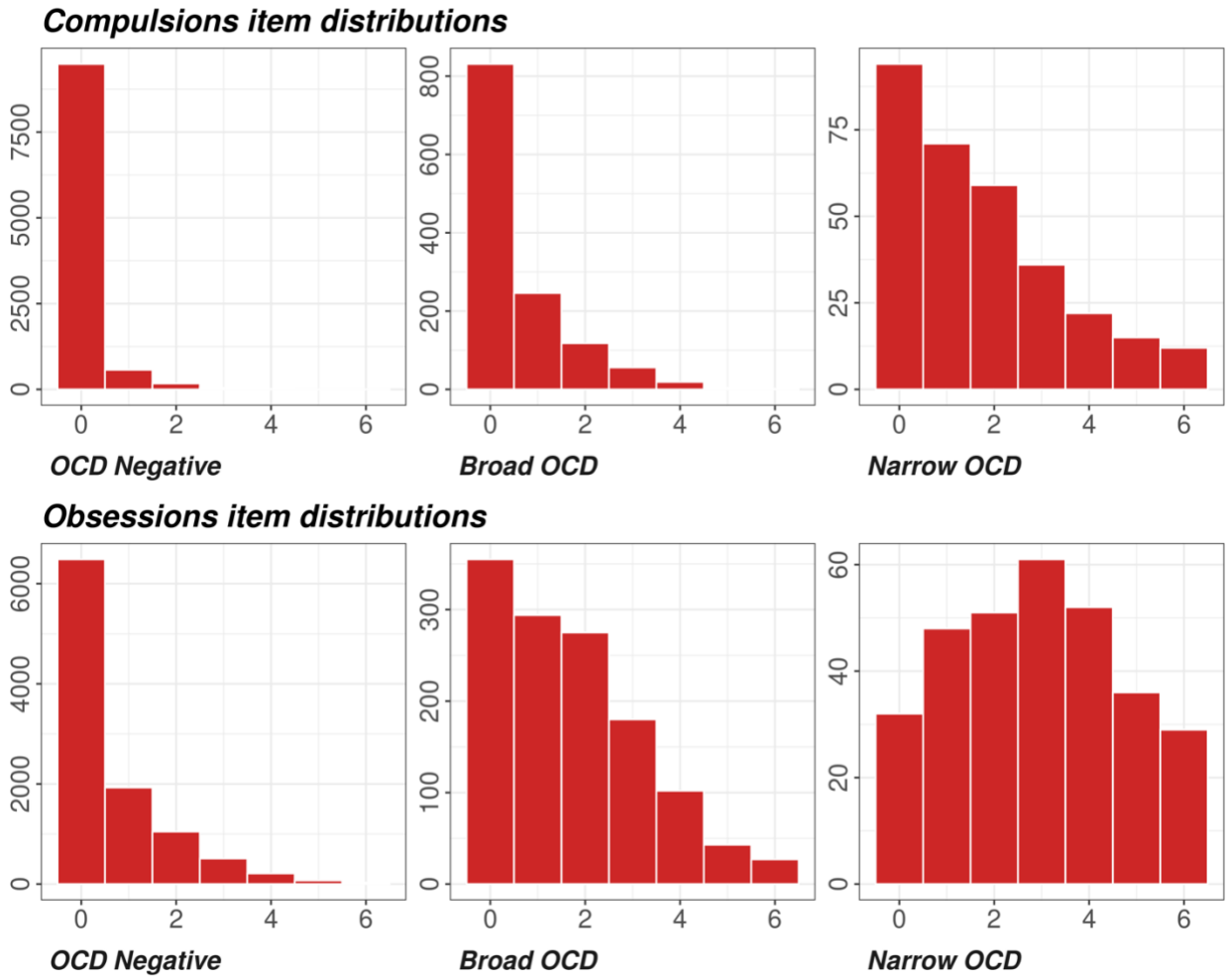

Figure 4-7. Histograms of compulsions and obsessions values stratified by OCD diagnosis. Compulsions (top) and obsessions (bottom) both show OCD diagnosis related skew.

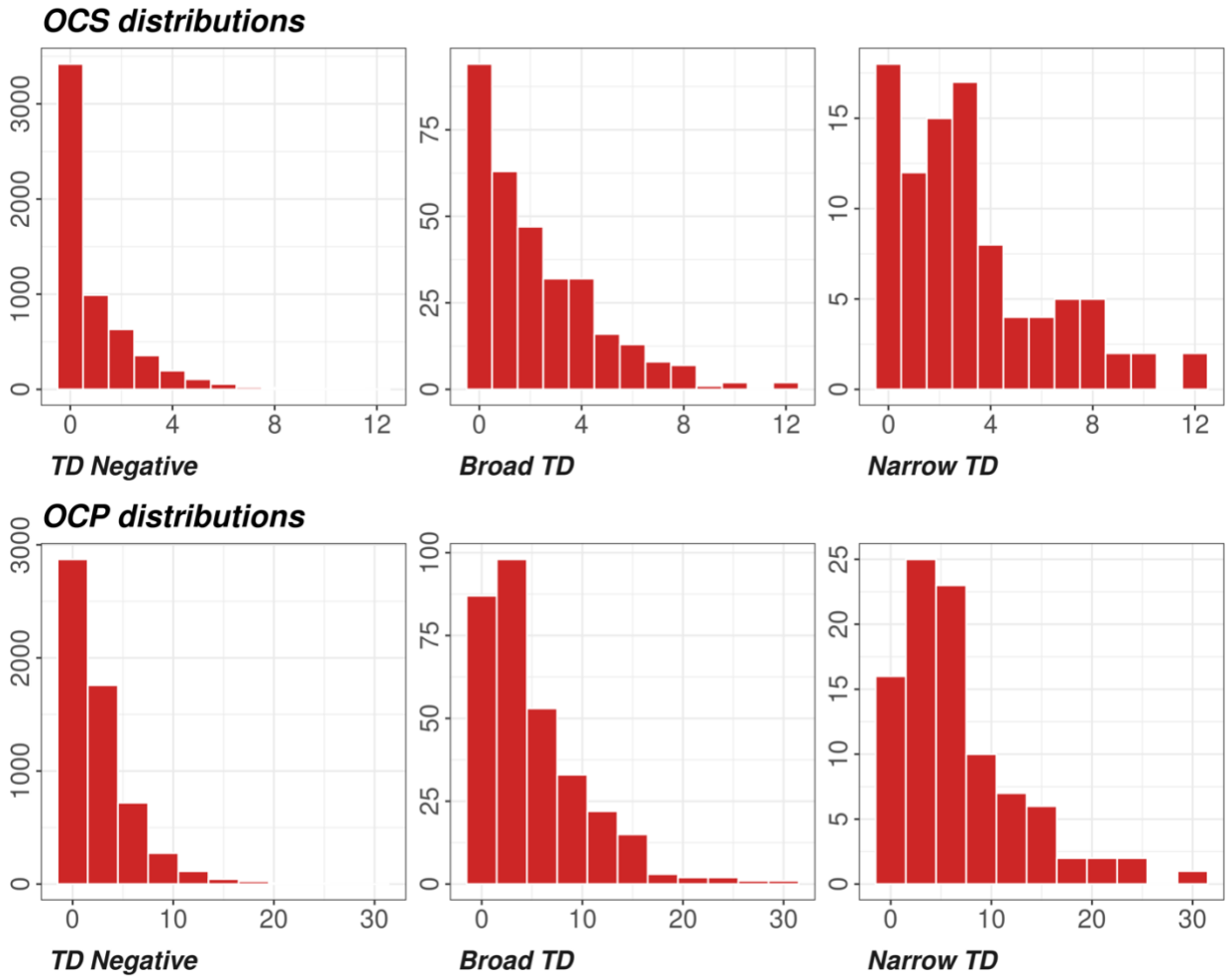

Figure 4-8. Histograms of OCS and OCP values stratified by TD diagnosis. OCS (top) and OCP (bottom) both show TD diagnosis related skew.

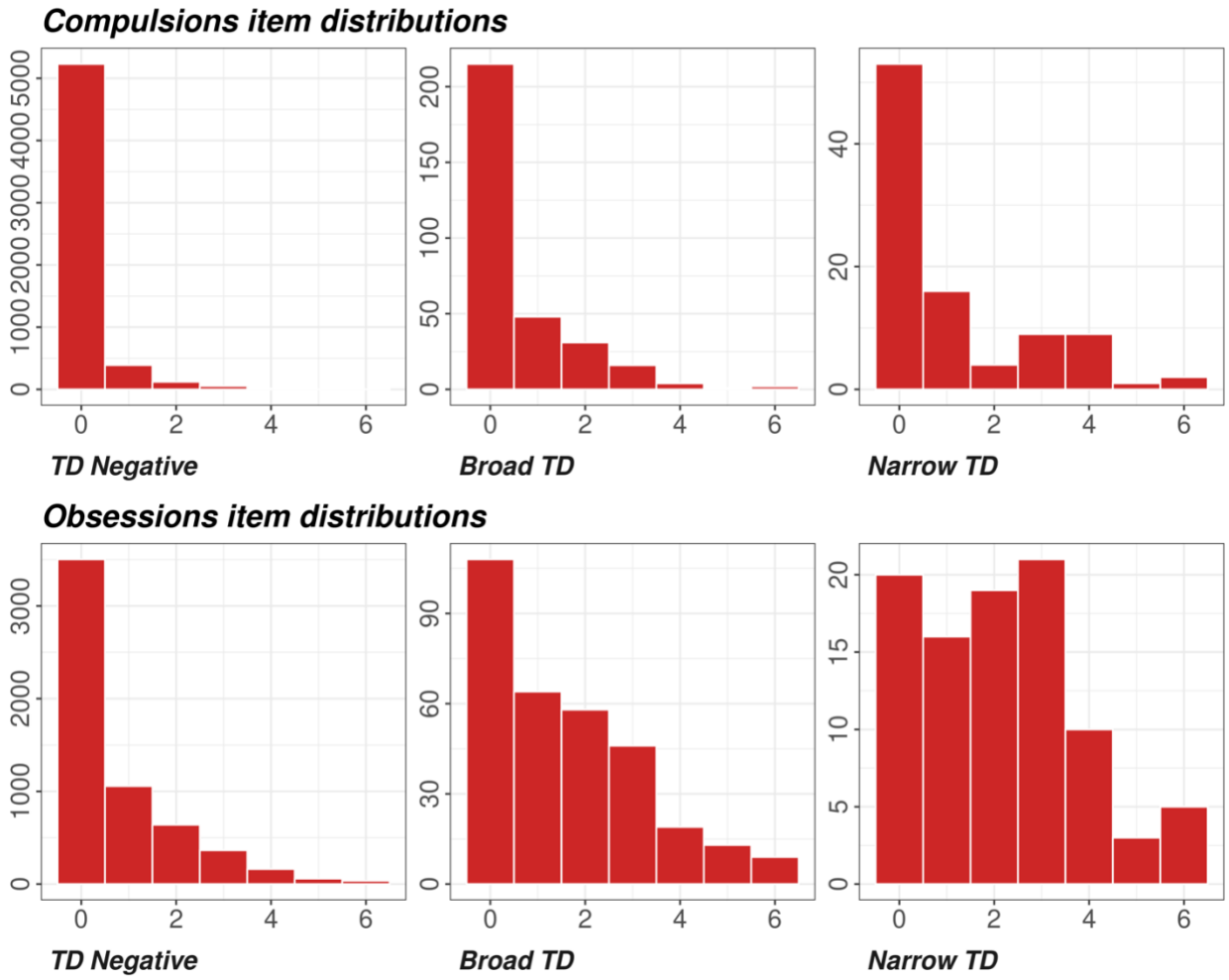

Figure 4-9. Histograms of compulsions and obsessions values stratified by TD diagnosis. Compulsions (top) and obsessions (bottom) both show TD diagnosis related skew.

Table 4-4. Summary statistics of logistic regressions of CBCL constructs on TD diagnoses.

TD Diagnosis	CBCL Construct	OR	Lower 95% CI	Upper 95% CI	P
Broad + narrow	OCS	1.38	1.33	1.45	$\ll 0.001$
Narrow	OCS	1.44	1.35	1.55	$\ll 0.001$
Broad + narrow	OCP	1.16	1.13	1.18	$\ll 0.001$
Narrow	OCP	1.17	1.13	1.21	$\ll 0.001$
Broad + narrow	Compulsions	1.96	1.77	2.16	$\ll 0.001$
Narrow	Compulsions	2.07	1.80	2.37	$\ll 0.001$
Broad + narrow	Obsessions	1.49	1.40	1.58	$\ll 0.001$
Narrow	Obsessions	1.61	1.44	1.80	$\ll 0.001$

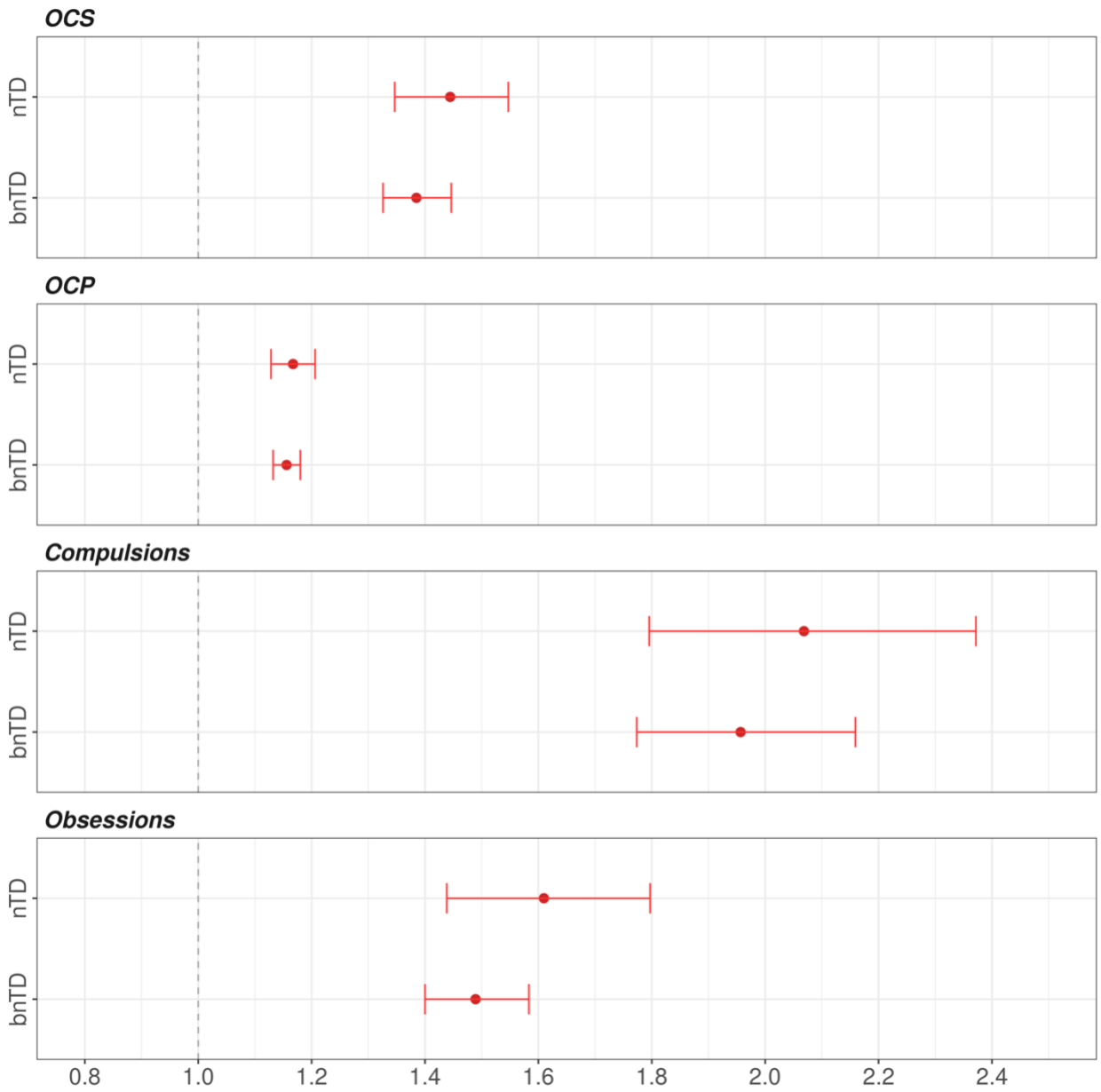

Figure 4-10. Graphical representation of logistic regressions of CBCL constructs on TD diagnoses. Dashed line indicates no risk (OR = 1). All associations were statistically significant at $p \ll 0.001$

CHAPTER 5
GENETIC ARCHITECTURE OF OCRD PHENOTYPES IN ABCD STUDY

Background

As discussed in Chapters 2 and 3, there is substantial evidence of genetic effects underlying OCRD phenotypes, yet exact causative variants remain to be discovered. The main reasons for this include limited sample sizes, as well as complex and small genetic effect sizes. One way to overcome this limitation is by meta-analyzing numerous small studies. For that reason, it is beneficial to conduct GWAS of OCD in the ABCD sample despite a limited number of cases, as resulting summary statistics may contribute to a large collection of OCD GWAS studies to be meta-analyzed, such as is done in the Psychiatric Genomics Consortium (PGC). Another reason to conduct such studies in this sample with such expansive phenotyping is that it allows for in-depth analysis of the genetic architecture of OCD in this unique population sample of American children and adolescents.

The ABCD study is a large scale, longitudinal brain development and child health study (n = 11,924), which, in addition to the phenotype data discussed in Chapter 4, also includes genotype data. In this chapter, I explore genetic variability in OCRD phenotypes in the ABCD sample, as well as cross-disorder genetic associations. This unique, population sample of children/adolescents has the potential to provide a valuable insight into pediatric OCD and other childhood onset psychiatric disorders.

Methods

Genotype Data

Genotyping data were derived from Affymetrix NIDA (National Institute for Drugs and Addiction) SmokeScreen Array (Baurley et al, 2016). The smokescreen array

consists of 733,293 markers optimized for addiction research. DNA was obtained from whole blood and saliva. Genotyping data are provided as pre-processed PLINK files (.bim, .bed, .fam) for 11,099 individuals with 516,598 markers (Chang et al., 2015). All markers have already been referenced and oriented to the positive strand.

Quality Control of Genotype Data

Initial exploration of the ABCD cohort revealed 38 individuals with problematic sex (38 with sex missing in the phenotype file, with 1 of them also being genetically too ambiguous to call). Further diagnostics revealed that no psychiatric data were available for these 38 samples, thus they were removed from the further analysis. After this removal, no samples had issues with or discrepancies between reported phenotypic and genetic sex.

Genotyping rate is defined either as the percent of successfully genotyped SNPs across an entire sample (Sample genotyping rate, GR_S), or the percent of samples with successfully genotyped SNP (SNP genotyping rate, GR_{SNP}). As a rule of thumb, GR_S and GR_{SNP} should both be > 0.98. However, the ABCD data had rates less than 0.98 which is why additional cleaning was required (Figure 5-1). Closer examination of GR_S and GR_{SNP} in a batch-wise fashion has shown that 2 blood batches underperformed compared to 4 saliva batches (Figure 5-2), thus samples from these 2 whole blood batches were removed from further analysis (totaling 207 samples, 2 of which were nOCD).

Subsequently, a serial filtering of genetic data based on genotyping rates was conducted in the following order to remove poorly genotyped SNPs and samples using a loop filtering iterating from i = 0.95 to i = 0.98 by 0.01, filtering SNPs with $GR_{SNP} \geq i$ first, then followed by filtering samples with $GR_S \geq i$. After genotyping rate filtering, non-

autosomal SNPs were removed from the dataset. The resulting dataset had 480,427 markers and 10,763 individuals.

Once poorly genotyped markers were removed, the dataset was LD-pruned to allow unbiased relationship estimations. Markers were pruned in PLINK using independent pairwise correlations in 200kb windows, sliding by 50 variants, and removing any markers with correlation of $r^2 \geq 0.15$. This reduced the dataset of LD-pruned markers to 152,163 markers. Subsequently, identity by state (IBS) and identity by descent (IBD) calculations were used to identify and remove related individuals. IBS/IBD was significant if the π metric, a metric representing the proportion IBD, was equivalent to the sum of proportions of half of one-allele IBD and two-allele IBD (Equation 5-1).

$$\hat{\pi} = \hat{p}[IBD = 2] \times \frac{1}{2}\, \hat{p}[IBD = 1] \tag{5-1}$$

Samples were removed in the following preferential order: for samples with genetic relatedness, if the OCD phenotypes mismatched (i.e., one individual had an OCD phenotype and the other didn't), then individuals without OCD were removed, if the OCD phenotypes matched, then individuals with lower genotyping rate were removed. After removal of related samples, 8,718 individuals with both OCD phenotype data and genotype data remained.

In the non-pruned data set, Hardy-Weinberg equilibrium (HWE) exact test statistics were calculated for the remaining markers and individuals (Wigginton et al., 2005). Because this sample was subjected to several GWASes, both case-control (ccGWAS) and quantitative (qGWAS) design, and using linear mixed models (LMM), samples were not filtered with respect to case status. Instead, a cutoff of $p_{HWE} \leq 10^{-7}$

was used to remove markers in Hardy-Weinberg disequilibrium. Markers that strongly deviate from HWE are likely a result of genotyping errors. A total 413,359 markers remained after filtering. Distribution of p_{HWE} is shown in Figure 5-5. While in ccGWAS studies SNPs that have $p_{HWE} \leq 10^{-6}$ among cases and $p_{HWE} \leq 10^{-10}$ among controls are usually the ones to be filtered out, due to multiple GWASes on multiple phenotypes in this study, I opted in for a cutoff that's between those two, biased more towards cases (thus more conservative approach). Based on Figure 5-5, no apparent inflection point occurs at these ranges, indicating that this decision should not result in influential biasing SNPs being retained in the analysis.

Heterozygosity, calculated as the inbreeding coefficient F_{HET}, is an indicator of extent of inbreeding for an individual. Individuals may be extreme outliers for heterozygosity for reasons such inbreeding, genotyping issues, or sample cross-contamination. Such individuals are traditionally removed from the analysis due to their potential for influencing associations. Individuals with F_{HET} scores that were not within 4 standard deviations of the mean were removed from the analysis. The overall distribution of heterozygosity in the ABCD sample is shown in Figure 5-6. After removing extreme outliers, 8,713 individuals remained in the study.

Minor allele frequency, MAF, indicates the marker-wise population frequency of the less frequent allele. As GWAS analyses primarily focus on the contribution of common variants to polygenic risk, and markers with extremely low MAF are more susceptible to genotyping errors, their power to detect associations is limited – especially in smaller samples. The overall distribution of MAF with considered

thresholds (0.01, 0.02, 0.03, 0.04, 0.05) is shown in Figure 5-7. Based on these results, a cutoff of 0.05 was selected. After filtering, 294,757 markers were retained for analysis.

Ultimately, insertion/deletion variants were also filtered out. The final dataset was composed of 8,713 individuals and 291,622 markers. This dataset was next used for phasing and imputation.

Phasing and Imputation

Phasing, or haplotype estimation, is a process utilizing positional Burrows-Wheeler transform (PBWT) to choose a local subset of haplotypes using the 1000 Genomes Project phase 3 data (1kGPp3) as a reference panel, which are then resolved using Markov chain Monte Carlo (MCMC) algorithm (Marchini, 2019; The 1000 Genomes Project Consortium, 2015). Phasing in this study was done using SHAPEIT4 (Delaneau et. al, 2019). The reference panel was filtered to keep only bi-allelic SNPs, and MAF of at least 0.001 across all superpopulations (admixed American, African, East Asian, European, and South Asian). Parameters for phasing were: expected error rate in the phase sets of 0.0001; MCMC iteration configuration of 10b,1p,1b,1p,1b,1p,1b,1p,10m (b: burn-in iteration used for haplotype sampling, p: pruning iteration used to sample haplotype and trim unlikely paths in the genotype graphs using transition probabilities, m: main iteration used to sample new haplotypes storing average transition probabilities – to be used in final iteration stage to produce final estimates); PBWT depth of 8.

Genotype imputation is a process where phased genotypes are used in conjunction with reference haplotypes to estimate values of missing genotypes of non-genotyped loci (Marchini, 2019). Imputation increases the number of testable SNPs in association studies; thus, genotype imputation increases power and resolution needed

to better approximate phenotype-associated loci and facilitate fine-mapping of causal variants. Imputation also improves the quality of downstream meta-analyses as overlapping SNPs are necessary to jointly analyze multiple GWASes (de Bakker et al., 2008). The same reference data used for phasing were also used for imputation. To facilitate faster and more accurate imputation, the genome was split into 5Mb chunks flanked by a 250kB buffer. Additionally, genetic data were transformed into IMP5 files to speed up imputation. Imputation itself was done using IMPUTE5 (Rubinacci et al, 2020). IMPUTE5 relies on PBWT and forward-backward HMM algorithms to impute the genotypes. Parameters for imputation were: PBWT depth of 8.

Resulting imputed genome chunks were merged, then filtered to keep only the imputed SNPs with INFO > 0.8 and MAF > 0.01. The INFO score is a measure of the relative statistical information about the SNP allele frequency from the imputed data, and as such, it is a measure of imputation quality (Marchini & Howie, 2010). After phasing, imputation, and filtering, the final number of SNPs was 5,322,421.

Global Ancestry

Global ancestry estimation was estimated with FRAPOSA software, using bias-adjusted principal component analysis (PCA), which calculates the asymptotic shrinkage bias from simple projections and then adjusts for bias which is useful to account for any residual relatedness in the samples (Zhang et al., 2020; Dey & Lee, 2019). The 1kGPp3 data were used as reference. Parameters for ancestry PCA were: method set to online adjusted projection; number of principal components of 20. Samples were classified by superpopulation, keeping only individuals classified as admixed Americans (AMR), Africans (AFR), or Europeans (EUR).

Figures 5-8 and 5-9 show the population and superpopulation stratification of ABCD study in controls, bOCD, and nOCD, with respect to the 1kGPp3 reference set. The first two principal components explain 65.71% of variance and show decent, yet incomplete separation of super populations, indicating additional principal components needing to be considered. Figures 5-10 and 5-11 show the biplots and density plots of the first 4 principal components, explaining 78.73% of the variance. These figures show decent separation for African and European participants, yet not so much for admixed American participants. This could be potentially concerning in downstream analyses of admixed Americans. Final cohort sample sizes were: 1,633 AMR, 1,695 AFR, and 5,164 EUR.

Covariate PCA

Principal components are often used as covariates in GWAS associations to account for population structure. While global ancestry information was used to create ancestral cohorts for independent GWASes, covariates needed to be formed without projecting sample data onto the reference panel to preserve sample-specific information such as allele frequencies and LD-structure. Since the Smokescreen array used to genotype ABCD samples contains a large number of custom SNPs selected for smoking and addiction relevant loci which are mostly European-biased, principal components and genomic relationship matrix construction only LD-pruned backbone markers were used. Twenty principal components were derived using PLINK for covariates in GWAS association tests.

Genomic Relationship Matrix

The genomic relationship matrix (GRM) is a structured matrix summarizing kinship between samples that is commonly used in linear mixed model GWASes. First,

GWASTools R package was used to convert available genetic data into GDS format (Gogarten et al., 2012). Subsequently, mutually unrelated individuals were identified using pairwise measures of genetic relatedness, or kinship coefficients (Manichaikul et al., 2010). The unrelated individuals were then used to construct ancestry principal components followed by projection of related individuals onto the derived principal components (Conomos et al., 2015), this projection is graphically represented in the principal component biplots in Figure 5-12. This was followed by construction of ancestry-aware genomic relationship matrices to provide accurate relatedness estimates due only to recent family structure (Conomos et al., 2016). Figure 5-13 demonstrates that this process successfully accounts for any ancestry substructure within the relationship matrix. The resulting GRM is used as a random effects covariate in GWAS linear mixed model association tests.

Phenotypes

Phenotypes were constructed for OCD (nOCD and bnOCD) and OCS (CBCL OCS and CBCL OCP) according to the methods described in Chapter 4. Sample IDs were annotated with the OCD and OCS phenotype information, the first 20 covariate principal components, sex information, and batch information and saved for future use in association analyses.

Case-Control Matching

Case-control imbalanced data has been shown to negatively affect GWAS by causing large numbers of spurious associations when the number of controls is considerably greater than the number of cases (Zhou et al., 2018). To circumvent this issue, for ccGWAS analyses, samples can be clustered by sex and ancestry similarity to get a best-matching subset of controls from the study and reduce the case-control

imbalance. For each ancestry cohort, participants were clustered in PLINK using IBS and sex information and controls chosen for GWAS in a 4:1 case matching ratio (control: case).

Association Testing

GWAS association testing was done using linear mixed modeling in GENESIS (Gogarten et al., 2019). Null models were first fitted using a subset of individuals-based ancestry and case-control (if ccGWAS) clusters. In the null models, I included GRM as random effects covariates, and sex, batch, and the first 20 principal components as fixed effect covariates – thus allowing to effectively control for population structure and relatedness (Chen et al., 2016). In the case of ccGWAS for OCD, the link function used was binomial, whereas for quantitative (q) qGWAS for OCS the link function used was Poisson. The convergence threshold for the average information REML parameter was set to 0.0001. The maximum number of iterations allowed to reach this convergence was set to 100. Variance component terms that converged to 0 were removed from the model. Due to the large number of SNPs tested, association testing was carried out in 5000-SNP blocks.

ccGWAS was performed for both nOCD and bnOCD. For each phenotype, separate ccGWASes were performed for AMR, AFR, EUR, and the combined sample (to be referred to as MEGA). qGWASes were performed for OCS, for AMR, AFR, EUR, and MEGA. Table 5-1 summarizes samples for each GWAS association, total sample sizes, including number of cases and controls for ccGWAS, and percent of the sample that is genetically and assigned at birth female.

I report Manhattan and quantile-quantile (QQ) plots for each GWAS, as well as the genomic inflation factor, λ_{GC}. λ_{GC} is a commonly used method of checking for

systematic biases in GWAS – where $\lambda_{GC} > 1$ indicates presence of a systematic bias that might need to be addressed. λ_{GC} can be mathematically defined according to the Equation 5-2.

$$\lambda_{GC} = \frac{\mathrm{Med}(\chi^2_{GWAS})}{\chi^2_{df=1,\ \alpha=0.05}} \tag{5-2}$$

Gene Annotation and Ontology Analysis

Gene annotation was done using GRCh37 BioMart web server at www.grch37.ensembl.org (Smedley et al., 2009). Briefly, SNPs were filtered to keep only those associated with the trait at $p < 10^{-5}$ for a given GWAS. Subsequently, the list of SNPs was submitted to the aforementioned web server and a list of gene names was obtained for further analysis.

GTEx annotation was done using the GTEx web server at www.gtexportal.org (Lonsdale et al., 2013). All genes of interest were looked up on the portal, tissue expression data was saved after filtering for tissues of interest (central and peripheral nervous system tissues, cardiac, smooth, and skeletal muscle tissues, and pituitary gland) in the form of a violin plot.

GO analysis was conducted using the GO web server at www.geneontology.org (Ashburner et al., 2000; Gene Ontology Consortium et al., 2021; Mi e al., 2018). Briefly, a list of gene names of interest is supplied and analyzed. GO terms with $p<0.05$, after FDR are presented. For the GO analysis discussed in Chapter 2, a list of genes from high-throughput genome-wide studies was collected from previous studies and jointly analyzed. For the GO analysis discussed in Chapter 5 GWAS analyses, a list of genes overlapping markers associated with a given phenotype $p < 10^{-5}$ was collected and analyzed per each individual GWAS.

Polygenic Risk Score Analysis

The polygenic risk score (PRS) is a commonly utilized tool to measure disease risk due to genetic factors and can be defined as the total genetic burden of the disease due to common variants (Meisner & Chatterjee, 2019). From the Equation 1-2, PRS is defined as the sum of products of risk alleles at numerous loci with the weights derived from GWAS summary statistics reports at those loci (Choi et al., 2020). Thus, PRS predictions require 2 independent samples: a discovery sample from which the weights can be determined, and a target sample for testing the calculated PRS association with targeted phenotype. While requirement of two independent samples makes the test demanding, the two phenotypes tested do not need to be the same. Thus, testing PRS models derived from one phenotype can be used to examine another phenotype, effectively probing the extent of shared genetic architecture between the two phenotypes. There are several PRS tests run in this study. Namely, I used weights from the nOCD ccGWAS to test the PRS models in independent bOCD and OCS samples, weights from the OCS qGWAS to test the PRS models in the independent nOCD sample, and finally, weights derived from publicly available PGC GWAS summary statistics to test PRS models in nOCD in the ABCD sample. Table 5-2 summarizes these experiments, indicating the combinations of discovery samples with the target samples.

PRS testing was done using the PRS-PCA approach (Coombes et al., 2020). Briefly, for each discovery sample of interest, summary statistics files are obtained. Subsequently, ORs are transformed into beta estimates by taking their natural logarithms for all summary statistics files obtained from ccGWASes. Subsequently, data are clumped in PLINK, with following parameters: clump p-value threshold of 1, LD

threshold for clumping of 0.2, and physical distance threshold for clumping of 500kb. The resulting valid SNPs are extracted and used to generate PRS. PRS are generated using PLINK for following p-value ranges: 0-0.001, 0-0.01, 0-0.05, 0-0.1, 0-0.2, 0-0.3, 0-0.4, 0-0.5, 0-0.75, and 0-1. Due to sample size restrictions (at least 100 cases), the PGC PRS analyses were limited to EUR and MEGA cohorts, and other PRS analyses where AMR and AFR cohorts are included should be interpreted with caution. For each cohort and trait of interest, PRS for all thresholds are scaled, followed by PCA.

Loadings are calculated and used to derive PRS-PCA scores for each individual (here I only focus on PRS-PCA from the loadings of principal components 1 and 2). Subsequently, for each target trait (nOCD, bOCD, bnOCD, or OCS) and ancestral cohort (AMR, AFR, EUR, or MEGA), and discovery sample (nOCD, bOCD, OCS from AMR, AFR, EUR, or MEGA, and PGC studies), Nagelkerke's pseudo R_2, R_N^2 was calculated. I focused on only OCS as its relationship with OCD was stronger than that of OCP. To calculate R_N^2, I first fit a null model regressing target trait onto sex, the first 20 principal components for population stratification control, and batch. Then I fit a full model including all covariates from above adding PRS-PCA1 and PRS-PCA2 scores. Then, I used the likelihoods from these two models to derive R_N^2. For each scenario, OR, SE_{OR}, and p for PRS are taken from the full model for the PRS-PCA term. p values are then FDR corrected. The repeated subsampling analysis is done by deriving R_N^2 for a subset of randomly drawn 120 cases and 600 controls, 500 times. If the target trait is coming from binomial distribution (nOCD, bOCD, bnOCD) then the Gaussian regression is used, if the target trait is coming from Poisson distribution (OCS), then the Poisson

regression is used. R_N^2 can be mathematically represented according to the Equation 5-3.

$$R_N^2 = \frac{1 - \left(\frac{L(M_{NULL})}{L(M_{FULL})}\right)^{2/N}}{1 - L(M_{NULL})^{2/N}} \quad (5\text{-}3)$$

PGC summary statistics were obtained from the PGC web site at www.med.unc.edu/pgc/download-results and include: ADHD (Demontis et al., 2019), AN (Watson et al., 2019), a compound anxiety disorder ccGWAS (ANX; Otowa et al., 2016), ASD (Grove et al., 2019), BPD (Mullins et al., 2021), a cross disorder / psychopathology phenotype consisting of 11 psychiatric disorders, including OCD and TS (Lee et al., 2019), MDD (Howard et al., 2019), OCD (IOCDF-GC & OCGAS, 2017), PD (Forstner et al., 2021), PTSD (Nievergelt et al., 2019), SCZ (Pardiñas et al., 2018), and TS (Yu et al., 2019).

Heritability and Genetic Correlations

For within-sample analyses, heritabilites were estimated using Genome-wide Complex Trait Analysis (GCTA) and individual-level data (Yang et al., 2011). Using the same package, per-ancestry genetic correlations between bnOCD and OCS were also estimated (Lee et al., 2012). Analyses were limited to the EUR and MEGA cohorts due to sample size constraints.

For inter-sample analyses, heritability was estimated from MEGA and EUR bnOCD, MEGA and EUR OCS, and PGC GWASes summary statistics using LDSC (Bulik-Sullivan, Loh, et al., 2015). The same software was used to calculate the heritabilites and genetic correlations between EUR bnOCD and OCS, MEGA bnOCD and OCS, and PGC GWASes (Bulik-Sullivan, Finucane, et al., 2015).

Admixture Analysis

To assess unbiased estimation of ancestry patterns, the software ADMIXTURE was used to run analysis on LD-independent ($r^2 < 0.1$) loci for ancestry composition (Alexander et al., 2009), assuming 6 independent contributing populations (this being a numeric parameter passed into the software instructing the number of clusters to consider).

Results

nOCD ccGWAS

Across all ancestry cohorts (AMR, AFR, EUR, MEGA) examined, no genome-wide significant associations were found to associate with the narrow definition of OCD. Table 5-3 summarizes λ_{GC} for all GWASes, and both the whole SNP set and genotyped-only SNPset. Figure 5-14 summarizes Manhattan plots for the nOCD GWASes for all ancestry cohorts; no immediately apparent patterns that are common across ancestry cohorts can be noticed. QQ plots, Figure 5-15, for the nOCD GWASes show p-value distributions that follow expectations, without extreme biases or deviations.

Only one locus associated with nOCD at $p < 10^{-5}$ in 2 different ancestral cohorts (Table 5-4, Figure 5-20), namely rs76846589 from AMR (chr4:180,653,660, $p = 5.77 \times 10^{-6}$) in a 130kb proximity to rs55747917 from AFR (chr4:180,784,550, $p = 2.51 \times 10^{-6}$). Both SNPs are located in a relatively conserved intergenic region of the chromosome 4, yet neither SNP is known to be associated with any significant clinical or non-clinical phenotype in previous studies.

GO analysis of genes overlapped by markers with GWAS $p < 10^{-5}$ identified neuron to neuron synapse cellular components as significantly enriched at $p_{FDR} = 4.99 \times 10^{-2}$ in the AMR cohort, however that was the only significant finding across all 4

GWASes on nOCD (Table 5-5). The contributing genes were *ACTR2*, *MAGI2*, and *ALS2*, all three of which are expressed in the brain.

bnOCD ccGWAS

Across all ancestry cohorts (AMR, AFR, EUR, MEGA) examined, no genome-wide significant associations were found to associate with the wide (bnOCD), broad (bOCD) or narrow (nOCD) definition of OCD. Table 5-3 summarizes λ_{GC} for all GWASes, and both the whole SNP set and genotyped-only SNPset. Figure 5-16 summarizes Manhattan plots for the nOCD GWASes for all ancestry cohorts; no immediately apparent patterns that are common across ancestry cohorts can be noticed. QQ plots, Figure 5-17, for the nOCD GWASes show p distributions that follow expectations, without extreme biases or deviations.

No loci associated with bnOCD at $p < 10^{-5}$ were replicated in 2 independent ancestral cohorts. One locus, overlapping *CDKAL1* gene on chromosome 6 was in the same p-value trench in EUR cohort, and then also in a larger MEGA cohort, however EUR individuals are all included in MEGA as well, so this is not valid replication of association as much as it is an association robust to sample expansion (Table 5-4, Figure 5-20). The strongest hits in these two loci were rs36045545 from EUR (chr6:20,649,111, $p = 6.69 \times 10^{-6}$), which is about 11kb downstream to rs13203361 from MEGA (chr6:20,661,021, $p = 8.86 \times 10^{-5}$). CDKAL1 transcripts can be detected in brain tissues, but no associations with neurodevelopmental or psychiatric disorders have been noted. GO analysis yielded no significant associations after controlling for FDR (Table 5-5).

OCS qGWAS

Across all ancestry cohorts (AMR, AFR, EUR, MEGA) examined, no genome-wide significant associations were found to associate with the OCS phenotype. Table 5-3 summarizes λGC for all GWASes, including both the whole SNP set and genotyped-only SNPset. Figure 5-18 summarizes Manhattan plots for the nOCD GWASes for all ancestry cohorts; no immediately apparent patterns that are common across ancestry cohorts can be noticed. QQ plots, Figure 5-19, for the nOCD GWASes show p-value distributions that follow expectations, without extreme biases or deviations.

No loci associated with OCS at $p < 10^{-5}$ were replicated in 2 independent ancestral cohorts. Two loci separated by about 545kb can be found in AFR and EUR cohorts, but they are separated by a region of high recombination rate, so this is unlikely to be a replication. One locus found in the EUR cohort remained in the $p < 10^{-5}$ trench, with lowest p-values belonging to rs2418954 from EUR (chr10:108,892,622, $p = 5.54 \times 10^{-6}$), which is about 14kb upstream to rs7089127 from MEGA (chr10:108,878,616, $p = 7.47 \times 10^{-5}$).

GO analysis of the AMR cohort resulted in enrichment of SNPs in the $p < 10^{-5}$ trench overlapping genes involved in the integral component of lumenal side of endoplasmic reticulum (ER) membrane ($p_{FDR} = 1.29 \times 10^{-4}$) ER to Golgi transport vesicle membrane ($p_{FDR} = 5.39 \times 10^{-4}$), clathrin-coated endocytic vesicle membrane ($p_{FDR} = 6.54 \times 10^{-4}$) trans-Golgi network membrane ($p_{FDR} = 1.28 \times 10^{-3}$), and lysosomal membrane ($p_{FDR} = 3.54 \times 10^{-2}$) cellular compartment ontologies. Other cohorts resulted in no significant associations.

Cross-OCRD Trait GWAS

There were overall low rates of either within-disorder cross-ancestry (Table 5-4, Figure 5-20) or within-ancestry cross-disorder (Table 5-4, Figure 5-21) overlap in loci with GWAS $p < 10^{-5}$. One locus on chromosome 16 was found in this trench in AFR nOCD (rs887523, chr16:5,705,280, $p = 3.25 \times 10^{-6}$), AFR bnOCD (rs62016433, chr16:5,525,579, $p = 3.19 \times 10^{-6}$), and AMR OCS (rs12325273, chr16:7,253,286, $p = 2.54 \times 10^{-6}$). In all three GWASes, the SNP clusters overlapped with the *RBFOX1* gene (Figure 5-22). *RBFOX1* is ubiquitously expressed across the central nervous system and skeletal, but not cardiac and smooth, muscle (Figure 5-23). *RBFOX1* is an important gene implicated in numerous neurological and psychiatric conditions, including spinocerebellar ataxia, autism, developmental coordination disorder, and epilepsy (GeneCards, n.d.).

OCRD Trait PRS Analysis

Within-sample cross-trait PRS analysis yielded no statistically significant predictions, regardless of the discovery or target phenotype or ancestry (Figure 5-24). Among the nOCD discovery analyses, the strongest associations in terms of amount of variance explained, were seen when using PRS generated from nOCD AFR GWAS to predict nOCD AMR case status, at $R_N^2 = 0.0365$, $p_{FDR} = 0.7715$. Other noteworthy associations were between the nOCD EUR discovery, and nOCD AMR ($R_N^2 = 0.0311$, $p_{FDR} = 0.8252$) and nOCD AFR $R_N^2 = 0.0280$, $p_{FDR} = 0.7715$) target samples. These associations were also the strongest associations overall. nOCD discovery performed better than bnOCD discovery, despite their larger sample sizes (Table 5-1).

Using the OCS trait as the discovery sample also yielded no statistically significant associations. The strongest associations in terms of R_N^2 were from the qOCS

AMR discovery with nOCD AFR (R_N^2 = 0.0295, p_{FDR} = 0.7715) and nOCD AMR as the target samples (R_N^2 = 0.0191, p_{FDR} = 0.9178), and from the qOCS EUR discovery with nOCD AFR target (R_N^2 = 0.0266, p_{FDR} = 0.8307).

PGC PRS Analysis

PGC PRS analysis of cross-disorder diagnoses yielded no significant associations between PGC GWAS summary statistics PRS scores and OCD status (either nOCD, bOCD, or bnOCD) in the ABCD EUR and META cohorts after controlling for multiple testing. Overall, PCG PRS scores were better at predicting nOCD than bOCD or bnOCD, particularly in the EUR cohort. This effect was particularly pronounced in the ADHD, AN, ASD, BPD, MDD, OCD, PD, and SCZ derived PRS scores (Figure 5-25, 5-26). The opposite was true of ANX, cross-disorder / psychopathology, PTSD, and TS derived PRS scores (Figure 5-25, 5-26).

Nominally significant associations included MDD-derived PRS predicting into MEGA nOCD (R_N^2 = 0.0084, $p_{NOMINAL}$ = 0.0292), bOCD (R_N^2 = 0.0031, $p_{NOMINAL}$ = 0.0056), bnOCD (R_N^2 = 0.0031, $p_{NOMINAL}$ = 0.0014) and EUR bOCD (R_N^2 = 0.0038, $p_{NOMINAL}$ = 0.0211), bnOCD (R_N^2 = 0.0042, $p_{NOMINAL}$ = 0.0043); SCZ-derived PRS predicting into EUR bOCD (R_N^2 = 0.0038, $p_{NOMINAL}$ = 0.0205), bnOCD (R_N^2 = 0.0043, $p_{NOMINAL}$ = 0.0042); AN-derived PRS predicting into MEGA bOCD (R_N^2 = 0.0021, $p_{NOMINAL}$ = 0.0292), bnOCD (R_N^2 = 0.0019, $p_{NOMINAL}$ = 0.0169) and EUR bOCD (R_N^2 = 0.0040, $p_{NOMINAL}$ = 0.0170), bnOCD (R_N^2 = 0.0037, $p_{NOMINAL}$ = 0.0088); OCD-derived PRS predicting into MEGA bOCD (R_N^2 = 0.0025, $p_{NOMINAL}$ = 0.0141), bnOCD (R_N^2 = 0.0020, $p_{NOMINAL}$ = 0.0146); PTSD-derived PRS predicting into MEGA bOCD (R_N^2 = 0.0025, $p_{NOMINAL}$ = 0.0152), bnOCD (R_N^2 = 0.0015, $p_{NOMINAL}$ = 0.0436); ANX-derived PRS predicting into MEGA bOCD (R_N^2 = 0.0023, $p_{NOMINAL}$ = 0.0190), bnOCD (R_N^2 =

0.0019, $p_{NOMINAL}$ = 0.0195); TS-derived PRS predicting into MEGA bOCD (R_N^2 = 0.0023, $p_{NOMINAL}$ = 0.0221), bnOCD (R_N^2 = 0.0018, $p_{NOMINAL}$ = 0.0212); and cross-disorder / psychopathology derived PRS predicting into EUR bOCD (R_N^2 = 0.0034, $p_{NOMINAL}$ = 0.0301), bnOCD (R_N^2 = 0.0025, $p_{NOMINAL}$ = 0.0400). However, none of these associations were significant after multiple testing corrections, controlling for the FDR.

Repeated undersampling analysis shows large variability in R_N^2 with respect to target sample sizes (Figure 5-26). The higher R_N^2 in nOCD compared to bOCD and bnOCD remains among ASD, MDD, and SCZ - indicating effects of discovery sample sizes on PRS modeling.

Within Sample Heritability and Genetic Correlations

REML analysis is reported in Table 5-6. Briefly, only the EUR OCS trait among bnOCD individuals was successfully calculated at h_{SNP}^2 = 0.00153 and SE_{SNP} = 0.04872 on the liability scale. Due to low sample size (especially cases) and sample heterogeneity, h_{SNP}^2 could not be calculated for OCD traits. Hence, r_g could also not be calculated in this sample.

PGC Heritability and Genetic Correlations

LDSC heritability estimates for ABCD data returned negative heritabilites for all ABCD GWASes: EUR bnOCD and OCS, and MEGA bnOCD and OCS. As this heritability is out of bounds (minimum possible being 0), r_g could also not be calculated in these analyses.

Admixture Analysis

Admixture analysis shows low level of non-European admixture in EUR cohort (Figure 5-27). Low-to-moderate levels of non-African admixture were present in AFR

cohort, primarily from Europeans. High levels of non-American admixture were present in the AMR cohort, primarily coming from Europeans then Africans.

Discussion

The ABCD Study is a longitudinal population study of American adolescents, representing one of the most diverse samples for genetic study to date – with nationally representative sample in terms of ancestry/ethnicity and sex. The ABCD Study also features comprehensive phenotyping data and neuroimaging data, in addition to the genetic data, making it a treasure trove for various genomic explorations.

ccGWAS of both nOCD and bnOCD resulted in no genome-wide significant associations, which was expected to be the case given a) low sample sizes, b) phenotypically heterogenous samples in bnOCD, c) highly stratified sample, and d) high polygenicity and low penetrance of OCD variants. QC pipeline has successfully controlled for most confounders, as evident by the Manhattan plots, QQ plots, and non-deviant inflation factors λ_{GC} (Table 5-3). GO analysis of genes overlapped by markers with $p_{GWAS} < 10^{-5}$ identified neuron to neuron synapse cellular components as significantly enriched at $p_{FDR} = 4.99 \times 10^{-2}$ in AMR cohort (Table 5-5), with all three contributing genes (*ACTR2*, *MAGI2*, and *ALS2*) expressed in the brain.

qGWAS of OCS also resulted in no genome-wide significant associations. This is likely due to underrepresentation of high OCS scores in the ABCD Study, i.e., extreme right-skew. Using Poisson linear mixed modelling helped control for this skew, but the number of individuals on the above-zero spectrum of OCS scores remains inadequate for a well-powered qGWAS. Across all GWASes, *RBFOX1* repeatedly shows trending association with OCD and related symptoms, namely in AFR nOCD, AFR bnOCD, and

AMR OCD cohorts (Figure 5-22). *RBFOX1* gene is an important RNA-processing factor with high expression in the brain (Figure 5-23).

PRS analysis of OCRD traits within the ABCD Study has shown nOCD to be a better predictor of OCRD traits than bnOCD, despite its derivation from a smaller sample GWAS (Figure 5-24). This is likely due to lower phenotype heterogeneity and more accurate phenotype classification in nOCD cohorts. bnOCD was also outperformed by qOCS. While a few associations were present at nominal p-value, no statistically significant associations were present after controlling for FDR.

PRS analysis of PGC traits shows PCG PRS scores to better predict nOCD than bOCD or bnOCD, particularly in the EUR cohort, especially when derived from ADHD, AN, ASD, BPD, MDD, OCD, PD, and SCZ summary statistics (Figure 5-25). Opposite was true of ANX, cross-disorder / psychopathology, PTSD, and TS derived PRS scores (Figure 5-25). These patterns indicate lower phenotype heterogeneity within the nOCD as compared to bOCD sample. Repeated under sampling somewhat accounts for these swings, indicating a need for larger sample sizes (Figure 5-26).

Attempts at heritability estimation failed with estimated heritability being effectively 0 (in case of GCTA REML analysis of individual data) or negative i.e., out of bounds (in case of LDSC analysis of summary statistics). Subsequently no genetic correlations r_g could be calculated. There are several possible reasons for these issues, namely 1) low sample size, 2) phenotype heterogeneity and misclassification, 3) high level of admixture and population stratification, and 4) use of LMM for association tests. LMM can increase power in genetic tests, especially when it comes to highly structured data in terms of relatedness and ancestry – such as ABCD Study. However, LMMs

have not been validated for the use in heritability and genetic correlation estimates. A potential work-around is to run a simple linear regression in addition to LMM, simply for the purposes of heritability and genetic correlation analysis.

Admixture analysis of the ABCD Study has shown the high rates of admixture, specifically in the AMR cohort. European admixture in AFR cohort can likely be attributed to normal admixture of European DNA in African Americans (AFR). Conversely, this can also be said of admixed Americans (AMR), as they too have naturally high levels of European and African admixture. Nonetheless, this population structure introduces additional confounding and variability that reduces power of association tests and limits inference.

Based on my genetic analyses, nOCD is indeed a better approximation of true OCD diagnosis than bnOCD. Unfortunately, due to very low sample sizes, specifically in terms of the number of cases, power for genetic analyses of OCD in ABCD Study is very limited. These analyses can, however, be meta-analyzed into larger consortia to increase power and diversity in samples. The possibility of local ancestry-based inferences has been explored; however, the small sample sizes lead to a high rate of spurious associations.

ABCD Study remains a valuable resource for genet studies. Additional longitudinal data will be useful for further phenotype refinement and improvements in association. Some methodological changes, like running simple logistic modelling, might enable post-GWAS analyses, like heritability and genetic correlation estimation.

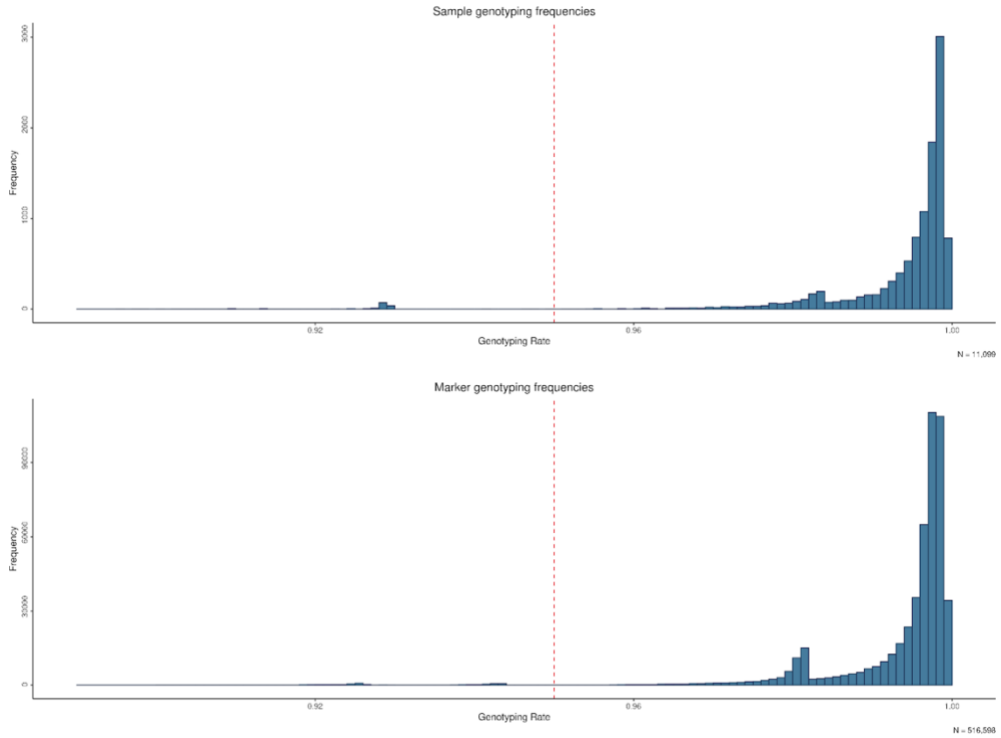

Figure 5-1. Sample (top) and marker/SNP (bottom) genotyping rates for ABCD study. Red lines represent *0.95* GR_S and GR_{SNP} cutoffs.

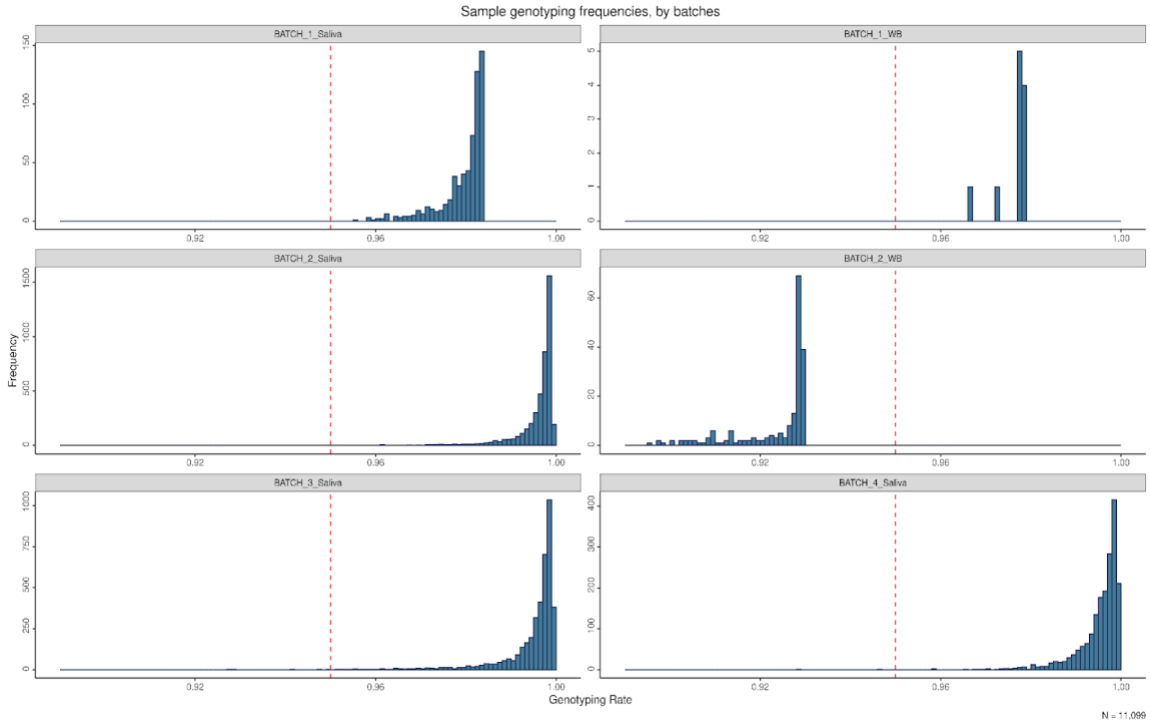

Figure 5-2. Batch-wise sample genotyping rates for ABCD study. Batches (top to bottom, left to right): saliva 1, whole blood 1, saliva 2, whole blood 2, saliva 3, and saliva 4. Red lines represent *0.95* GR_S cutoff.

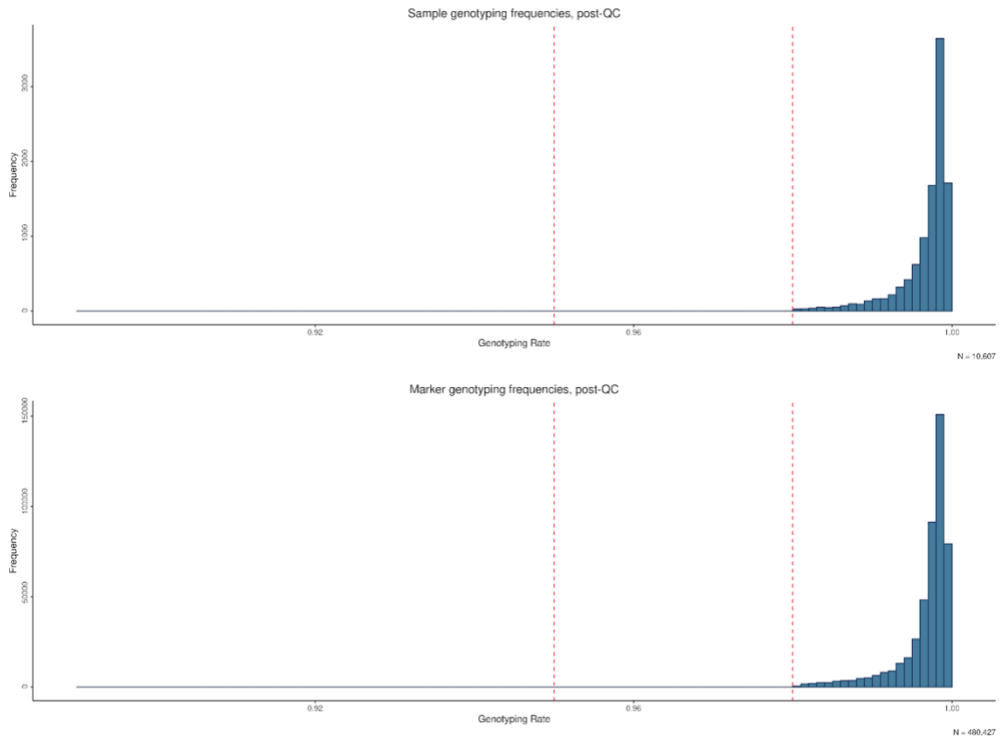

Figure 5-3. Post-QC sample (top) and marker/SNP (bottom) genotyping rates for ABCD study. Red lines represent *0.95* and *0.98* GR_S and GR_{SNP} cutoffs.

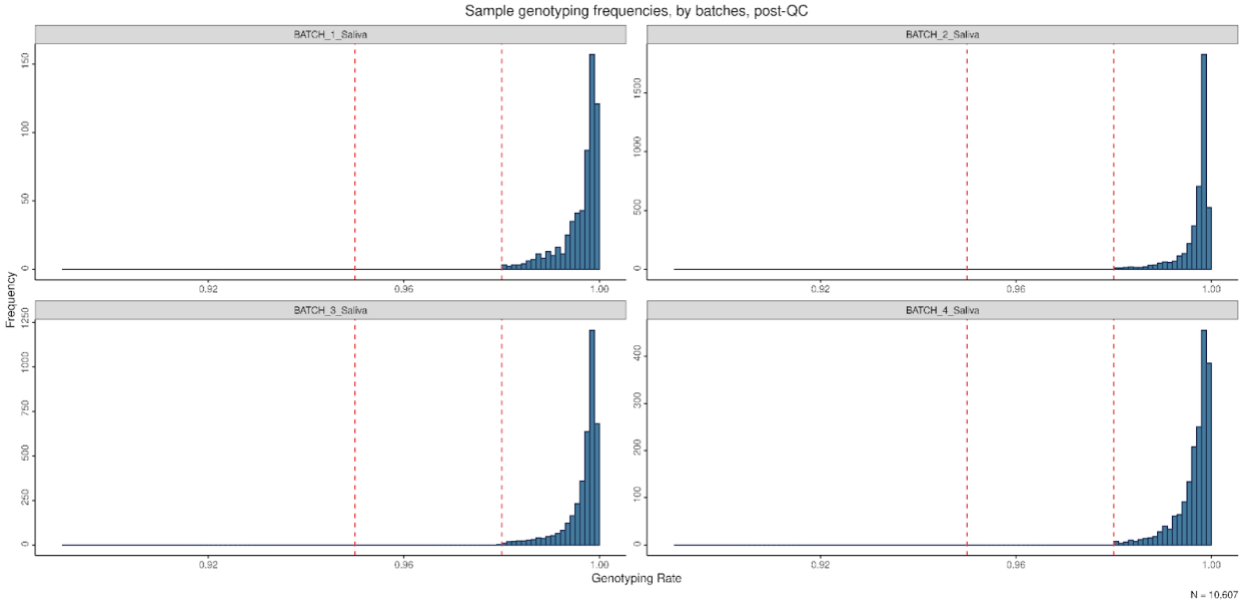

Figure 5-4. Post-QC batch-wise sample genotyping rates for ABCD study. Batches (top to bottom, left to right): saliva 1, saliva 2, saliva 3, and saliva 4. Red lines represent *0.95* and *0.98* GR_S cutoffs.

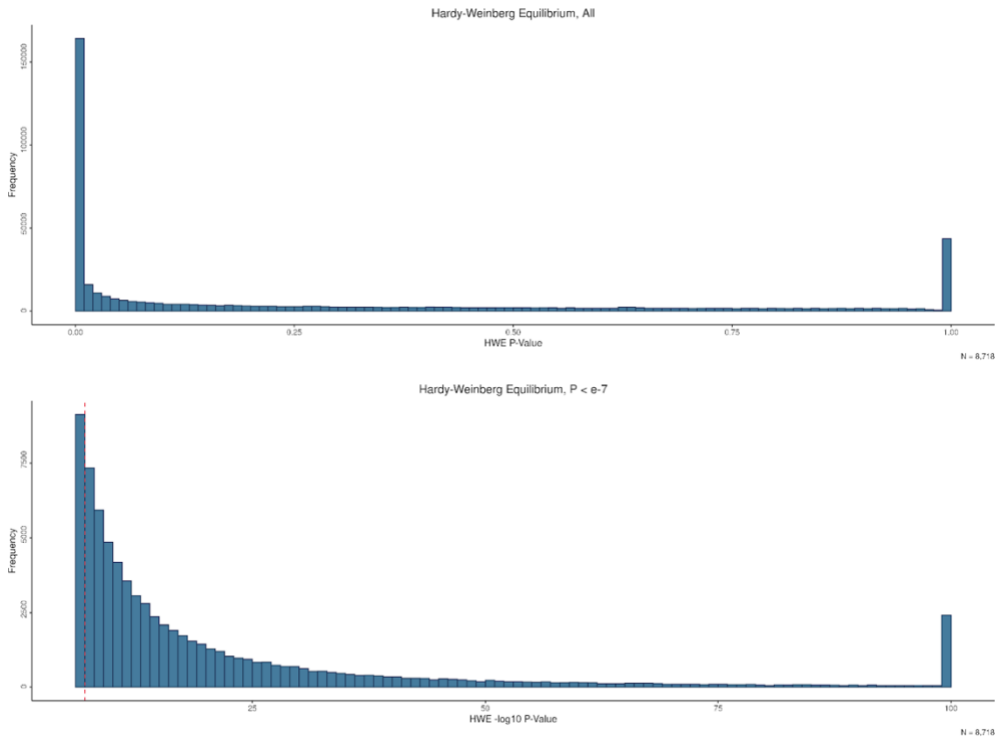

Figure 5-5. Distribution of p_{HWE} values in ABCD study. Distribution of all markers (top) and only those markers in high Hardy-Weinberg disequilibrium (bottom) with $p_{HWE} \leq 10^{-7}$.

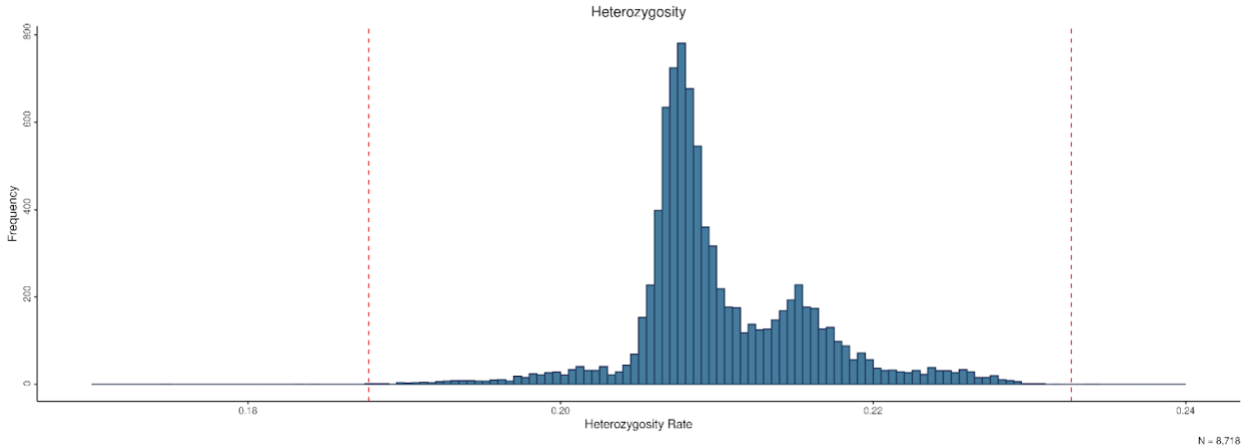

Figure 5-6. Distribution of F_{HET} values in ABCD study. Red lines indicate $mean(F_{HET}) \pm 4 * sd(F_{HET})$ cutoffs.

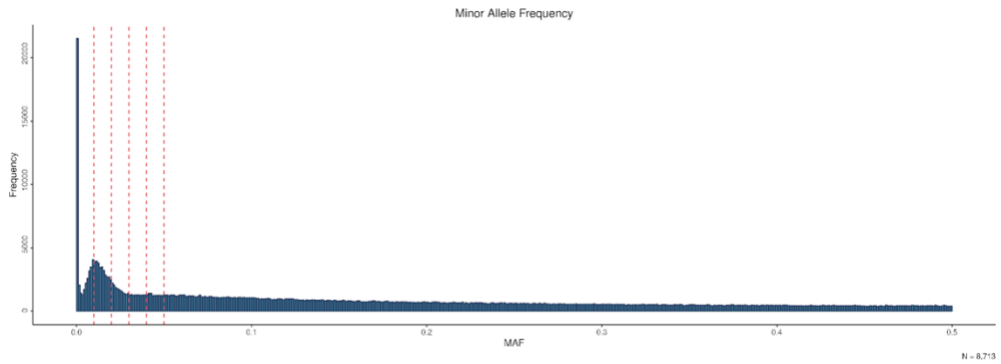

Figure 5-7. Distribution of *MAF* values in ABCD study. Red lines indicate cutoffs at *0.01, 0.02, 0.03, 0.04,* and *0.05*.

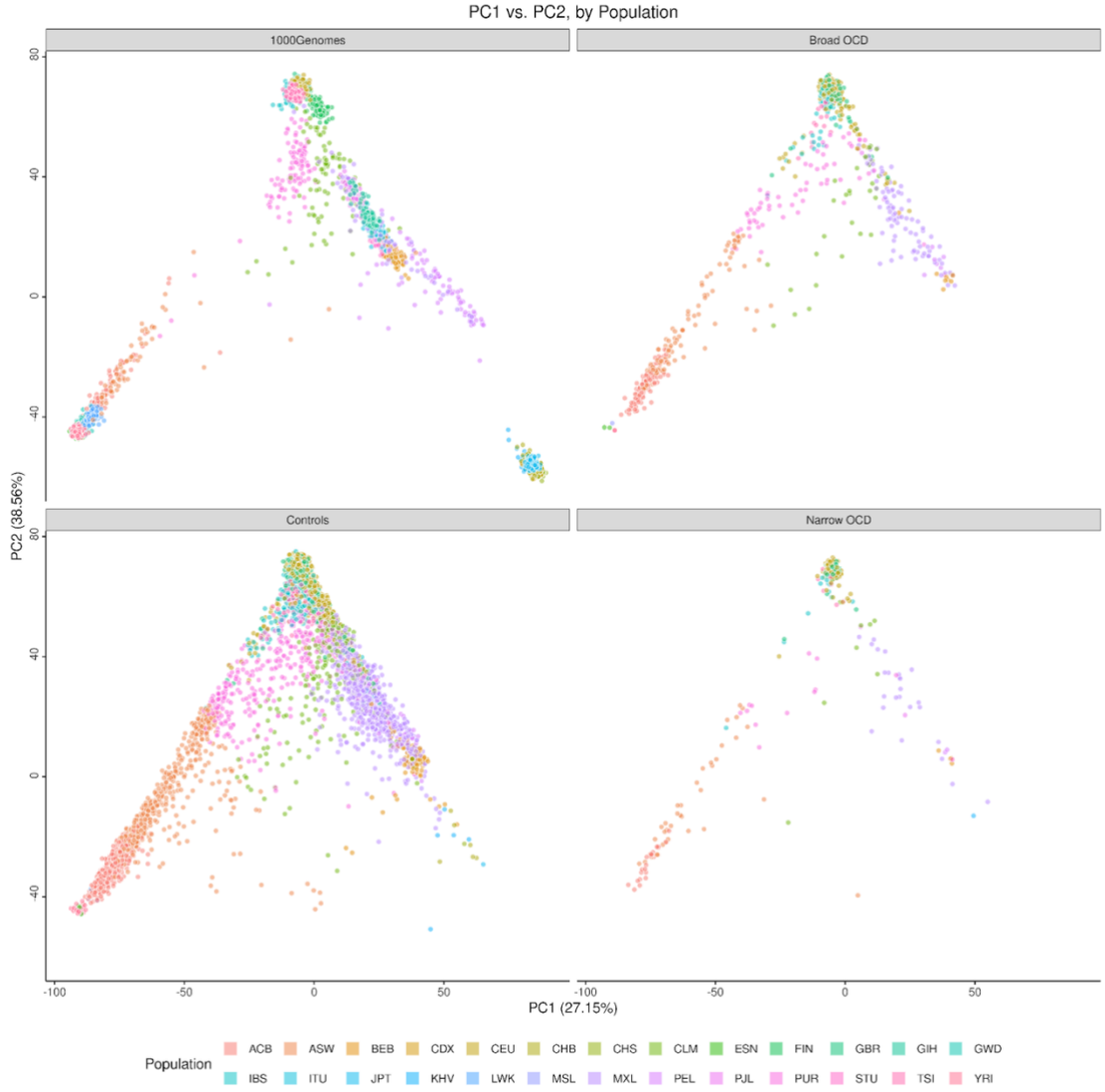

Figure 5-8. Population stratification of ABCD data compared to 1kGPp3 reference. Visualized are principal component 1 vs. 2 biplots, with respect to 26 populations explaining 65.71% of variance.

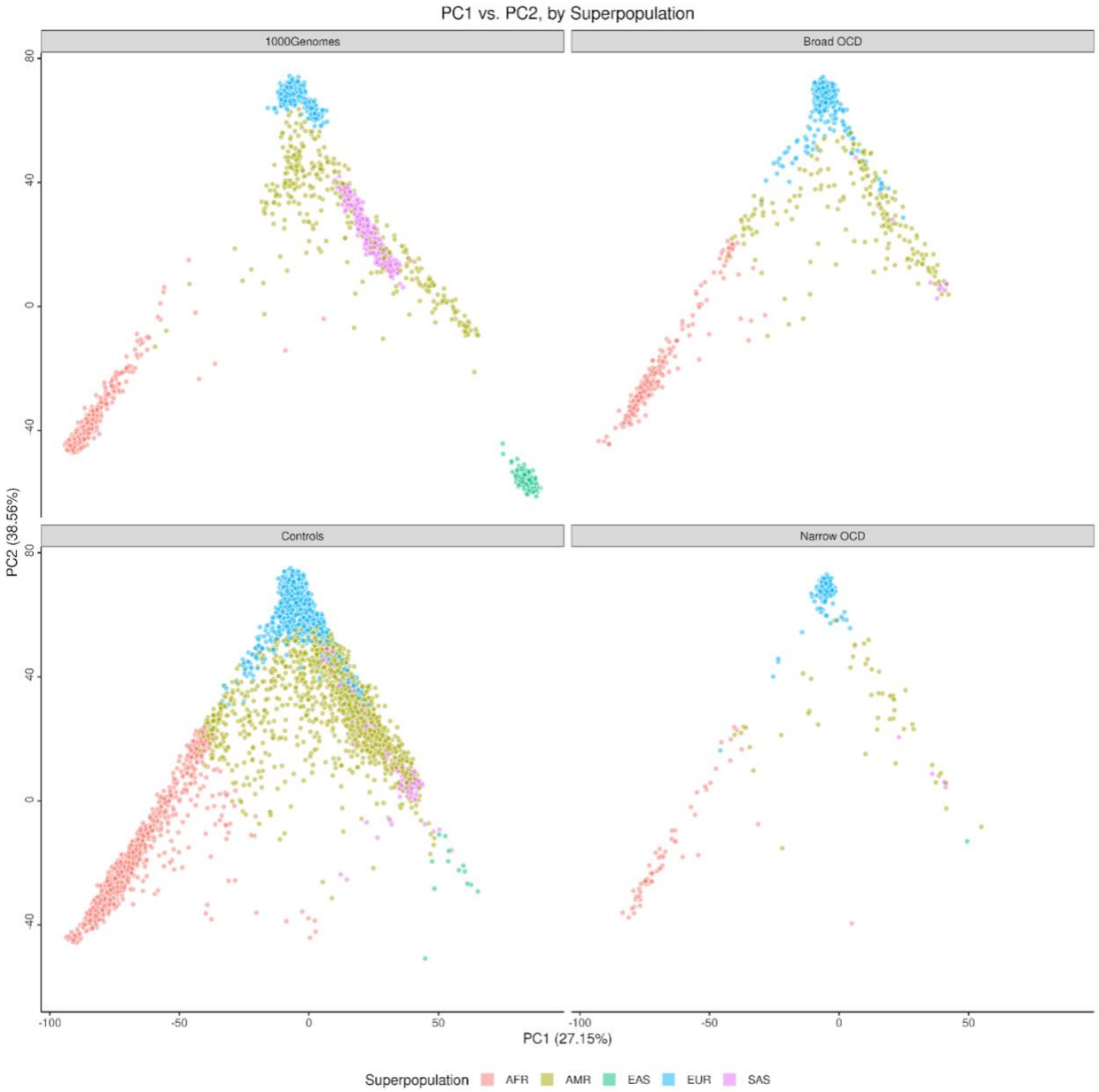

Figure 5-9. Superpopulation stratification of ABCD data compared to 1kGPp3 reference. Visualized are principal component 1 vs. 2 biplots, with respect to 5 superpopulations explaining 65.71% of variance.

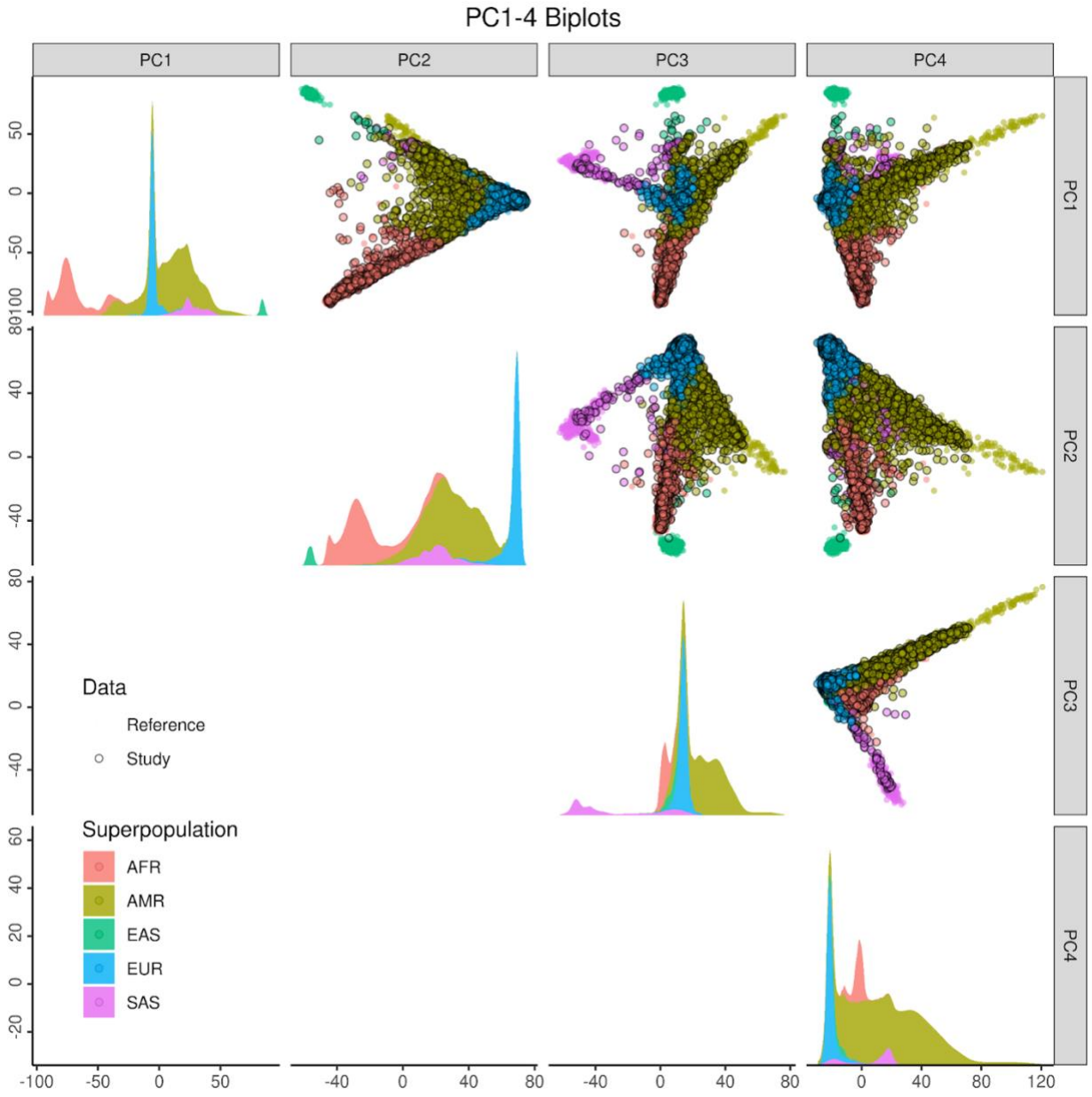

Figure 5-10. Population structure of ABCD study across the first 4 principal components. Colors represent superpopulations, whereas circles indicate ABCD study participants (any OCD status). The diagonal shows density plots for each principal component.

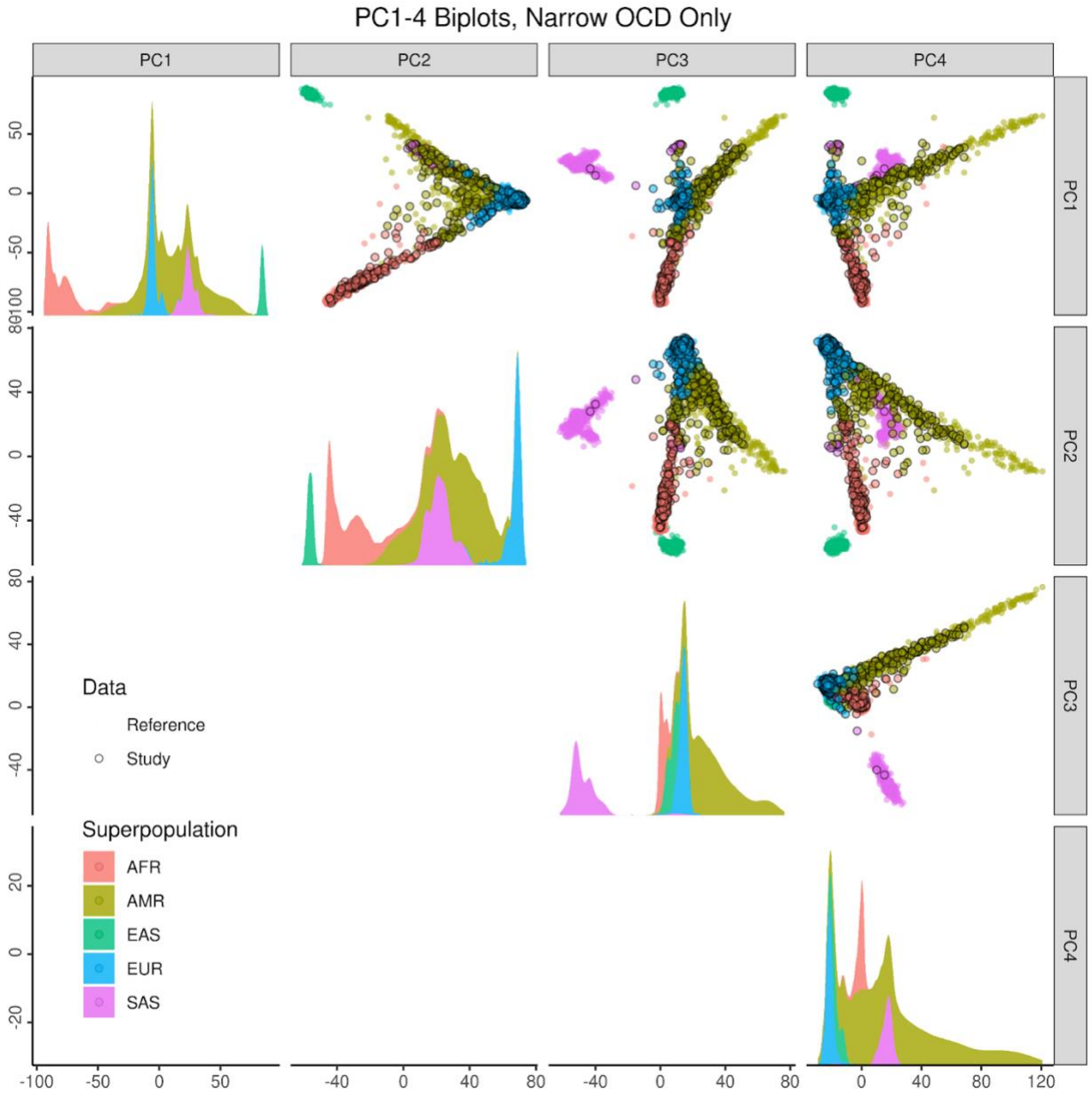

Figure 5-11. Population structure of nOCD samples across the first 4 principal components. Colors represent superpopulations, whereas circles indicate nOCD participants from the ABCD study. The diagonal shows density plots for each principal component.

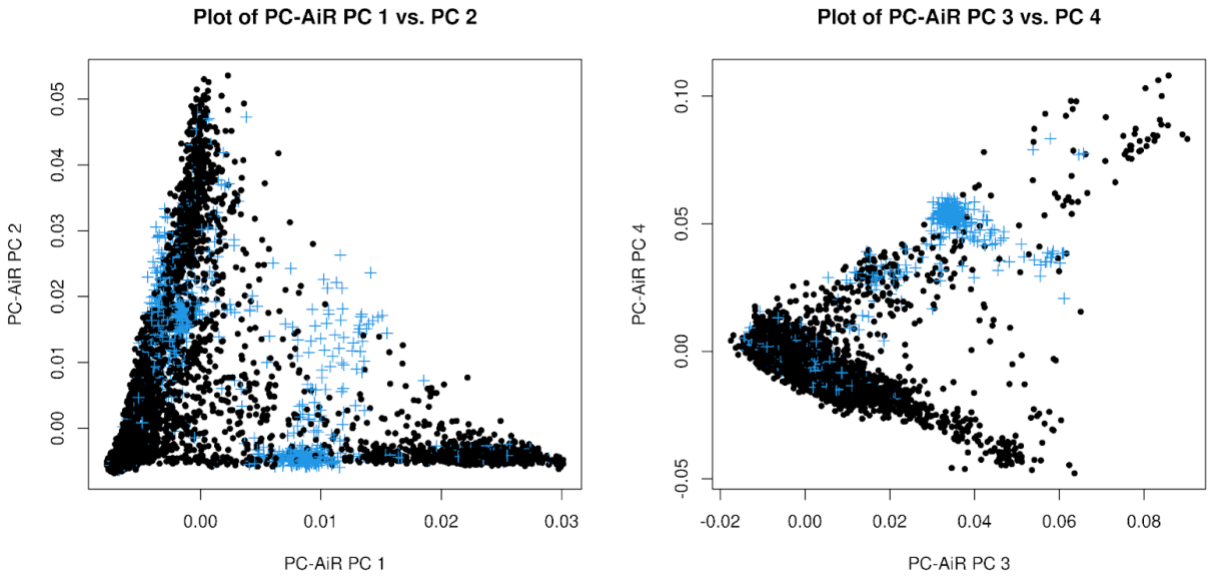

Figure 5-12. Relationship-aware population stratification biplots. Black dots represent individuals from unrelated subset, blue pluses represent individuals from the related subset.

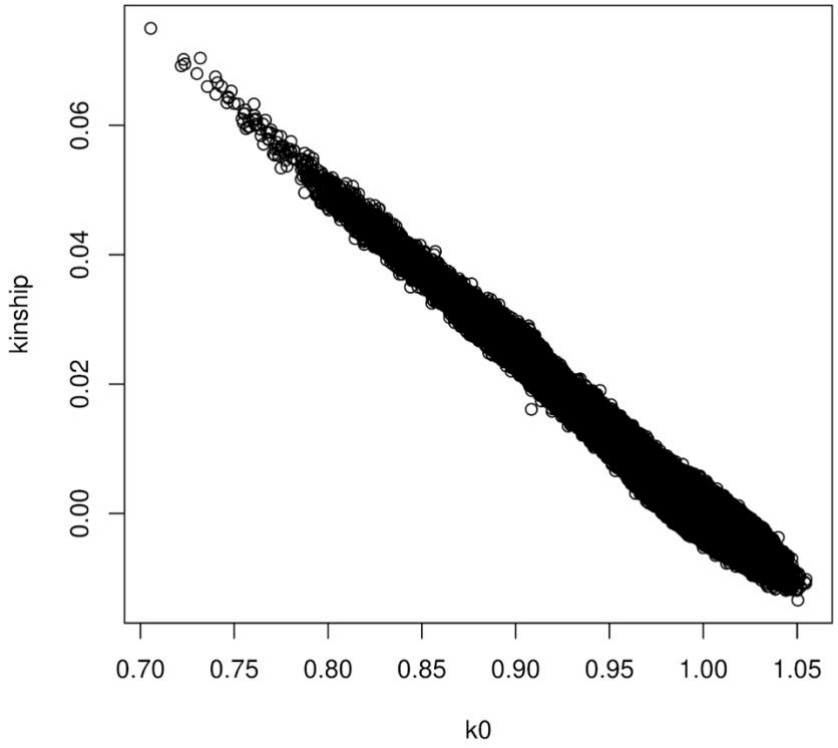

Figure 5-13. Population structure aware relationships. As expected, the less related individuals are, the higher proportion of no shared (k0) alleles.

Table 5-1. GWAS sample summaries.

Phenotype	Ancestry	N_{CASES}	$N_{CONTROLS}$	N_{TOTAL}	% Female
ccGWAS					
nOCD	AMR	49	196	245	38.78
	AFR	59	236	295	35.59
	EUR	155	660	825	35.19
	MEGA	273	1,092	1,365	35.92
bnOCD	AMR	239	956	1,195	47.70
	AFR	286	1,144	1,430	41.61
	EUR	820	3,280	4,100	42.22
	MEGA	1,345	5,380	6,725	43.06
qGWAS					
OCS	AMR	-	-	1,388	52.45
	AFR	-	-	1,400	48.14
	EUR	-	-	4,339	50.88
	MEGA	-	-	7,127	50.65

Table 5-2. Summary of PRS experiments.

Discovery	Target sample (*ABCD*)			
	nOCD	bOCD	bnOCD	OCS
ABCD				
nOCD	B	A	B	A
bOCD	A	B	B	A
OCS	A	B	B	B
PGC				
ADHD	A	A	A	A
AN	A	A	A	A
ANX	A	A	A	A
ASD	A	A	A	A
BPD	A	A	A	A
CD	A	A	A	A
MDD	A	A	A	A
OCD	A	A	A	A
PD	A	A	A	A
PTSD	A	A	A	A
SCZ	A	A	A	A
TS	A	A	A	A

A: target samples include all ancestry cohorts. B: target samples only include non-matching ancestry cohorts.

Table 5-3. Summary of GWAS genomic inflation factor, λ_{GC}.

Phenotype	Ancestry	N_{SNP}	λ_{GC}	$\lambda_{GC\text{-}GENOTYPED}$
nOCD	AMR	5,322,421	1.008	0.996
	AFR	5,322,421	1.061	1.080
	EUR	5,322,421	0.993	0.987
	MEGA	5,322,421	1.003	0.990
bnOCD	AMR	5,322,421	0.994	0.993
	AFR	5,322,421	1.000	1.007
	EUR	5,322,421	0.987	0.998
	MEGA	5,322,421	0.979	0.977
OCS	AMR	5,322,421	1.008	1.004
	AFR	5,322,421	0.998	0.990
	EUR	5,322,421	0.981	0.986
	MEGA	5,322,421	0.988	0.982

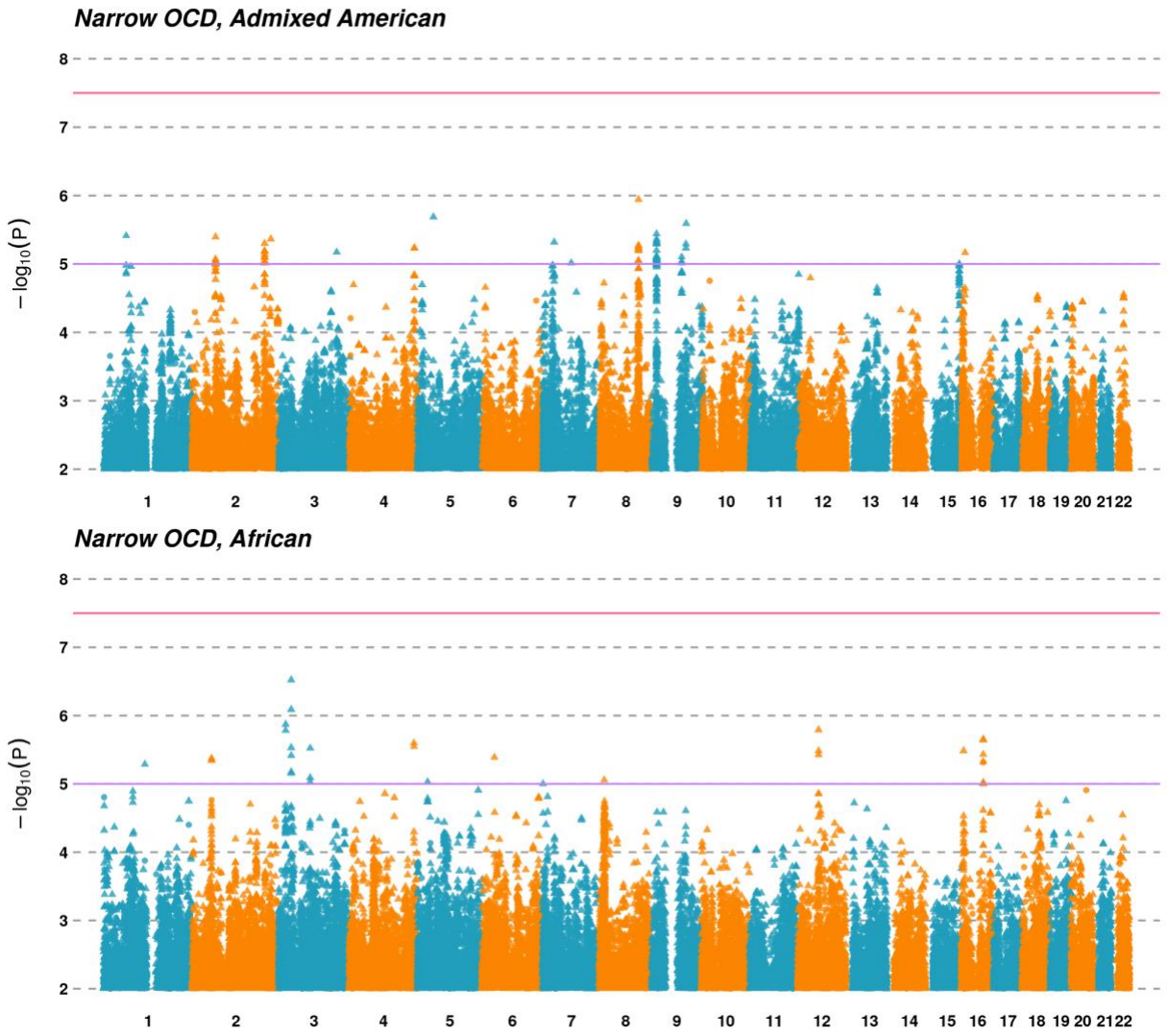

Figure 5-14. nOCD GWAS Manhattan plots.

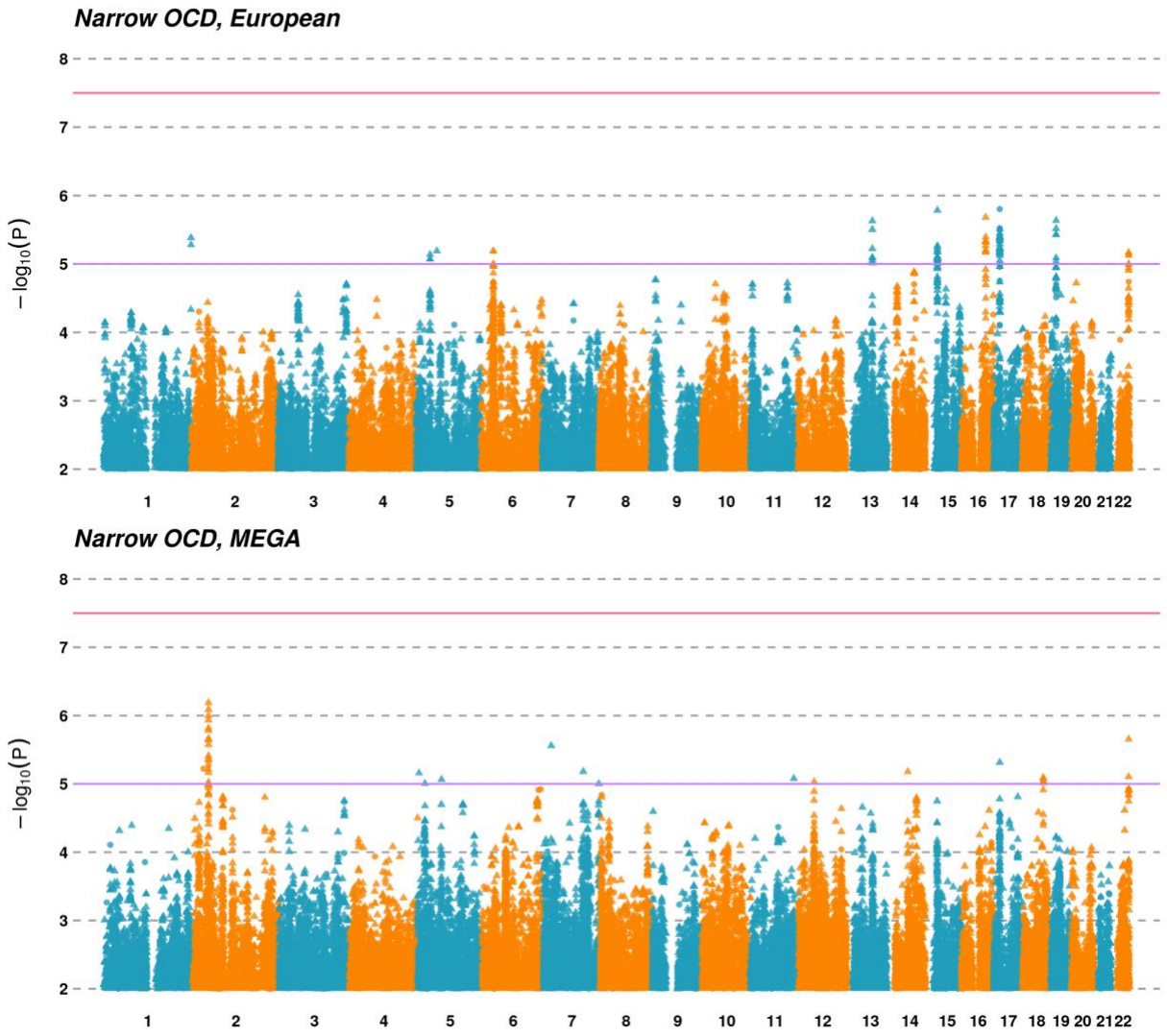

Figure 5-14. Continued.

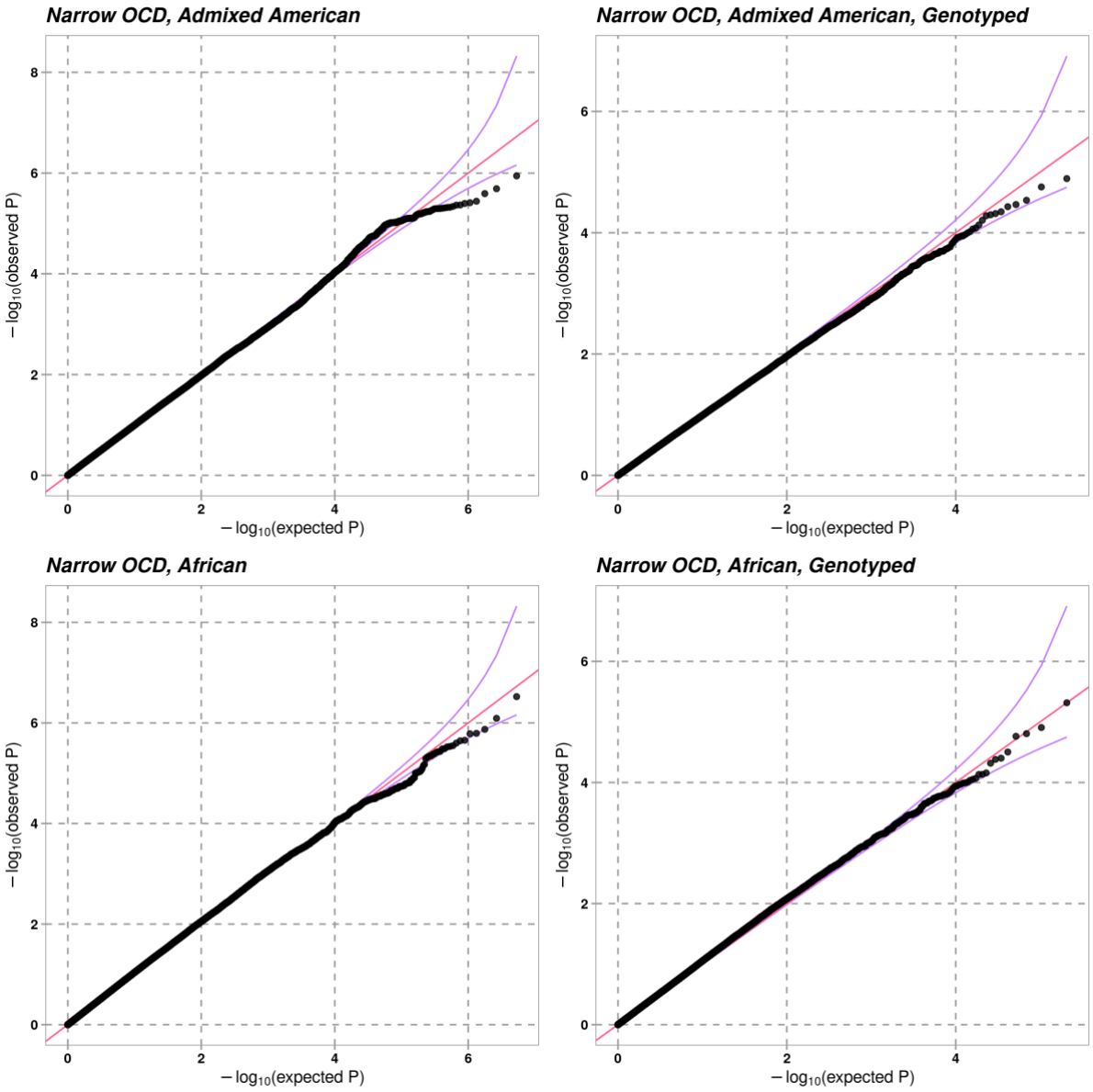

Figure 5-15. nOCD GWAS QQ plots.

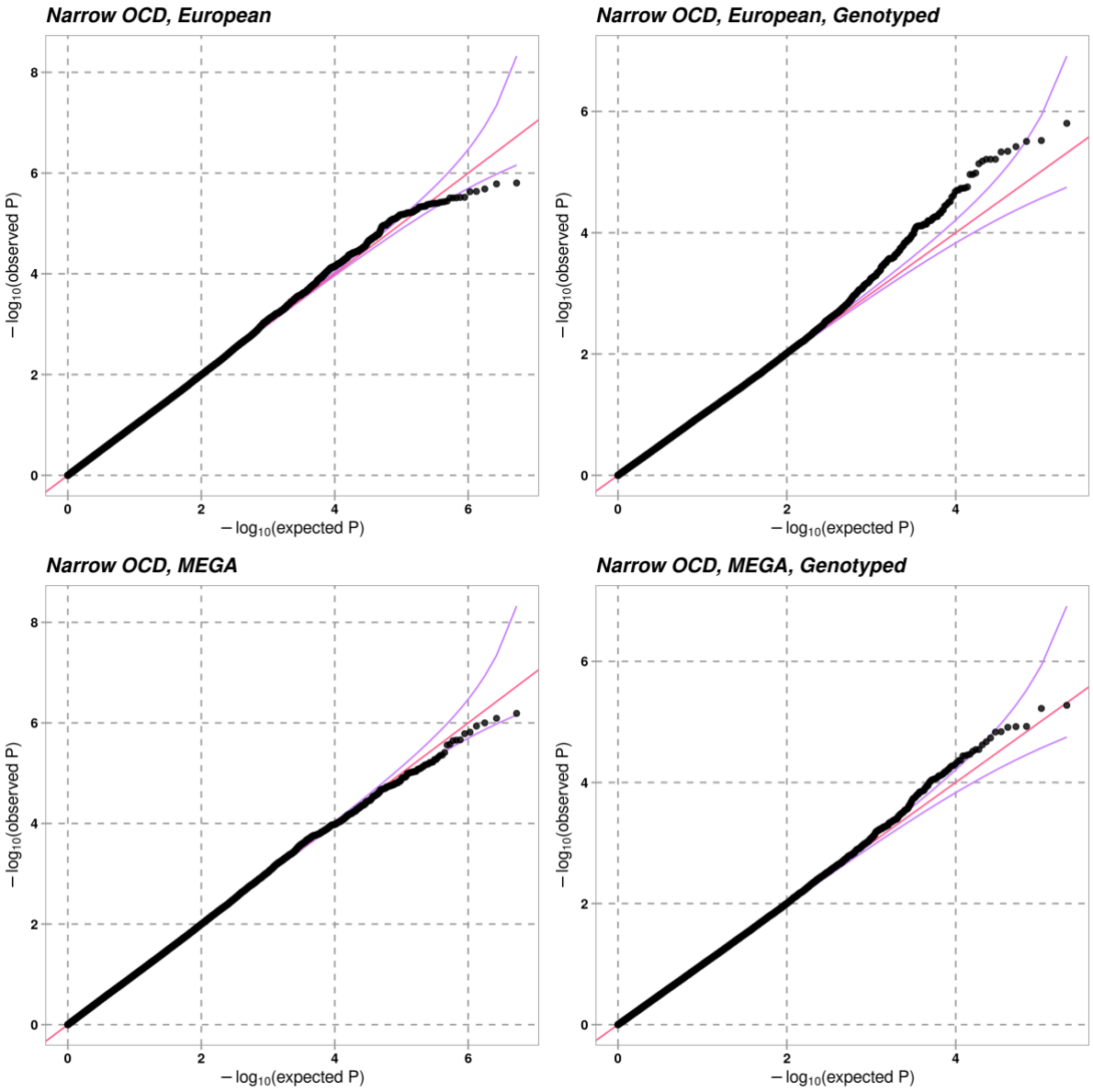

Figure 5-15. Continued.

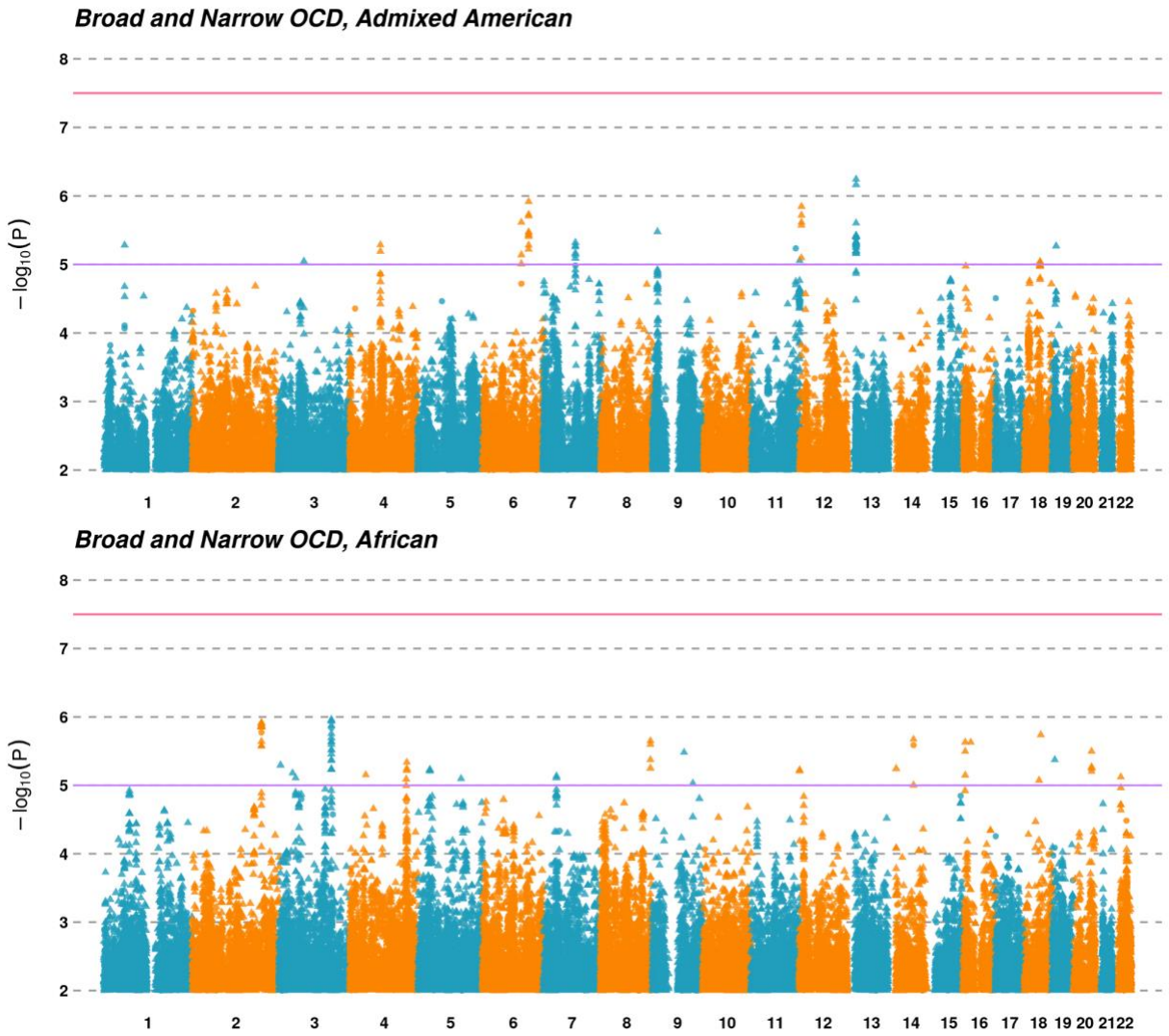

Figure 5-16. bnOCD GWAS Manhattan plots.

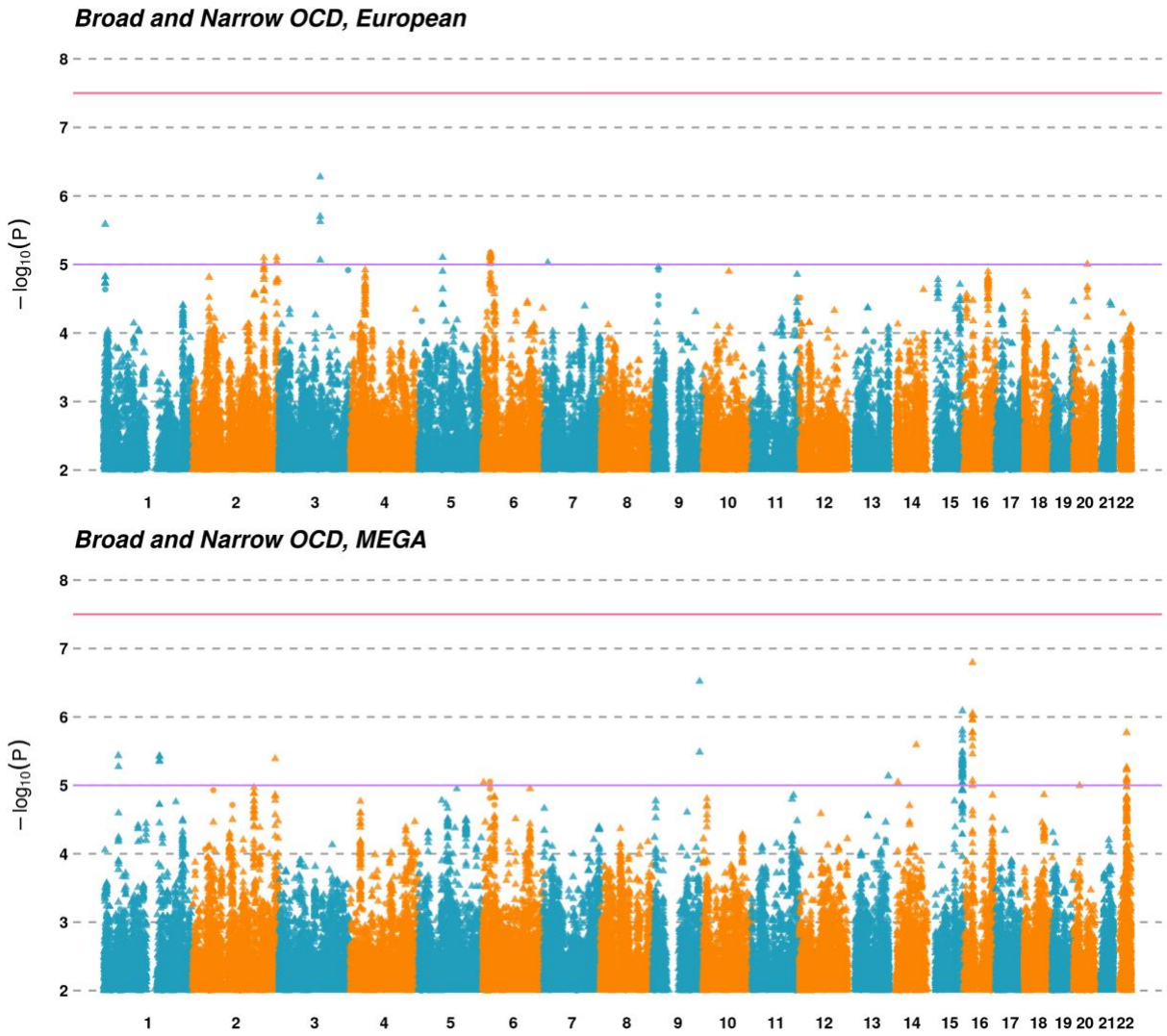

Figure 5-16. Continued.

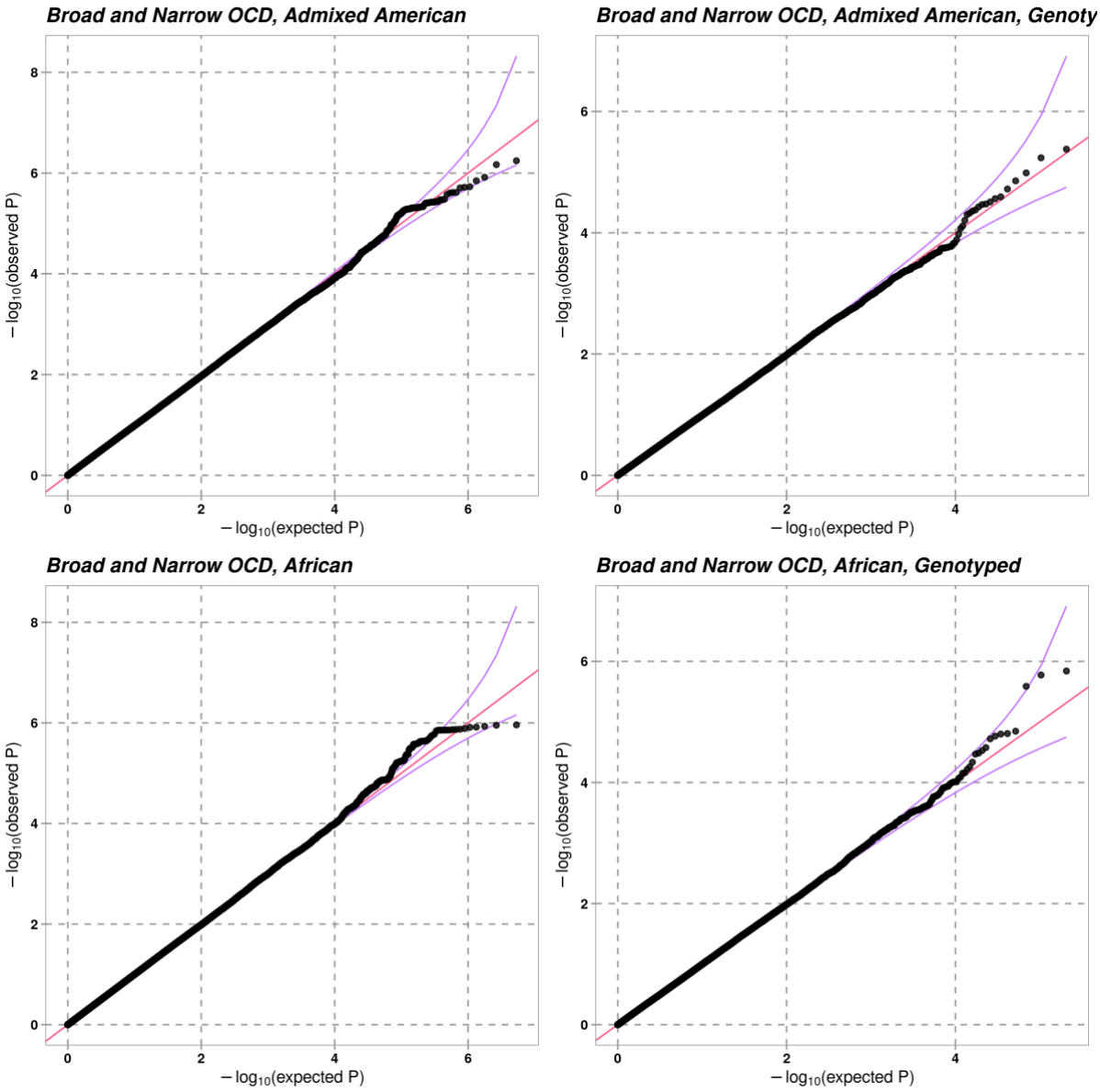

Figure 5-17. bnOCD GWAS QQ plots.

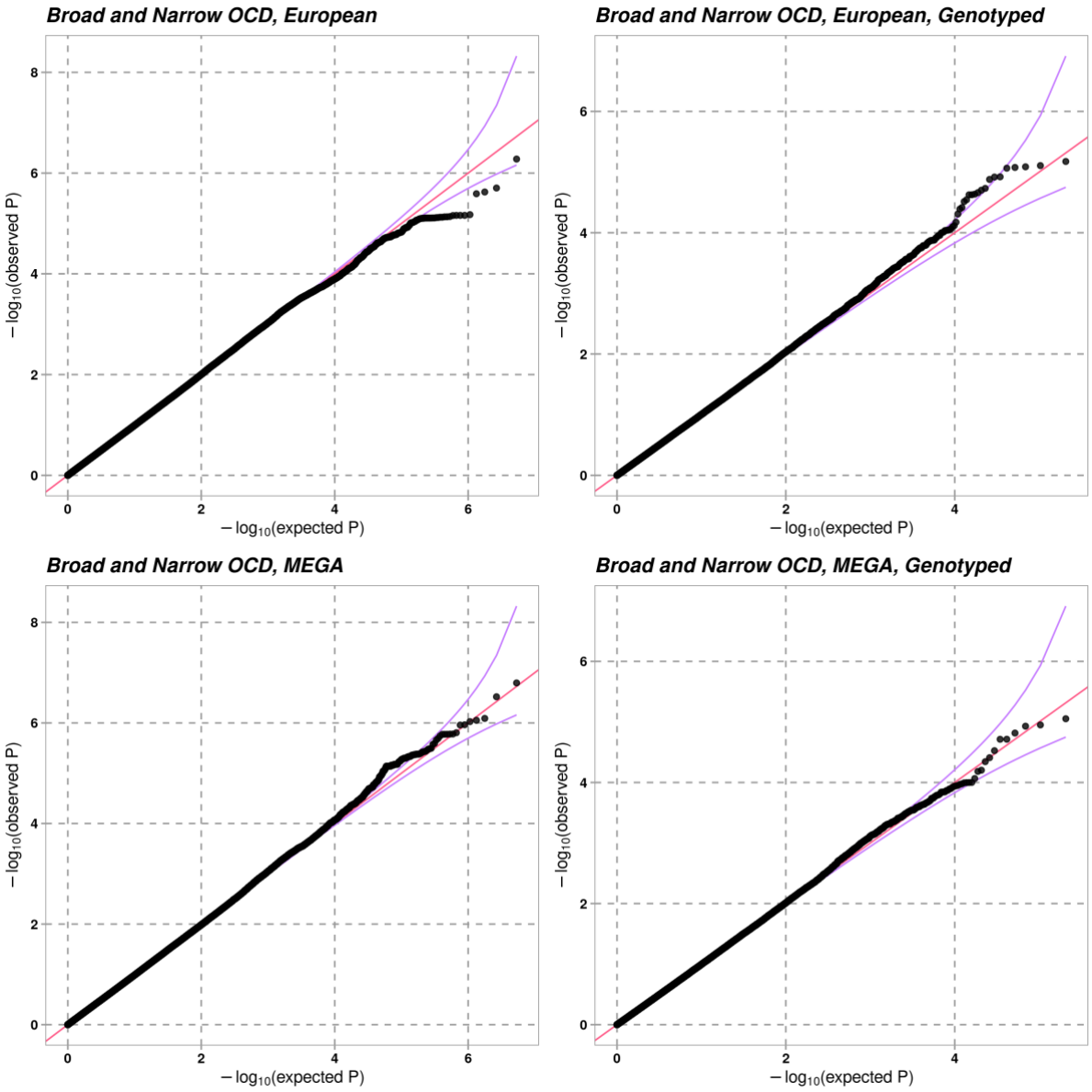

Figure 5-17. Continued.

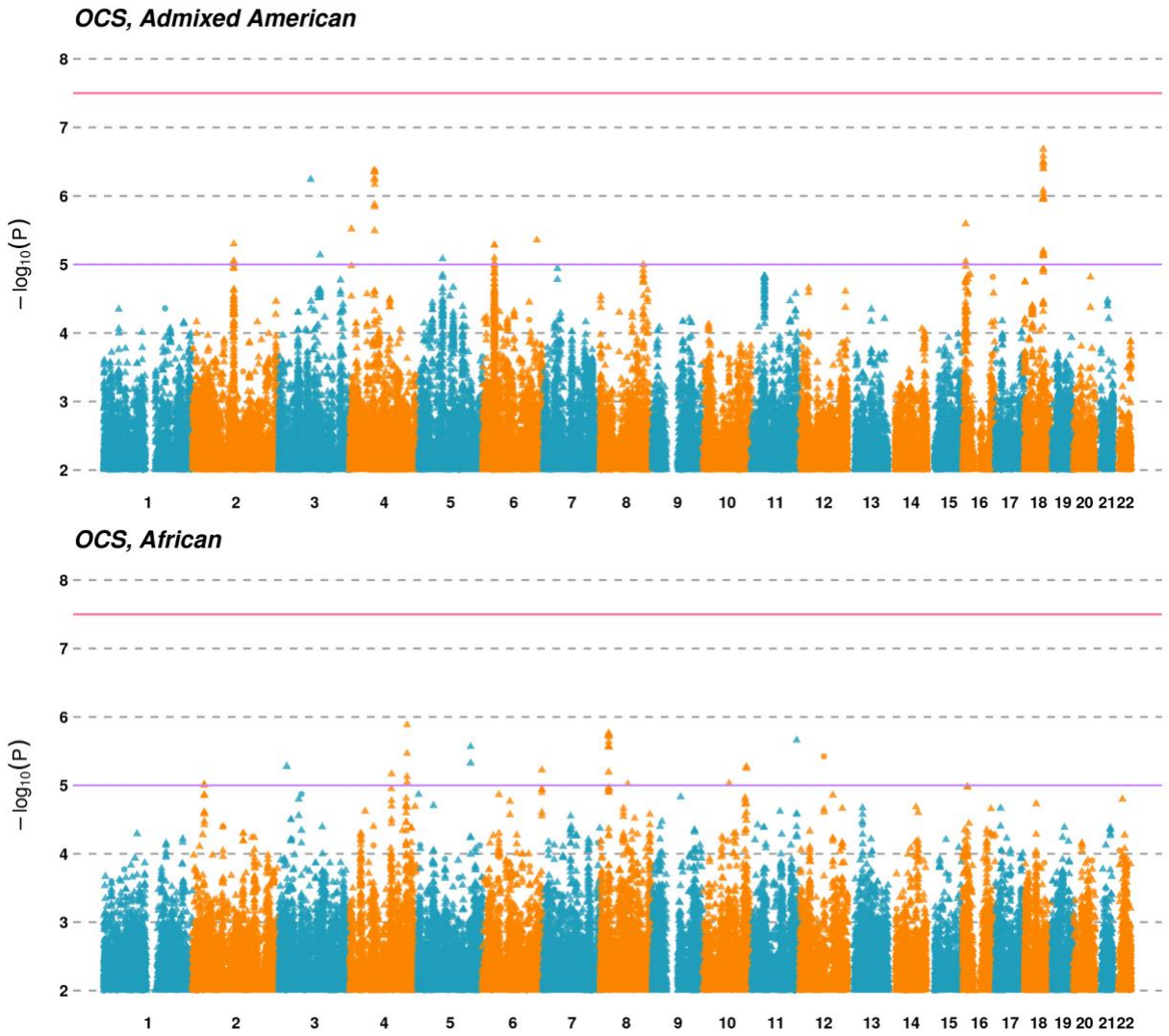

Figure 5-18. OCS Manhattan plots.

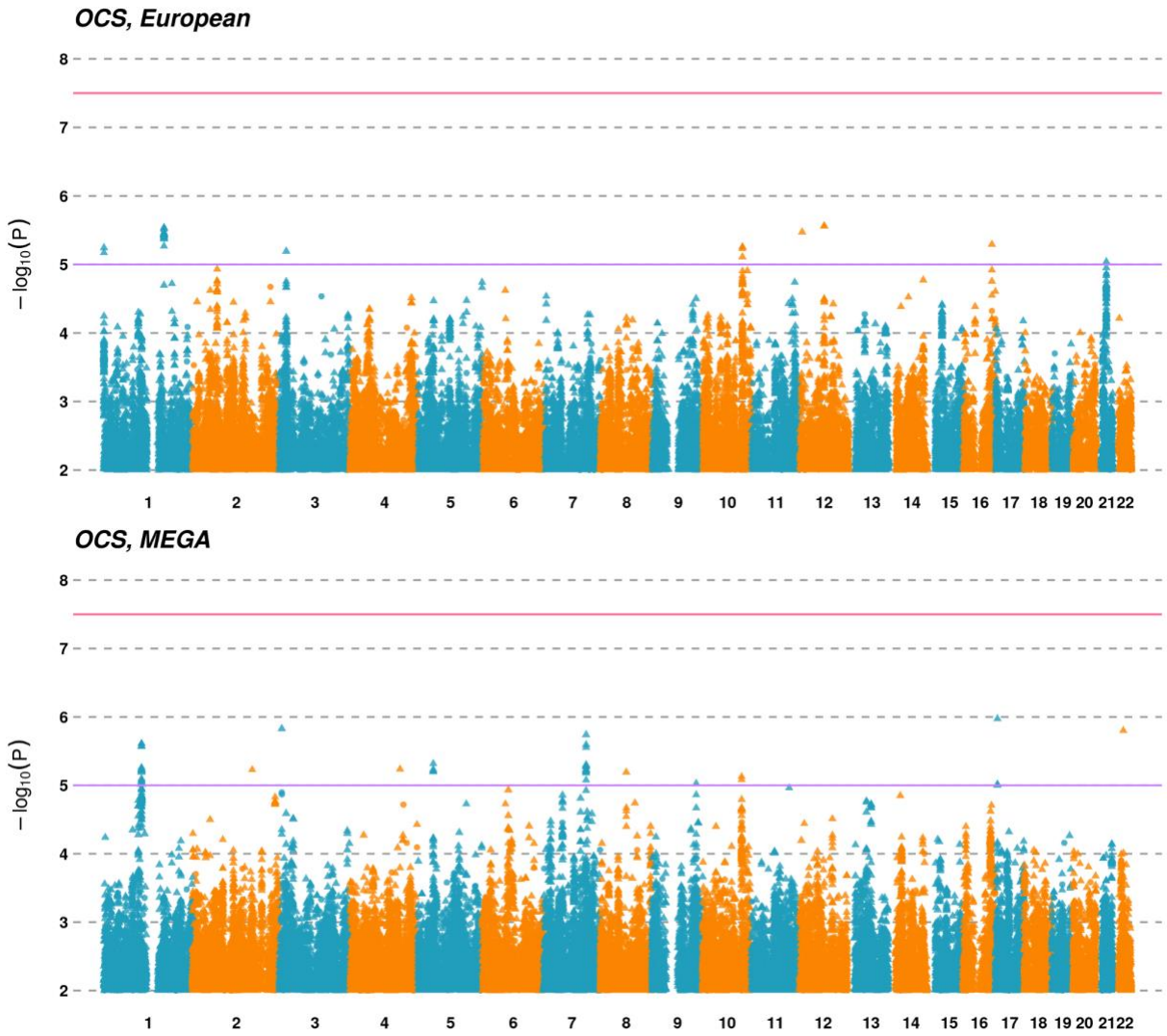

Figure 5-18. Continued.

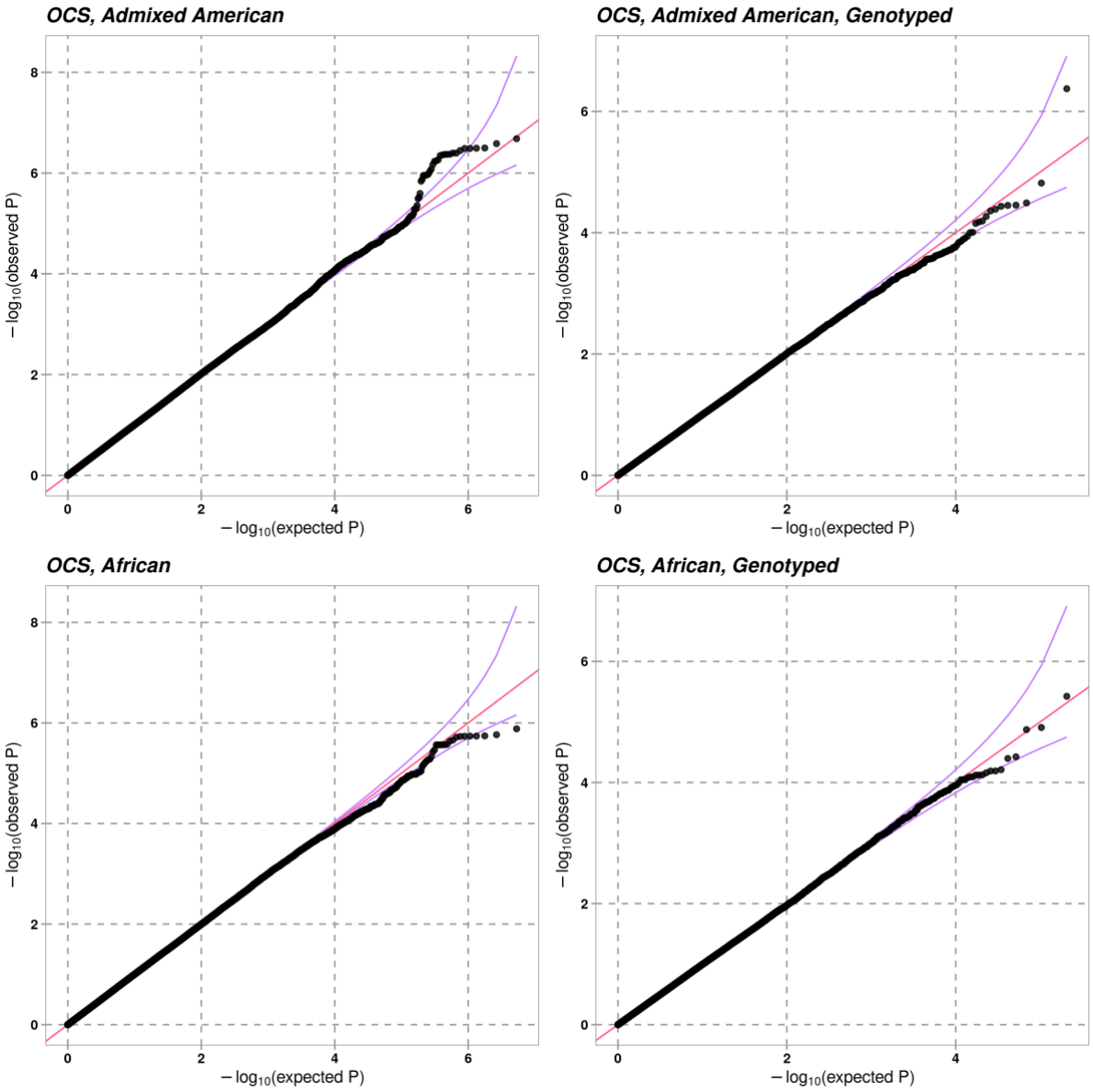

Figure 5-19. OCS GWAS QQ plots.

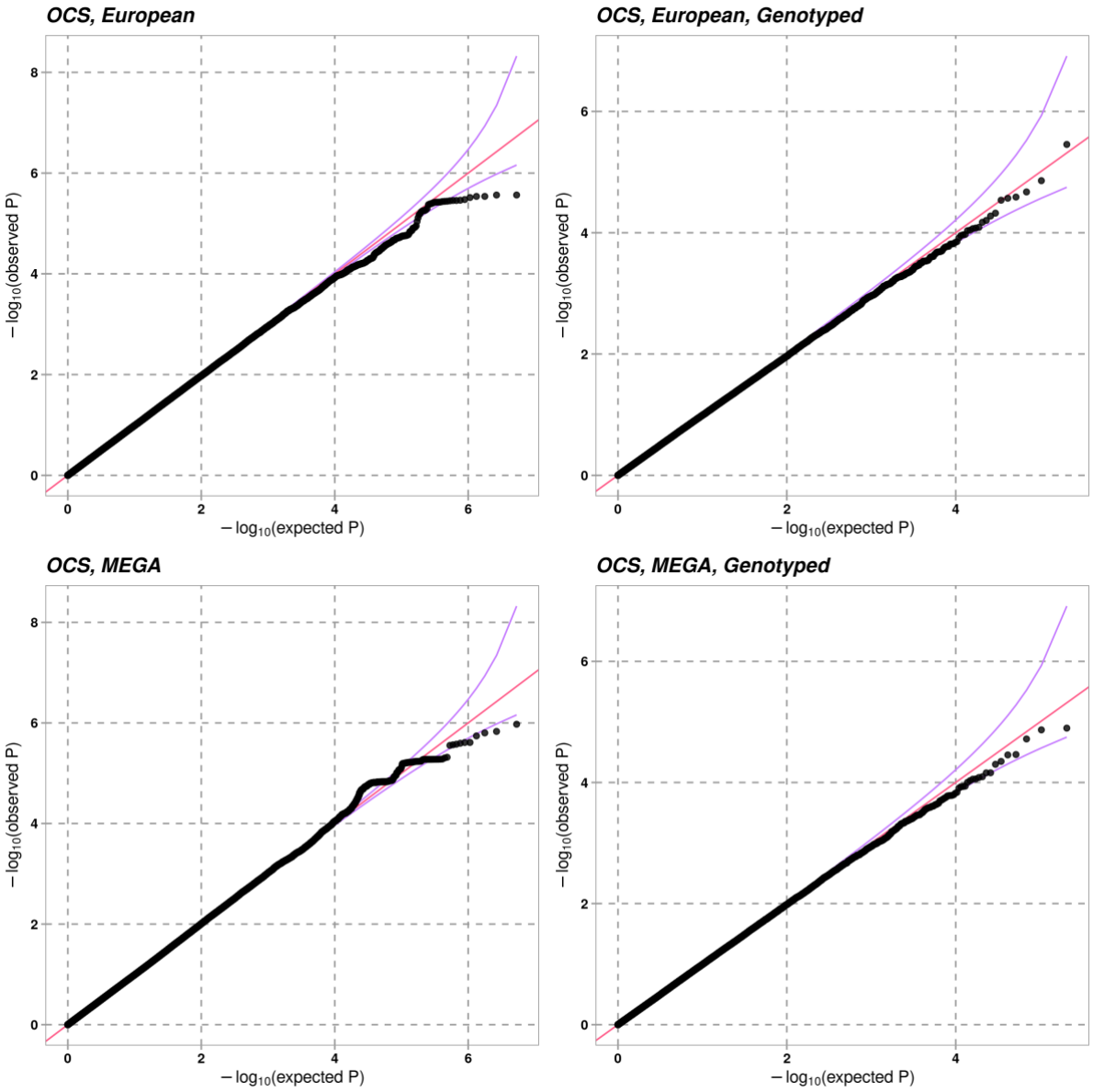

Figure 5-19. Continued.

Table 5-4. Summary of GWAS top hit loci, $p < 10^{-5}$.

Chr	Mb	nOCD				bnOCD				OCS			
		AM	AF	EU	ME	AM	AF	EU	ME	AM	AF	EU	ME
1	1	.	.	.	.	.	.	.	.	.	.	G	.
	4	.	.	.	.	.	.	I	.	.	.	.	.
	41	.	.	.	.	.	.	.	G	.	.	.	.
	58	.	.	.	.	G	.	.	.	.	.	.	.
	64	I	.	.	.	.	.	.	.	.	.	.	.
	106	.	.	.	.	.	.	.	.	.	.	.	I
	115	.	G	.	.	.	.	.	.	.	.	.	.
	156	.	.	.	.	.	.	.	G	.	.	.	.
	168	.	.	.	.	.	.	.	.	.	.	I	.
	246	.	.	G	.	.	.	.	.	.	.	.	.
2	30	.	.	.	I	.	.	.	.	.	.	.	.
	33	.	.	.	.	.	.	.	.	.	G	.	.
	45	.	.	.	I	.	.	.	.	.	.	.	.
	65	I	.	.	.	.	.	.	.	.	.	.	.
	115	.	.	.	.	.	.	.	.	G	.	.	.
	128	.	.	.	G	.	.	.	.	.	.	.	.
	166	.	.	.	.	.	.	.	.	.	.	.	G
	193	.	.	.	.	.	G	.	.	.	.	.	.
	200	.	.	.	.	.	.	G	.	.	.	.	.
	220	I	.	.	.	.	.	.	.	.	.	.	.
	231	.	.	.	.	.	.	.	I	.	.	.	.
	235	.	.	.	.	.	.	I	.	.	.	.	.
3	3	.	.	.	.	.	I	.	.	.	.	.	.
	6	.	.	.	.	.	.	.	.	.	.	.	I
	18	.	I	.	.	.	.	.	.	.	.	G	.
	21	.	.	.	.	.	.	.	.	.	I	.	.
	34	.	I	.	.	.	.	.	.	.	.	.	.
	37	.	.	.	.	.	G	.	.	.	.	.	.
	45	.	.	.	.	.	G	.	.	.	.	.	.
	70	.	.	.	.	I	.	.	.	.	.	.	.
	87	.	I	.	.	.	.	.	.	I	.	.	.
	114	.	.	.	.	.	.	G	.	G	.	.	.
	145	.	.	.	.	.	I	.	.	.	.	.	.
	161	I	.	.	.	.	.	.	.	.	.	.	.
4	3	.	.	.	.	.	.	.	.	I	.	.	.
	43	.	.	.	.	.	I	.	.	.	.	.	.

Note: Bolded italicized rows indicate a locus that associated with a phenotype with $p < 10^{-5}$ in more than one GWAS. AM = AMR, AF = AFR, EU = EUR, ME = MEGA.

Table 5-4. Continued.

Chr	Mb	nOCD				bnOCD				OCS			
		AM	AF	EU	ME	AM	AF	EU	ME	AM	AF	EU	ME
4	68	.	.	.	.	.	.	.	.	G	.	.	.
	86	.	.	.	.	G	.	.	.	.	.	.	.
	139	.	.	.	.	.	.	.	.	.	.	.	I
	158	.	.	.	.	.	I	.	.	.	.	.	.
	160	.	.	.	.	.	.	.	.	.	G	.	.
	180	I	I	.	.	.	.	.	.	.	.	.	.
5	3	.	.	.	I	.	.	.	.	.	.	.	.
	20	.	.	.	G	.	.	.	.	.	.	.	.
	23	.	.	.	.	.	I	.	.	.	.	.	.
	28	.	I	.	.	.	.	.	.	.	.	.	.
	36	.	.	I	.	.	.	.	.	.	.	.	.
	42	.	.	.	.	.	.	.	.	.	.	.	I
	43	I	.	.	.	.	.	.	.	.	.	.	.
	55	.	.	I	.	.	.	.	.	.	.	.	.
	66	.	.	.	G	.	.	.	.	.	.	.	.
	67	.	.	.	.	.	.	I	.	I	.	.	.
	119	.	.	.	.	.	I	.	.	.	.	.	.
	147	.	.	.	.	.	.	.	.	.	I	.	.
6	2	.	.	.	.	.	.	.	I	.	.	.	.
	20	.	.	.	.	.	.	G	G	.	.	.	.
	32	.	.	.	.	.	.	.	.	G	.	.	.
	33	.	.	I	.	.	.	.	.	.	.	.	.
	35	.	G	.	.	.	.	.	.	.	.	.	.
	108	.	.	.	.	I	.	.	.	.	.	.	.
	129	.	.	.	.	G	.	.	.	.	.	.	.
	151	.	.	.	.	.	.	.	.	G	.	.	.
	166	.	.	.	.	.	.	.	.	.	G	.	.
7	0	.	G	.	.	.	.	.	.	.	.	.	.
	9	.	.	.	.	.	.	I	.	.	.	.	.
	21	.	.	.	I	.	.	.	.	.	.	.	.
	30	I	.	.	.	.	.	.	.	.	.	.	.
	35	.	.	.	.	.	I	.	.	.	.	.	.
	77	G	.	.	.	.	.	.	.	.	.	.	.
	88	.	.	.	.	G	.	.	.	.	.	.	.
	110	.	.	.	G	.	.	.	.	.	.	.	.
	118	.	.	.	.	.	.	.	.	.	.	.	I

Note: Bolded italicized rows indicate a locus that associated with a phenotype with $p < 10^{-5}$ in more than one GWAS. AM = AMR, AF = AFR, EU = EUR, ME = MEGA.

Table 5-4. Continued.

Chr	Mb	nOCD				bnOCD				OCS			
		AM	AF	EU	ME	AM	AF	EU	ME	AM	AF	EU	ME
7	153	.	.	.	G	.	.	.	.	.	.	.	.
8	13	.	G	.	.	.	.	.	.	.	.	.	.
	23	.	.	.	.	.	.	.	.	.	I	.	.
	72	.	.	.	.	.	.	.	.	.	.	.	G
	77	.	.	.	.	.	.	.	.	.	I	.	.
	108	I	.	.	.	.	.	.	.	.	.	.	.
	139	.	.	.	.	.	I	.	.	.	.	.	.
9	12	I	.	.	.	.	.	.	.	.	.	.	.
	13	.	.	.	.	I	.	.	.	.	.	.	.
	88	.	.	.	.	.	I	.	.	.	.	.	.
	95	G	.	.	.	.	.	.	.	.	.	.	.
	112	.	.	.	.	.	G	.	.	.	.	.	.
	122	.	.	.	.	.	.	.	.	.	.	.	I
	130	.	.	.	.	.	.	.	G	.	.	.	.
10	72	.	.	.	.	.	.	.	.	.	I	.	.
	108	.	.	.	.	.	.	.	.	.	.	G	G
	121	.	.	.	.	.	.	.	.	.	G	.	.
11	120	.	.	.	G	.	.	.	.	.	.	.	.
	123	.	.	.	.	G	.	.	.	.	.	.	.
	126	.	.	.	.	.	.	.	.	.	G	.	.
	134	.	.	.	.	G	.	.	.	.	.	.	.
12	0	.	.	.	.	.	G	.	.	.	.	.	.
	4	.	.	.	.	G	.	.	.	.	.	.	.
	5	.	.	.	.	.	.	.	.	.	.	G	.
	42	.	.	.	G	.	.	.	.	.	.	.	.
	57	.	G	.	.	.	.	.	.	.	.	.	.
	67	.	.	.	.	.	.	.	.	.	I	G	.
13	23	.	.	.	.	G	.	.	.	.	.	.	.
	73	.	.	I	.	.	.	.	.	.	.	.	.
	113	.	.	.	.	.	.	.	I	.	.	.	.
14	21	.	.	.	.	.	G	.	.	.	.	.	.
	24	.	.	.	.	.	.	.	G	.	.	.	.
	56	.	.	.	G	.	.	.	.	.	.	.	.
	70	.	.	.	.	.	G	.	.	.	.	.	.
	76	.	.	.	.	.	.	.	I	.	.	.	.
15	33	.	.	G	.	.	.	.	.	.	.	.	.

Note: Bolded italicized rows indicate a locus that associated with a phenotype with $p < 10^{-5}$ in more than one GWAS. AM = AMR, AF = AFR, EU = EUR, ME = MEGA.

Table 5-4. Continued.

Chr	Mb	nOCD				bnOCD				OCS			
		AM	AF	EU	ME	AM	AF	EU	ME	AM	AF	EU	ME
	94	I	.	.	.	.	.	.	.	.	.	.	.
	98	.	.	.	.	.	.	.	I	.	.	.	.
16	5	.	G	.	.	.	G	.	.	G	.	.	.
	8	I	.	.	.	.	.	.	.	.	.	.	.
	20	.	.	.	.	.	.	.	.	G	.	.	.
	21	.	.	.	.	.	G	.	.	.	.	.	.
	24	.	.	.	.	.	.	.	I	.	.	.	.
	60	.	I	.	.	.	.	.	.	.	.	.	.
	66	.	.	G	.	.	.	.	.	.	.	.	.
	78	.	.	.	.	.	.	.	.	.	.	G	.
17	5	.	.	.	.	.	.	.	.	.	.	.	I
	16	.	.	G	G	.	.	.	.	.	.	.	.
18	40	.	.	.	.	.	I	.	.	.	.	.	.
	45	.	.	.	.	.	G	.	.	.	.	.	.
	43	.	.	.	.	I	.	.	.	.	.	.	.
	52	.	.	.	.	.	.	.	.	G	.	.	.
	57	.	.	.	I	.	.	.	.	.	.	.	.
19	6	.	.	.	.	.	I	.	.	.	.	.	.
	10	.	.	.	.	I	.	.	.	.	.	.	.
	15	.	.	G	.	.	.	.	.	.	.	.	.
20	38	.	.	.	.	.	.	I	.	.	.	.	.
	50	.	.	.	.	.	G	.	.	.	.	.	.
21	29	.	.	.	.	.	.	.	.	.	.	I	.
22	22	.	.	.	.	.	I	.	.	.	.	.	.
	29	.	.	.	.	.	.	.	.	.	.	.	G
	38	.	.	.	.	.	.	.	G	.	.	.	.
	49	.	.	I	I	.	.	.	.	.	.	.	.

Note: Bolded italicized rows indicate a locus that associated with a phenotype with $p < 10^{-5}$ in more than one GWAS. AM = AMR, AF = AFR, EU = EUR, ME = MEGA.

Figure 5-20. Ancestry overlapping loci with $p < 10^{-5}$.

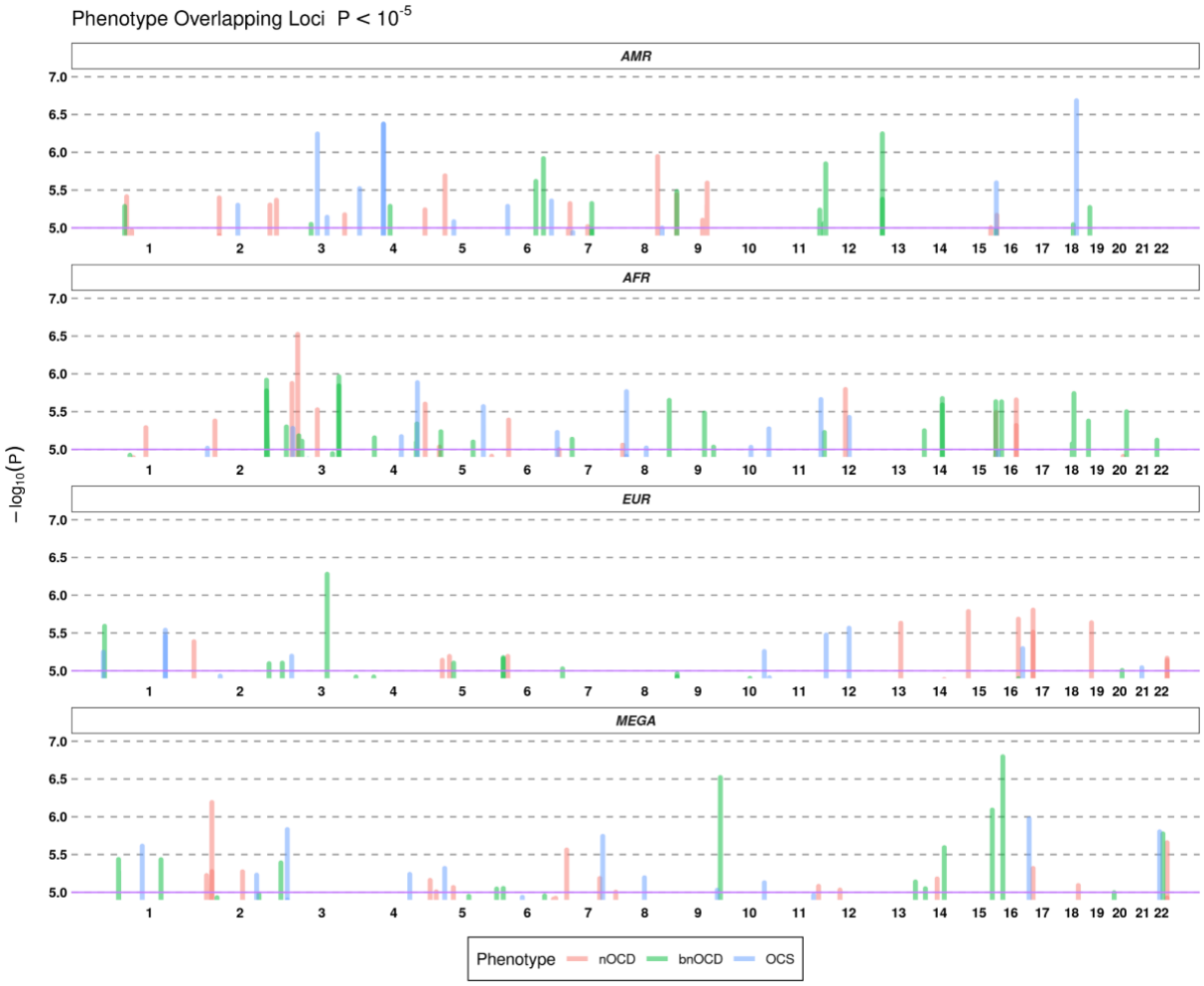

Figure 5-21. Phenotype overlapping loci with $p < 10^{-5}$.

Table 5-5. GWAS GO Analysis summary.

GO Term	nOCD				bnOCD				OCS			
	AM	AF	EU	ME	AM	AF	EU	ME	AM	AF	EU	ME
Neuron to neuron synapse	10^{-2}	-	-	-	-	-	-	-	-	-	-	-
Integral component of lumenal side of endoplasmic reticulum membrane	-	-	-	-	-	-	-	-	10^{-4}	-	-	-
ER to Golgi transport vesicle membrane	-	-	-	-	-	-	-	-	10^{-4}	-	-	-
Clathrin-coated endocytic vesicle membrane	-	-	-	-	-	-	-	-	10^{-4}	-	-	-
Trans-Golgi network membrane	-	-	-	-	-	-	-	-	10^{-3}	-	-	-
Lysosomal membrane	-	-	-	-	-	-	-	-	10^{-2}	-	-	-

Note: AM = AMR, AF = AFR, EU = EUR, ME = MEGA.

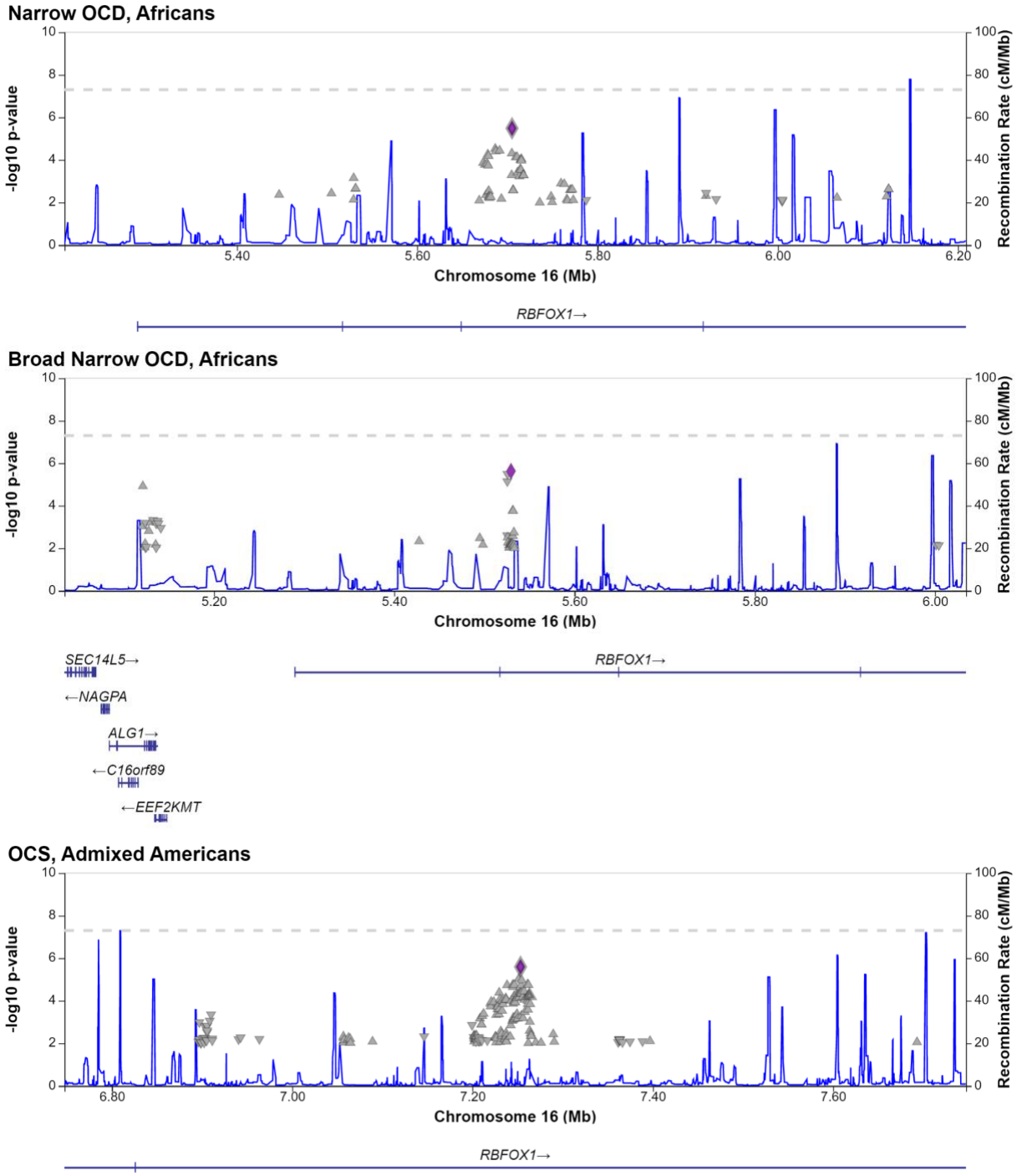

Figure 5-22. LocusZoom plot with $p < 10^{-5}$ loci overlapping *RBFOX1* gene.

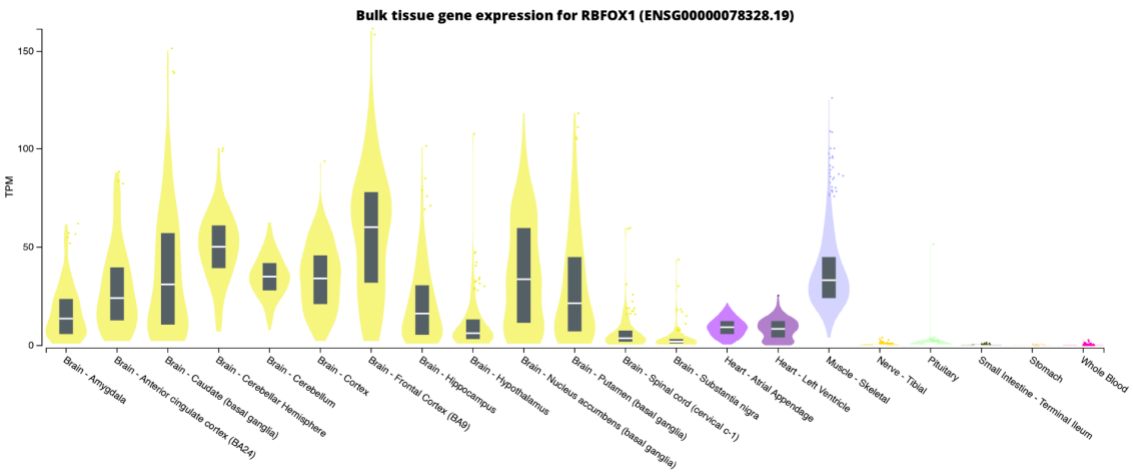

Figure 5-23. GTEx plot for RBFOX1.

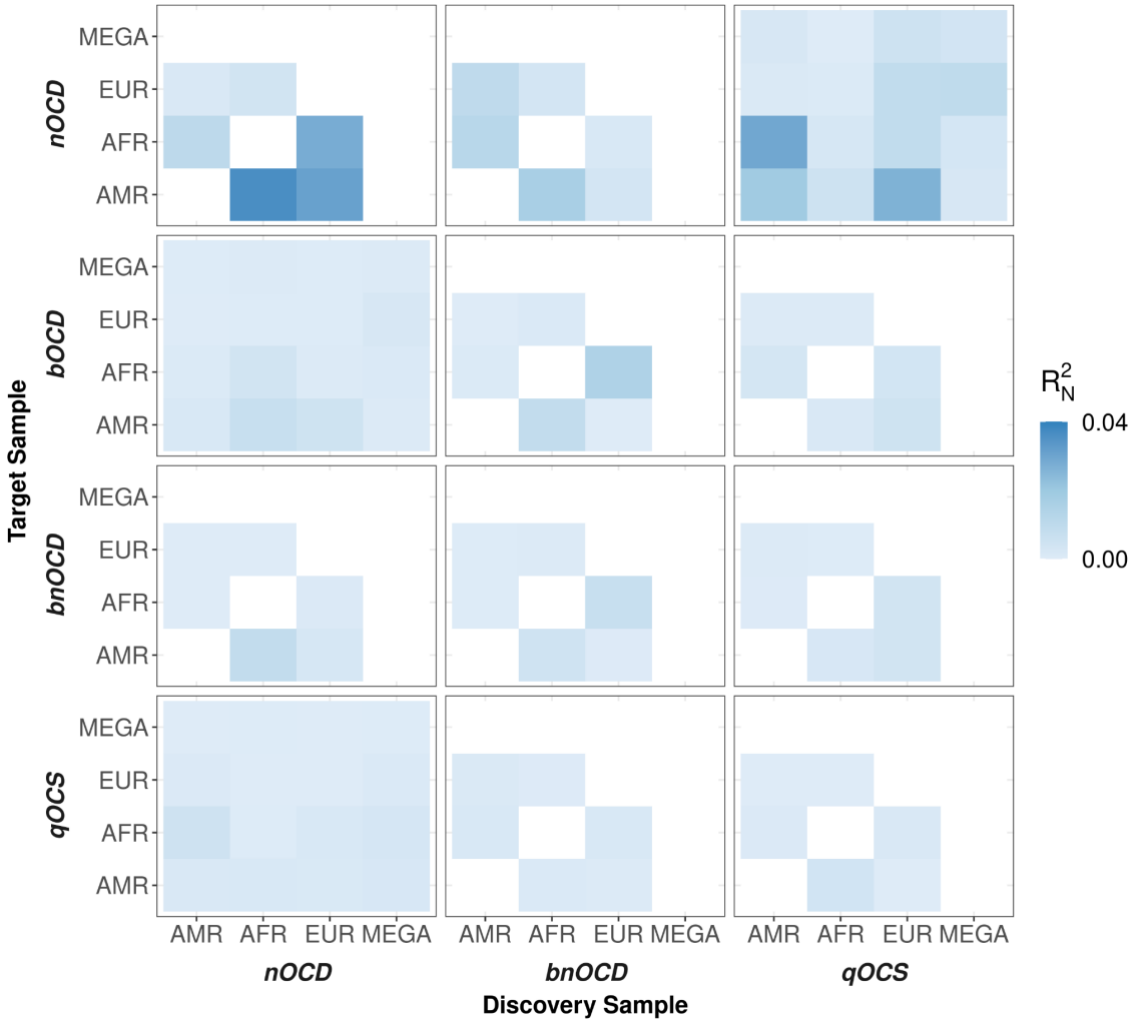

Figure 5-24. ABCD PRS analysis summary heatmap of R_N^2 statistics.

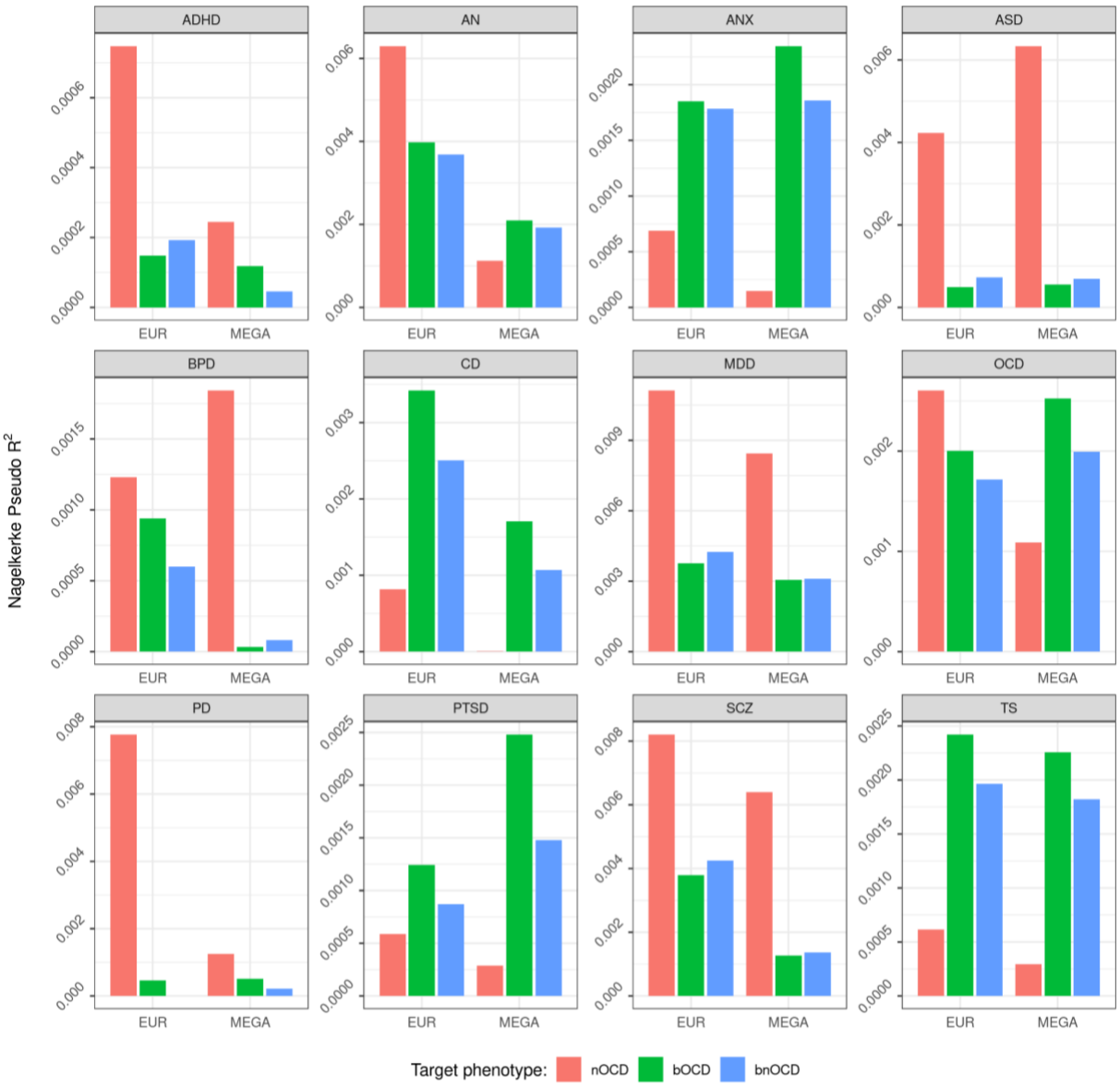

Figure 5-25. PGC PRS-PCA analysis on full target samples. CD = cross-disorder / psychopathology meta-GWAS.

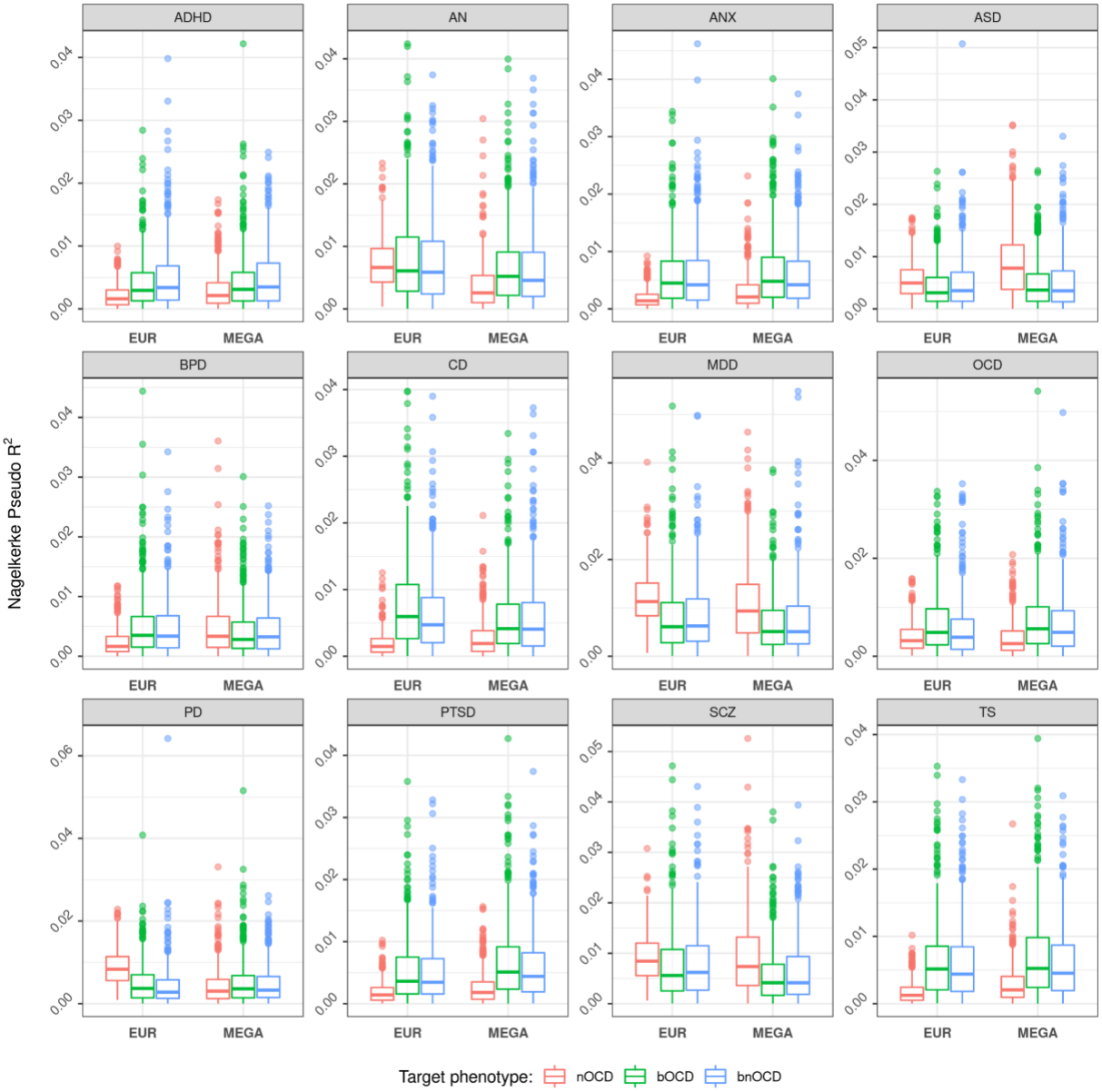

Figure 5-26. PGC PRS-PCA analysis by repeated undersampling. CD = cross-disorder / psychopathology meta-GWAS.

Table 5-6. REML analysis of ABCD sample.

Cohort	Ancestry	h^2_{OCD}	SE_{OCD}	h^2_{OCS}	SE_{OCS}	r_g	SE_g
bnOCD	EUR	0.00000	0.06090	0.00153	0.04872	-	-
bnOCD	MEGA	0.00000	0.04145	0.00000	0.03579	-	-

Note: Trait correlations were not computable due to sample size limitations.

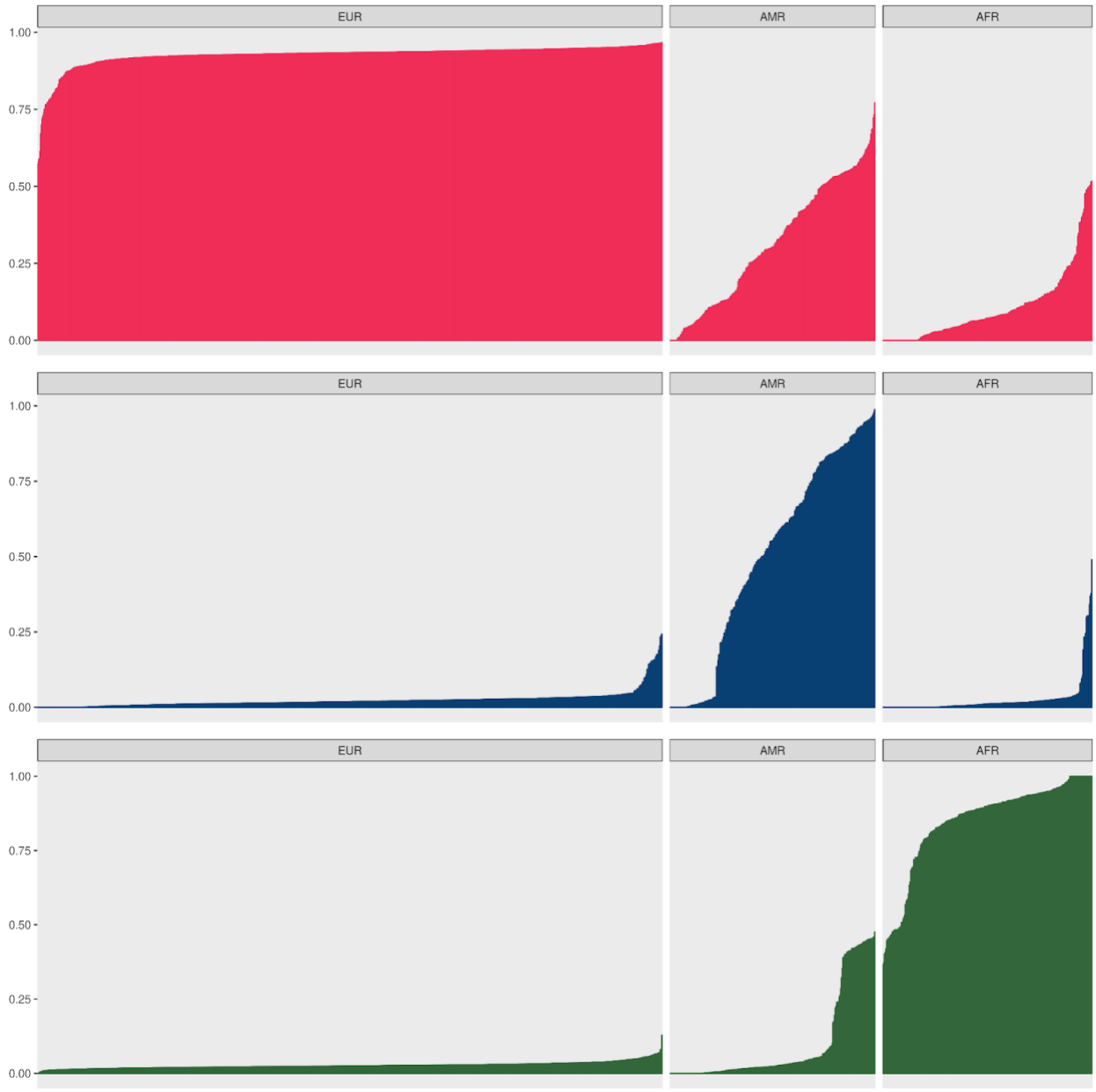

Figure 5-27. Admixture analysis of ABCD Study.

CHAPTER 6
CNV ANALYSIS OF TOURETTE SYNDROME FAMILIES

Background

Shortly after the advent and wide-spread utilization of GWAS, hundreds of genetic variants underlying complex diseases and traits were identified, yet they usually conferred small magnitudes of risk and explained a fraction of estimated genetic variance from the family studies (Zuk et al., 2012). Possible sources of missing genetic variation underlying this missing heritability problem could be due to large-effect rare variants that can only be detected by sequencing or structural variants which affect dosage of a given allele but not its sequence composition (Manolio et al., 2009). One such structural variant type includes CNVs. CNVs form faster than any other mutation in the human genome, likely due to its ability to escape repair mechanisms, and usually arise through non-homologous end joining, replication, and replication of non-contiguous DNA segments – usually mediated by cellular stress which can cause repair of broken replication fork switch from high-fidelity replication mechanisms to low-fidelity ones (Hastings et al., 2009). Indeed, CNVs are found to make up 4.8-9.5% of the human genome, and can be very polymorphic (Zarrei et al., 2015). However, the rare and large CNVs have been found to play an important role human disorders, especially those with severe psychiatric phenotypes like schizophrenia, ASD, and intellectual disability (Shaikh, 2017). In addition to being associated with disorders at large, CNVs were also found to underlie impairments in cognitive domains like memory and perceptual reasoning (Thygesen et al., 2021).

Detection of CNVs has its fair share of challenges – namely, existing microarray and sequencing technologies were developed with a goal of identifying sequence of

nucleotides in the human genome in a binary-like present/absent form, yet CNV detection relies on complex statistical and computational modelling of unbinarized signal measurements to determine dosage of a given nucleotide or segments thereof as it relates to the rest of the sample (Carter, 2007). Most reliable methods for detecting CNVs involve low-throughput qPCR, molecular copy counting, and paralog ratio testing (Li & Olivier, 2013). However, these methods, despite being high-resolution, sensitive, and specific, are not suitable for genome-wide high-throughput experiments required to detect weaker effects CNVs like those presumed to underlie complex genetic disorders.

Thus, numerous approaches have been developed to detect CNVs from both sequencing and microarray data, namely relying on HMM, expectation-maximization (EM) clustering, REML, various forms of regressions (under Gaussian, Poisson, and negative binomial assumptions), least absolute shrinkage and selection operator (LASSO) regressions, and machine learning (ML) approaches like deep learning and neural networks, and random forest algorithms (Li & Olivier, 2013; Macé et al., 2016; Hill & Unckless, 2019; Pounraja et al., 2019; Zhang et al., 2019; Zhuang et al., 2020; Glessner et al., 2021). For this project, the primary calling algorithm used in PennCNV, an integrated HMM-based approach to detecting CNVs in family samples from microarray data (Wang et al., 2007; Diskin et al., 2008; Wang et al., 2008). To increase validity and reliability of detected CNVs, a secondary caller QuantiSNP, can be used to refine the CNV call-set (Colella et al., 2007).

As discussed in Chapter 2, there are several studies that probed the role of CNVs in TS (Huang et al., 2017; Wang et al., 2018). In this chapter, I introduce a continuation of these efforts by analyzing impact of CNVs on OCRDs, specifically TS in

the largest ever family sample, focusing on both rare and common variation, as well as inherited and *de novo* CNV mutations.

Methods

Samples

TD

Sample and phenotyping. Sample with families of probands ascertained for TDs was collected and provided by the Tourette Association of America International Consortium for Genetics (TAAICG, 2007). TAAICG cohort consists of TD (primarily TS) probands collaboratively recruited by TS investigators from numerous sites throughout the North America, Europe, and Israel. Over 1,200 families were used as a starting sample. Diagnosis level information was available for TS, PMVT, TTD, OCD, and ADHD. Biosamples used for genetic experiments were obtained from either whole blood, saliva, and peripheral blood mononuclear cell lines. Individuals were diagnosed using a standardized and validated semi-structured direct interview (TAAICG tic and comorbid symptom inventory) and recruited by the clinicians at the respective sites, TS specialty clinics, and online (Darrow et al., 2015). Phenotypes were verified by multi-investigator best-estimate analysis (Leckman et al., 1982). Raw genotype data in form of .idat files were provided for processing and analysis.

Genotyping. The TAAICG samples were genotyped on a Infinium Global Screening Array-24 BeadChip, versions 1 (GSAv1) and 3 (GSAv3), across 4 waves of genotyping between 2020 and 2021. A total of 4,473 samples were genotyped (with a few samples being duplicates) in following waves:

1. GSAv1 genotyping of 2,349 samples (released on 01/08/2020),
2. GSAv1 genotyping of 867 samples (released on 02/11/2020),
3. GSAv1 genotyping of 1,174 samples (released on 08/18/2021),

4. GSAv3 genotyping of 83 samples (released on 08/23/2021).

Both GSAv1 and GSAv3 feature backbone panel of SNPs specifically created to combine multi-ethnic genome-wide content with curated clinical research variants markers for precision medicine research, and multi-disease drop-in panel of SNPs consisting of fine-mapping content derived from exome sequencing and meta-analyses of phenotype-specific consortia focused on the psychiatric, neurological, cancer, cardiometabolic, autoimmune, and anthropometric traits (Illumina, 2017; Illumina 2020b). GSAv1 features 642,824 backbone SNPs and 57,254 SNPs from the multi-disease drop-in panel. GSAv3 features 654,027 backbone SNPs and 76,039 SNPs from the multi-disease drop-in panel. However, Illumina has updated manifest files for GSAv1 since its release, to now feature 618,540 backbone SNPs only. For the CNV analysis purposes, only the backbone SNPs are considered as multi-disease drop-in panel varies across samples and is not available for the comparison sample. Raw .idat files were provided, together with .bpm SNP manifest file, and .csv sample sheet, importable and processable by Illumina's GenomeStudio software (Illumina, 2020a). Data were accessed through Terra (Terra, n.d.).

ASD and unaffected siblings

Sample and phenotyping. For comparison groups, I obtained family data composed of ASD probands and unaffected siblings recruited by the Simons Foundation Autism Research Initiative (SFARI), specifically from their Simons Foundation Powering Autism Research and Knowledge (SPARK) study (SPARK Consortium, 2018). As of February 2021, SPARK study enrolled 251,082 individuals, 91,477 of which have ASD diagnosis. Phenotyping for primary disorder of interest, ASD

is based on individual or parent (in case of underage participants) self-reported previous ASD diagnosis by a trained professional (physician, psychologist, or therapist). The unaffected siblings are, in fact, unscreened – therefore, some missed ASD cases are possible. Other phenotyping data includes background history and medical screen, CBCL, and several other inventories probing ASD-specific symptomatology that is not of immediate interest to this project.

Genotype data. As of February 2021, 27,615 participants had their genotyping data released. All genotyping was done on GSAv1, described in the previous section. Batch information was not available for these samples. Data were accessed through GlobusOnline (Foster, 2011; Allen et al., 2012).

Data Processing

Part of the QC for this dataset was previously conducted by a member of our team for the purposes of GWAS on TAAICG trios, Dongmei Yu. The pre-processing step included GWAS QC approaches as described in Chapter 5. Briefly, the following exclusion criteria were used to clean the samples: $GR_{SNP} < 0.98$; $GR_S < 0.98$; $P_{HWE} < 10^{-6}$ (among unaffected individuals); MAF < 0.01; $MISS_{Diff} > 0.02$; $MER_{SNP} > 0.05$; $MER_S > 0.03$; $|F_{HET}| > 0.2$, where $MISS_{Diff}$ represents differential missingness, and MER_{SNP} and MER_S represent SNP-wise and sample-wise Mendelian error rate. Additional steps included removal of incomplete trios, and cross-contaminated and unexpectedly related samples as determined by IBD. Remaining analyses in this chapter were all performed by me.

TAAICG

Input of 4,472 individuals was first filtered to remove incomplete trios, resulting in retention of 3,967 individuals. Wave 4 containing 83 individuals had an overall poor

genotyping performance and was genotyped on a completely different platform from the rest of the data, so it was excluded from the analysis. The remaining individuals were grouped in father-mother-offspring groups for trio calling, resulting in 1,376 unique trios from 3,815 individuals from 1,226 families. Summary is shown in Table 6-1.

SPARK

SPARK data input of 27,281 individuals was first filtered to remove all families where both parents and two or more children were not present, resulting in a sample of 14,045 individuals. Subsequently, families with missing ASD phenotypes were excluded, resulting in 13,857 retained participants. Next, 28 half siblings were also removed from the analysis. Samples without associated .idat files were removed, bringing sample size down to 13,697. Only complete families were kept for the analysis, resulting in removal of 176 participants. Finally, only those samples which had both ASD-affected and unaffected children were retained, bringing the final sample size down to 11,438. After closer inspection, some mismatching with available genotype files was present. After removing the mismatched individuals and affected families, 11,160 individuals were retained for the analysis. Samples were further stratified by proband status. In the ASD proband subsample, there were 2,893 unique trios from 8,341 individuals from 2,724 families. In the unaffected sibling subsample, there were 2,785 unique trios from 8,233 individuals from 2,724 families. Summary is shown in Table 6-1, TAAICG refers to the sample with TS probands, SPARK-ASD refers to the sample with ASD probands, and SPARK- SIB refers to the sample with unaffected sibling probands. Majority of the analyses will focus on differences between these three groups or batches of TAAICG.

Genetic Report formation

Based on preprocessed data, samples were imported into Illumina GenomeStudio and clustering was performed de novo (without utilization of canonical cluster files) with data-attached SNP manifest files. If batch information was available, clustering was done per-batch. After clustering was complete, non-autosomal SNPs, QC SNPs, and SNPs from the drop-in panel were removed from the analysis. For TAAICG data, the number of remaining SNPs was 600,679. For SPARK data, the number of remaining SNPs was 600,899. Genome-wide summaries of copy-number metrics relevant for this study are shown in Table 6-2. Listed are Med(LRR_{SD}), SD(LRR_{SD}), Med(BAF_{SD}), and SD(BAF_{SD}), where Med and SD stand for median and standard deviation, LRR_{SD} stands for sample-wise standard deviation of the log R ratio, and BAF_{SD} stands for sample-wise standard deviation of the B allele frequency.

Log R ratio (LRR) is a normalized signal intensity for each SNP on the microarray. For each SNP, two-color readout results in intensity values in two channels (where each color represents one of two alleles for a given SNP). These intensity values are polar transformed to obtain normalized intensity values (R) and allelic intensity ratios (θ).

LRR for i^{th} SNP and j^{th} individual is then calculated according to the Equation 6-1. B allele frequency (BAF) indicates relative quantity of one allele compared to another at a given locus, where B allele indicates non-reference allele, note these are not equivalent to MAF, as MAF is the frequency of a less prevalent allele in a given population. Mathematical formula for calculating BAF for i^{th} SNP and j^{th} individual is shown in Equation 6-2. Note that $R_{O,\,i,\,j}$ and $\theta_{O,\,i,\,j}$ are metrics observed for i^{th} SNP and j^{th} individual, whereas $R_{E,\,i}$, $\theta_{AA,\,i}$, $\theta_{AB,\,i}$, and $\theta_{BB,\,i}$ are metrics for i^{th} SNP either derived from

whole sample data or retrieved from canonical sample files. Important consideration for CNV analysis is that each SNP's metrics are affected by individual's data on that position, but not other SNPs in the dataset.

$$LRR_{i,j} = \log_2\left(\frac{R_{O,i,j}}{R_{E,i}}\right) \tag{6-1}$$

$$BAF_{i,j} = \begin{cases} 0 \ , & \theta_{O,i,j} < \theta_{AA,i} \\ \dfrac{\theta_{O,i,j} - \theta_{AA,i}}{2\left(\theta_{AB,i} - \theta_{AA,i}\right)} \ , & \theta_{AA,i} \le \theta_{O,i,j} < \theta_{AB,i} \\ \dfrac{1}{2} - \dfrac{\theta_{O,i,j} - \theta_{AB,i}}{2\left(\theta_{BB,i} - \theta_{AB,i}\right)} \ , & \theta_{AB,i} \le \theta_{O,i,j} < \theta_{BB,i} \\ 1 \ , & \theta_{O,i,j} \ge \theta_{BB,i} \end{cases} \tag{6-2}$$

No special QC was performed at this step, instead Illumina final report files were generated by exporting all autosomal backbone SNP values for all individuals for chromosome and position, nucleotide value for both alleles (A/T/G/C), $LRR_{i,j}$, and $BAF_{i,j}$. Final report files were then uploaded on HiPerGator for further analysis.

PennCNV Calling

PennCNV algorithm jointly calling CNVs for each set of trios was used as a primary algorithm in this study (Wang et al., 2007; Diskin et al., 2008; Wang et al., 2008). First, samples were unpacked from final report files using the split_illumina_report.pl script from the PennCNV package, version 1.0.5. Subsequently, the GC model file was generated using hg19 reference genome and cal_gc_snp.pl script. The GC model file summarizes percentage of G or C base pairs in a 500kb region flanking each SNP from the dataset, this information will then be used to adjust

the markers for the effects of genomic wave due to dense areas of G or C base pairs. Then a population frequency of B allele (PFB) file was formed using the compile_pfb.pl script. PFB file curates a list of SNPs to be used for CNV calling, together with the BAF and genomic coordinates. Lastly, HMM file is obtained which provides the expected signal intensity values and transition probabilities for different copy number states. Each sample was adjusted for genomic wave using the genomic_wave.pl script, which reduces the rate of false-positive CNV calls due to local GC content fluctuations.

PennCNV utilizes Vitebri algorithm to determine copy-number state at each SNP, then identifies stretches of markers deviating from expected values which are classified as CNVs. PennCNV has 6 possible outcomes for each marker and called CNVs, these states (for diploid organisms and autosomal loci), are:

0. Double deletion (can only be a homozygous deletion),
1. Single deletion (a heterozygous deletion),
2. Normal state,
3. Single duplication (a heterozygous duplication),
4. Double duplication (either a heterozygous triplication or a homozygous duplication),
5. Normal state with loss of heterozygosity.

Groups of trios were subsequently jointly processed through PennCNV for trio calling using the detect_cnv.pl -joint script. This process is computationally intensive, thus I used 108 cores with 7GB per core on HiPerGator to parallelize the joint calling process. Finally, since PennCNV has a tendency of artificially splitting large CNVs, clean_cnv.pl combineseg script was used to reanneal them and remove this technical artifact prior to post-calling QC. In addition to joint calling, probands were individually called to generate sample-level intensity QC reports (which do not get generated with joint mode).

Post-Calling QC

Once all the calling is done, merged CNVs, unmerged joint-called CNVs, and QC output summaries from individually called CNVs were imported into R, where they were wrangled to form CNV call-sets annotated with CNV calls, QC information, and *de novo* vs. inherited status. This is also a master CNV call-set which will be subjected to further analyses.

Separately, for each proband phenotype annotation files were prepared as follows. For TAAICG probands, the rows were indexed by proband ID, then the following information were obtained for each proband: batch, phenotypic and SNP sex, and diagnostic information for each proband and their parents (included TS, PMVT, PTD, TS not otherwise specified, OCD, and ADHD). For SPARK probands, the rows were indexed by proband ID, then the following information were obtained for each proband: sex, ASD diagnostic information for each proband and their parents.

CNVs covering small number of SNPs or small genomic regions are more likely to be unreliable or false positives. Thus, CNVs spanning less than 10 SNPs and less than 20kb are filtered out. After filtering, 25 trios from TAAICG group, 923 trios from SPARK-ASD group, and 900 trios from SPARK-SIB group had no CNVs, resulting in their removal from the study. Table 6-3 summarizes number of CNVs after filtering for size and SNP count for CNVs overall and *de novo* CNVs.

Subsequently, CNVs spanning telomeric regions (200kb from either end of the chromosomes), centromeric regions (+100kb on either end of the centromeres), and immunoglobulin and T-cell receptor regions were all removed. These regions are prone to copy-number variation and not of immediate interest, or likely to harbor CNVs with true effect on pathology. Out of 42,049 CNVs from the total call-set, 2,255 were

removed due to overlaps with centromere, 3,681 were removed due to overlaps with immunoglobulin or T-cell receptor regions, and 893 were removed due to overlaps with telomeric regions. After filtering CNVs overlapping these loci, average number of CNVs per individual was 6.872, with maximum at 976. Overall, there were 59 individuals who had over 50 called CNVs, 25 of which had over 100 called CNVs. Individuals with over 100 called CNVs were excluded from further analysis.

The remaining samples were annotated with QC metrics, LRR_{SD} and BAF_{SD}, and inspected to determine if additional filtering was necessary. Cutoff values for each metric were determined for each group by taking their median and adding it to three times its standard deviation. Distribution of LRR_{SD} and BAF_{SD} are shown in Figures 6-1 and 6-2, respectively, together with the visualized cut-offs. This resulted in exclusion of 511 trios across the board.

Annotations

Gene annotations were determined using biomaRt package in R (Durnick et al., 2005; Durinck et al., 2009). Briefly, genomic positions of CNVs were used to find any overlapping genes and annotate which genes are overlapped and if genes are overlapped.

Global Burden Analysis

Global burden metric was assessed using generalized linear modelling in R. Burden was examined for number of CNVs and average size of CNVs overall and stratified by de novo status and size bin (small CNVs < 100kb, intermediate CNVs < 500kb, large CNVs > 500kb). Associations were covaried for LRR_{SD}, sex, and 10 principal components. Regressions were fit for TAAICG probands using SPARK-ASD and SPARK-SIB as reference groups. Principal components for controlling for

population structure were determined using the same methods described in Chapter 5. I used first 4 principal components for modelling global burden.

Incidence Rate Ratio

De novo Incidence rate ratios (IRR) were calculated for TAAICG probands using SPARK-ASD and PARK-SIB as reference groups. Grouped IRRs were calculated for number of CNVs stratified by size bin (small CNVs < 100kb, intermediate CNVs < 500kb, large CNVs > 500kb). IRRs were calculated according to Equation 6-3:

$$IRR_{Dx} = \frac{N_{\text{De Novo CNVs Dx}} \div N_{\text{Dx}}}{N_{\text{De Novo CNVs Ctrl}} \div N_{\text{Controls}}} \tag{6-3}$$

where $N_{\text{De Novo CNVs Dx}}$ refers to number of *de novo* CNVs in TS cases (TAAICG), and the $N_{\text{De Novo CNVs Ctrl}}$ variable refers to the number of *de novo* CNVs in comparison group (SPARK-ASD or SPARK-SIB). N_{Controls} and N_{Dx} refer to the total number of individuals SPARK-ASD or SPARK-SIB, and TAAICG, respectively. The 95% confidence intervals were constructed according to Equations 6-4 and 6-5. Association tests were performed according to Equation 6-6.

$$95\%\ CI\ [IRR_{Dx}] = e^{\{\ln(IRR_{Dx}) \pm 1.96 \times SE[\ln(IRR_{Dx})]\}} \tag{6-4}$$

$$SE\ [\ln(IRR_{Dx})] = \sqrt{\frac{1}{N_{\text{De Novo CNVs Dx}}} + \frac{1}{N_{\text{De Novo CNVs Ctrl}}}} \tag{6-5}$$

$$z_{Dx} = \frac{\ln(IRR_{Dx})}{SE[\ln(IRR_{Dx})]} \tag{6-6}$$

Gene Tests

Association for prior identified genes were evaluated, including previously associated *NRXN1* and *CNTN6*. This analysis was focused on specifically rare CNVs, i.e. CNVs which occurred in over 1% (67) of the total combined sample were excluded from the analysis. In total, 6,940 CNVs were intergenic and excluded from this analysis, the remaining 26,442 CNVs were genic. Among the genic CNVs, 14,301 were too common and thus removed from the analysis, leaving 12,141 CNVs for burden and gene-based analysis. For gene tests specifically, 8,992 genes in total were testable. From these, 5,132 were protein coding and used as final set of genes to test. Fisher's exact tests were used to compare number of CNVs spanning given gene in TS (using ASD and unaffected siblings as controls) and ASD (using unaffected siblings as controls). Prior to test, CNVs were stratified by type (deletions or duplications). One-sided p-values were obtained and controlled for FDR to account for multiple testing. Resulting significantly associated genes were subjected to GO analysis (as described in GO methods section of Chapter 5).

Results

Incidence Rate Ratios

Analysis of *de novo* IRRs has shown TS probands to have lower *de novo* incidence rates compared to both ASD probands and unaffected siblings of ASD probands. With $IRR_{TS\text{-}ASD} = 0.62$ ($p_{TS\text{-}ASD} < 0.001$) and $IRR_{TS\text{-}SIB} = 0.66$ ($p_{TS\text{-}SIB} < 0.001$). Genic incidence rates were higher in ASD than TS ($IRR_{TS\text{-}ASD} = 0.91$; $p_{TS\text{-}ASD} = 0.012$). Genic *de novo* incidence rates were higher in ASD than both TS and unaffected siblings of ASD, with $IRR_{TS\text{-}ASD} = 0.74$ ($p_{TS\text{-}ASD} = 0.001$) and $IRR_{ASD\text{-}SIB} = 1.33$ ($p_{ASD\text{-}SIB} < 0.001$),

respectively. Increased incidence of either *de novo*, genic, or *de novo* genic CNVs was not observed in TS samples (Table 6-4 and Table 6-5).

Burden Tests

Burden tests controlling for LRR_{SD}, sex, and first 4 ancestry PCs has shown variable, yet consistently higher burden of CNV sizes and numbers for ASD and unaffected siblings compared to TS probands. Stratifications for burden tests were done based on *de novo* status (considering all CNVs or *de novo* CNVs only), mutation type (all CNVs, only deletions, and only duplications), genic overlap and frequency (all called CNVs or rare genic CNVs), and size bins. All CNV global burden tests are visualized in series of figures starting with Figure 6-3 through Figure 6-26.

Overall, unexpected patterns emerge in global burden tests. Namely, consistent patterns of associations between CNV numbers and SPARK samples emerge, contrasted by consistent patterns of association between average CNV sizes and TAAICG sample. When looking at rare genic CNVs specifically, this pattern is less pronounced, yet persistent. Such structure indicates substantial batch effects and warrants caution when interpreting global burdens in this study.

Gene Associations

Analysis of genes spanned by rare CNVs resulted in 94 genes significantly associated with TS as compared to ASD probands and their siblings ($p_{FDR} < 0.05$). Table 6-7 is showing 41 of those genes, associating with TS-case status at $p_{FDR} < 0.01$. Graphical representation of all associations comparing TS to unaffected siblings of ASD probands (Figure 6-27), TS to ASD probands (Figure 6-28), and ASD probands to their unaffected siblings (Figure 6-29), are shown in Miami plots – indicating strength of association (in terms of $-\log_{10}(p_{FDR})$), stratified by type of mutation (duplications above

and deletions below abscissa). GO analysis of significantly associated genes showed no significant trends in terms of biological processes, molecular functions, or cellular components clusters.

Previously identified candidates *CNTN6* and *NRXN1* were successfully validated in this study. For *NRXN1* deletions, odds ratios were $OR_{TS\text{-}ASD}$ = 1.68 (p_{FDR} = 0.10) and $OR_{TS\text{-}SIB}$ = 4.84 (p_{FDR} = 0.002). No significant relationship was observed when stratified by *de novo* status or size of CNVs. For *CNTN6* duplications, odds ratios were $OR_{TS\text{-}ASD}$ = 6.48 (p_{FDR} = 0.0009) and $OR_{TS\text{-}SIB}$ = 6.22 (p_{FDR} = 0.0009). No significant relationship was observed when stratified by *de novo* status or size of CNVs.

Other notable associations from Table 6-7 include *PCNT*, *SSTR5*, *TBP*, and *TRPM1* which are all genes previously implicated in neuropsychiatric conditions like bipolar disorder, depression, and schizophrenia.

In analysis of ASD probands with respect to their unaffected siblings, only 21 genes were associated at $p_{FDR} < 0.05$, and only 9 were associated at $p_{FDR} < 0.01$.

Discussion

CNV analysis in TS and ASD data has shown to be challenging, generally defying expectations of higher CNV risk and burden in CNV metrics among the TS probands compared to the unaffected ASD siblings, but not necessarily the affected ASD probands. Namely, *de novo* IRR measurements were consistently higher for ASD probands and their unaffected siblings, with the risk ratio being higher for ASD-affected probands. This is likely since siblings of ASD probands, even though they don't have ASD diagnosis, still carry many ASD-related genomic risk variants – CNVs included.

Another, more plausible explanation, is phenotype misclassification – due to the absence of ASD-assessment of the unaffected siblings of ASD probands, there might

be an influential amount of undiagnosed ASD cases among the unaffected siblings. This could potentially explain the unexpected trends in global burden.

I have successfully validated gene-specific risk of TS by replicating findings from Huang et al. (2017) reports of *NRXN1* deletions and *CNTN6* duplications as risk variants for TS. *NRXN1* risk was not significant when compared to ASD group as opposed to the unaffected sibling group. This is likely because *NRXN1* deletions also carry marked risk for ASD. This was not true for *CNTN6*, which appears to be TS-specific effector. Both *NRXN1* and *CNTN6* mutations were predominantly inherited.

Burden analysis has shown CNV numbers and average CNV sizes to carry risk predominantly for ASD, rather than TS. However, this relationship was different when it came to average CNV sizes of *de novo* mutations, which were found to carry risk for TS when compared to unaffected siblings of ASD probands. Re-assessment of CNV merging pipeline and post-calling quality control indicate these effects are not due to technical or random error, but persistent batch effects between the two consortia sample collections (TAAICG and SPARK). These batch effects represent an important obstacle for associations and their interpretations that needs to be addressed.

Potential reasons for these batch effects can be boiled down to (1) insufficiently controlled lower-level batch effects, (2) phenotype misclassifications, and (3) intrinsic differences between the two samples that introduce variability during sample collection or wet lab processing. Insufficiently controlled lower-level batch effects could result in higher rates of “interruptions” to CNV calling and result in artifacts observed in this study. However, lower-level batch effects are difficult to control for due to missing data. For example, TAAICG samples were independently clustered based on genotyping

waves. SPARK data, however, were not processed in same manner due to absence of those information. These batches can be potentially identified using unsupervised machine learning approaches, which is one of the avenues currently explored.

Phenotype misclassification can cause substantial skews in associations, especially in genetic studies of psychiatric phenotypes – as it was discussed in length in previous chapters. ASD phenotypes in SPARK datasets were based on self-reports of existing professional diagnosis, meaning that non-ASD individuals might be false negatives with less-penetrant forms of autism or have not yet been diagnosed. One potential way to address this issue is to stratify SPARK sample by other phenotypic measures that might be available (such as CBCL). These measures could be used to exclude extreme symptomatic cases within unaffected sibling group, potentially yielding more robust phenotypes. Lastly, intrinsic differences between the two samples that are not due to controllable factors would indicate that these two cohorts are ultimately analytically incompatible, and different samples altogether should be used. One obstacle to using different samples, however, includes different genotyping platform with different genome coverage density. I have developed a method that uses proximity-based matching of loci, which could address such obstacles, but that is beyond the scope of this aim and subject to future experimentation.

Unfortunately, the confounding batch effects have made it difficult to analyze and interpret CNV data. However, there are some notable silver linings. Namely, *NRXN1* and *CNTN6* validation, and identification of 39 additional genes that might play a role in TS pathogenesis. Similar in previous reports (like Huang et al., 2017), *NRXN1* deletions and *CNTN6* duplications occur in 2.11% of TS patients (about 1% each) in the TAAICG

sample, underlying their importance in molecular processes that could contribute to TS pathology. A lot of additional work remains to be done to fully understand TS and its pathomechanisms, and outcomes of this aim clearly emphasize the importance of methodology and sample characteristics in such efforts.

Table 6-1. Sample summaries.

Sample	Participants	Probands	Families	Mothers	Fathers
TAAICG	3,815	1,376	1,226	1,221	1,219
SPARK-ASD	8,341	2,893	2,724	2,724	2,724
SPARK-SIB	8,233	2,785	2,724	2,724	2,724

Table 6-2. Copy number metric summaries.

Sample	Batch	Med(LRR_{SD})	SD(LRR_{SD})	Med(BAF_{SD})	SD(BAF_{SD})
TAAICG	1	0.1403	0.0416	0.0385	0.0064
	2	0.1607	0.0715	0.0385	0.0087
	3	0.1435	0.0560	0.0309	0.0065
	Joint	0.1439	0.0530	0.0381	0.0070
SPARK	Joint	0.1300	0.0809	0.0411	0.0241

Table 6-3. CNV post-call summary.

Metric	TAAICG	SPARK-ASD	SPARK-SIB
CNV Calls			
Any	8,522 (6.11)	18,138 (9.21)	15,389 (8.16)
Double deletions	28 (0.02)	16 (0.01)	13 (0.01)
Single deletions	4,053 (3.00)	7,335 (3.72)	6,186 (3.28)
Single duplications	4,412 (3.27)	10,608 (5.38)	9,102 (4.83)
Double duplications	29 (0.02)	179 (0.09)	88 (0.05)
de novo			
Any	3,995 (2.96)	10,135 (5.14)	8,116 (4.31)
Double deletions	26 (0.02)	16 (0.01)	13 (0.01)
Single deletions	2,061 (1.53)	3,787 (1.92)	2,976 (1.58)
Single duplications	1,905 (1.41)	6,282 (3.19)	5,100 (2.71)
Double duplications	3 (0.00)	50 (0.03)	27 (0.01)

Note: number in the parenthesis shows sample-wise rates.

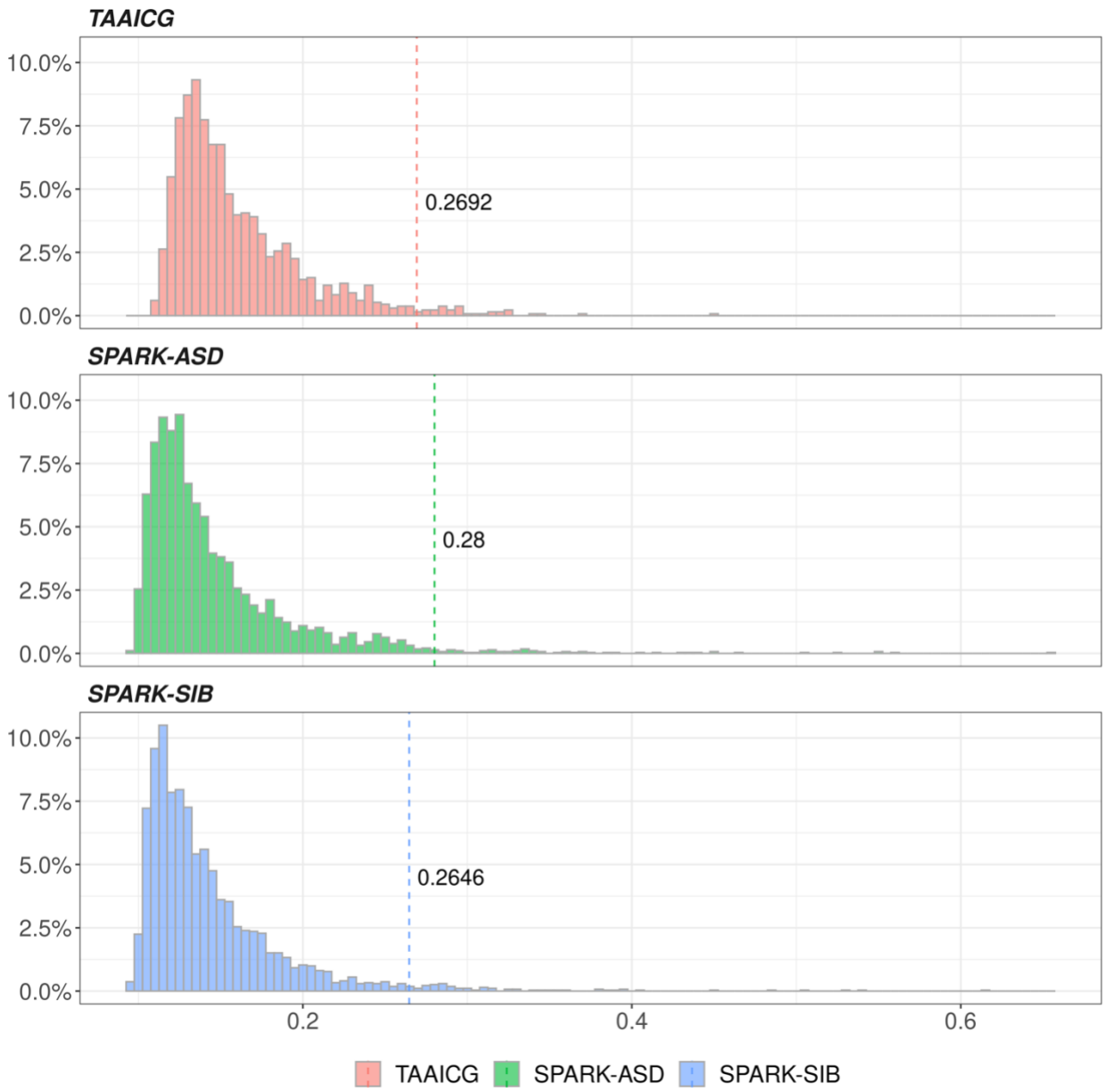

Figure 6-1. LRR_{SD} distributions across samples with cutoff-values for sample exclusion.

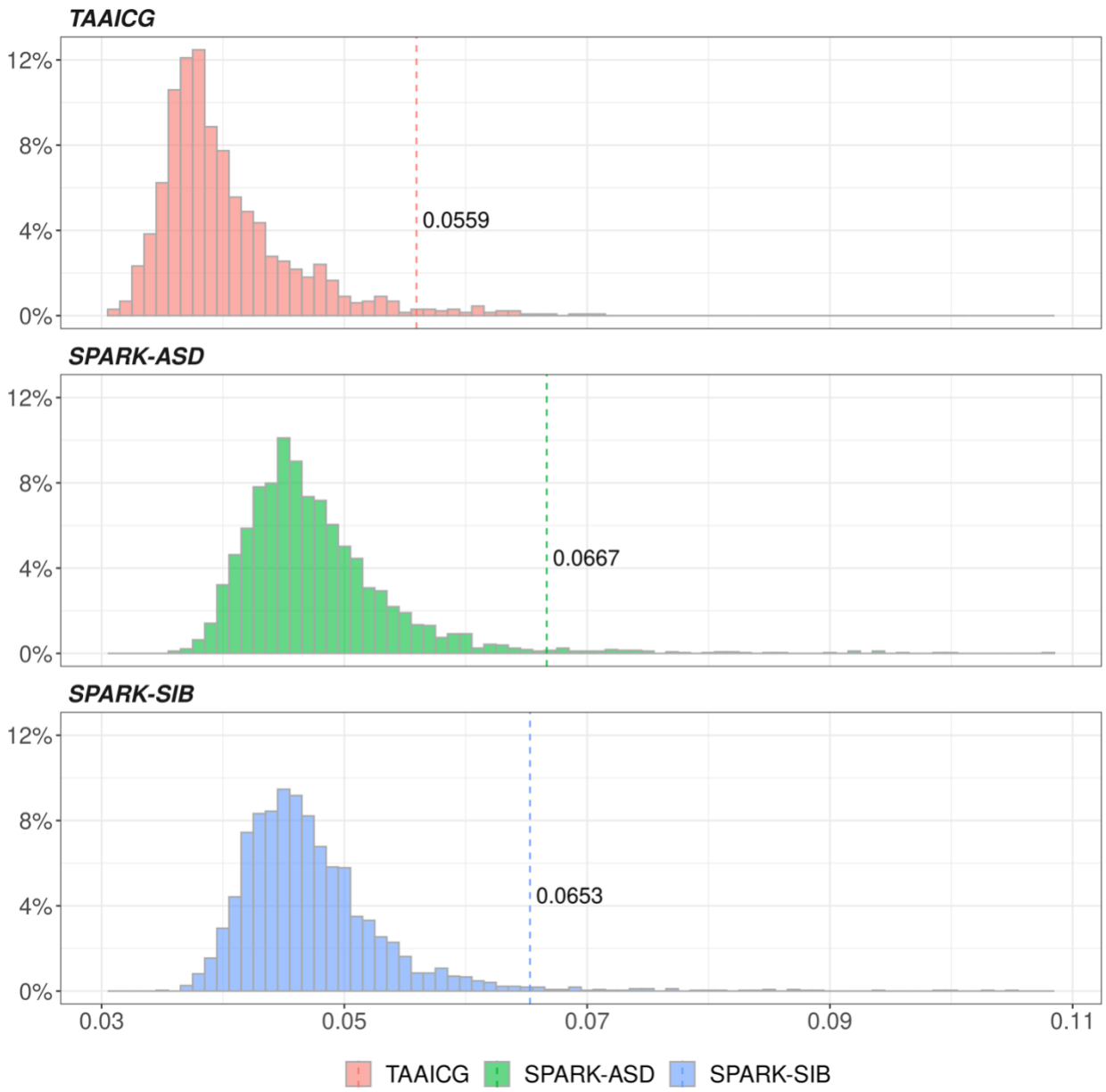

Figure 6-2. BAF_{SD} distributions across samples with cutoff-values for sample exclusion.

Table 6-4. Incidence rates of *de novo* and genic CNVs, by grup.

Group	Individuals with *de novo* CNV	genic CNV	*de novo* and genic CNV
TAAICG	377 (29.45%)	932 (72.81%)	136 (10.62%)
SPARK-ASD	1,107 (40.21%)	2,052 (74.54%)	380 (13.80%)
SPARK-SIB	1,024 (38.88%)	1,951 (74.07%)	283 (10.74%)

Table 6-5. Incidence rate ratios of *de novo* and genic CNVs.

Comparison	TS (ref. ASD)	TS (ref. SIB)	ASD (ref. SIB)
de novo CNV			
IRR	0.62	0.66	1.06
SE [ln(IRR)]	0.06	0.06	0.04
95% CI [IRR]	(0.55, 0.70)	(0.58, 0.74)	(0.97, 1.15)
p_z	6.45×10^{-16}	1.40×10^{-12}	0.90
genic CNV			
IRR	0.91	0.94	1.02
SE [ln(IRR)]	0.04	0.04	0.03
95% CI [IRR]	(0.85, 0.99)	(0.86, 1.01)	(0.96, 1.09)
p_z	0.01	0.05	0.22
de novo and genic CNV			
IRR	0.74	0.99	1.33
SE [ln(IRR)]	0.10	0.10	0.08
95% CI [IRR]	(0.61, 0.90)	(0.80, 1.21)	(1.14, 1.55)
p_z	1.44×10^{-3}	0.45	1.39×10^{-4}

Note: TS – TAAICG, ASD – SPARK-ASD, and SIB – SPARK-SIB.

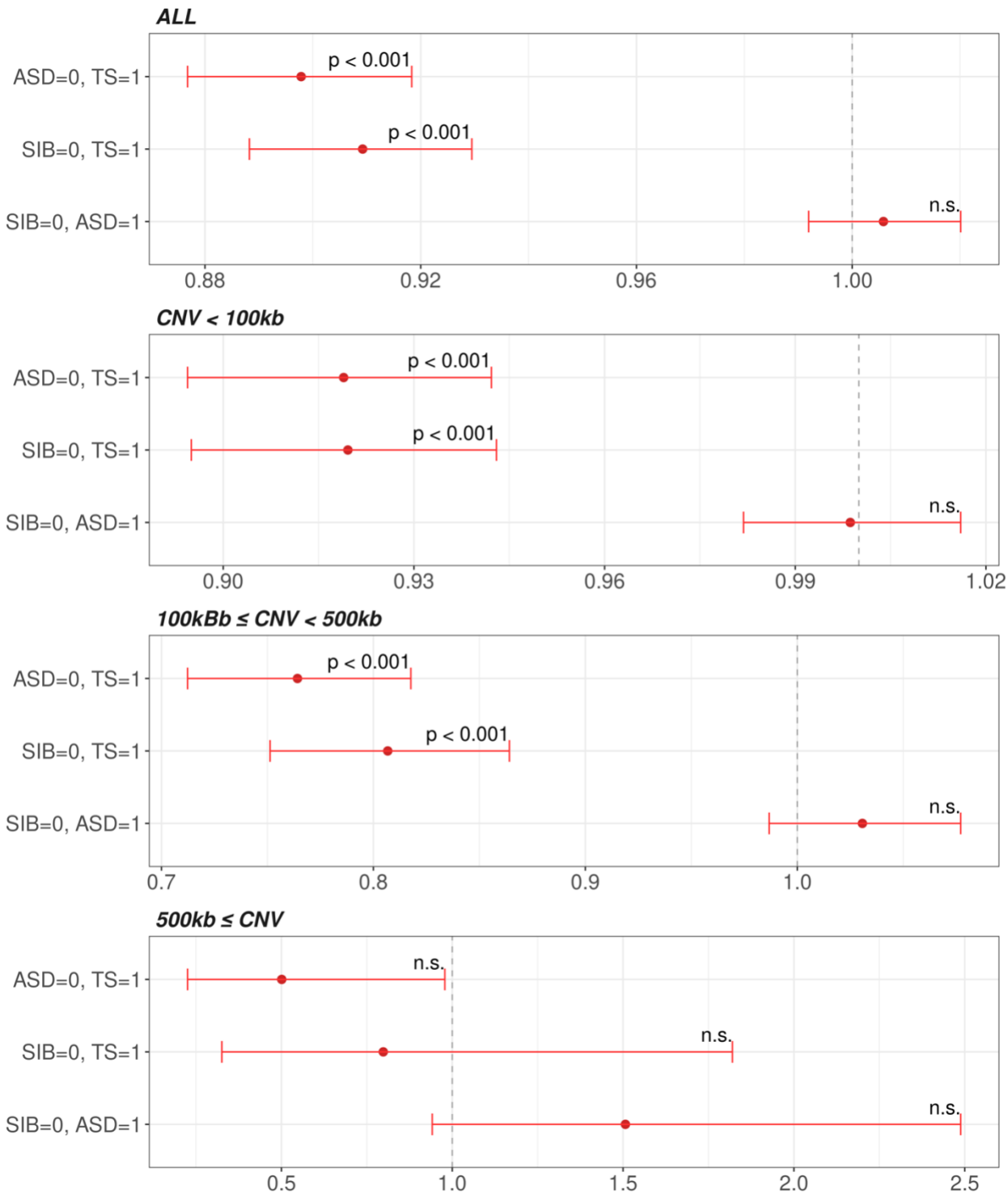

Figure 6-3. OR plots of CNV count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

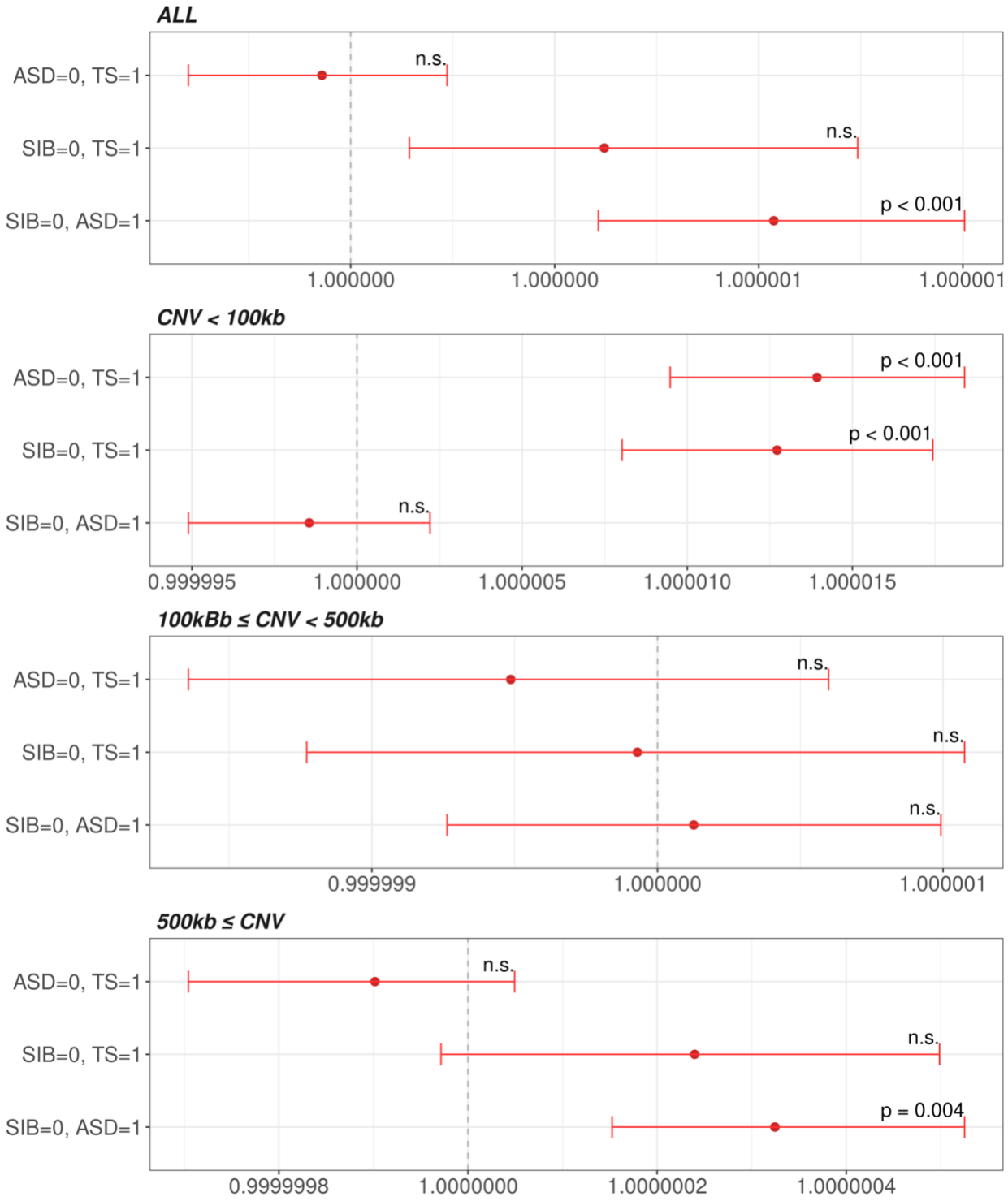

Figure 6-4. OR plots of average CNV size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

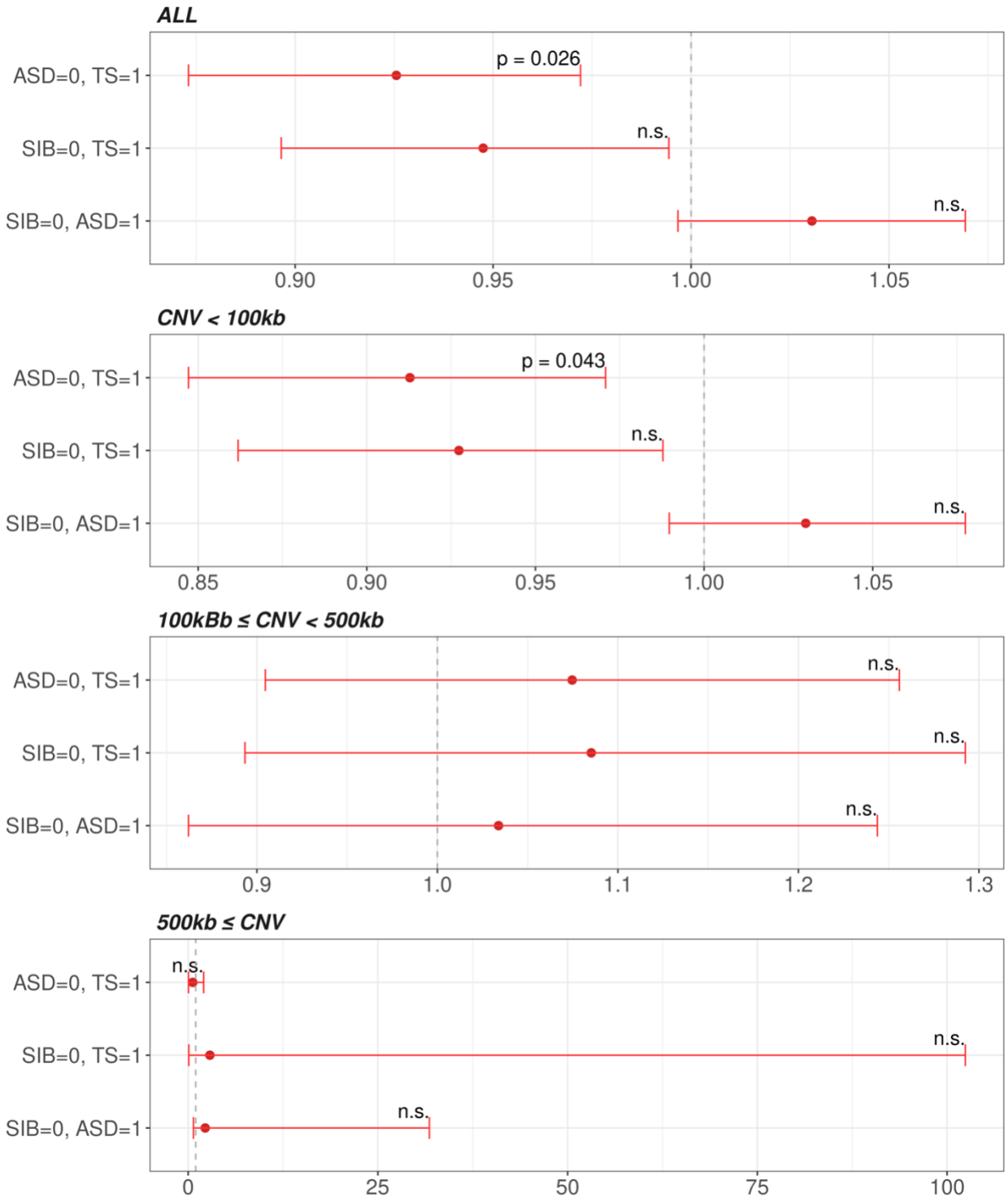

Figure 6-5. OR plots of *de novo* CNV count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

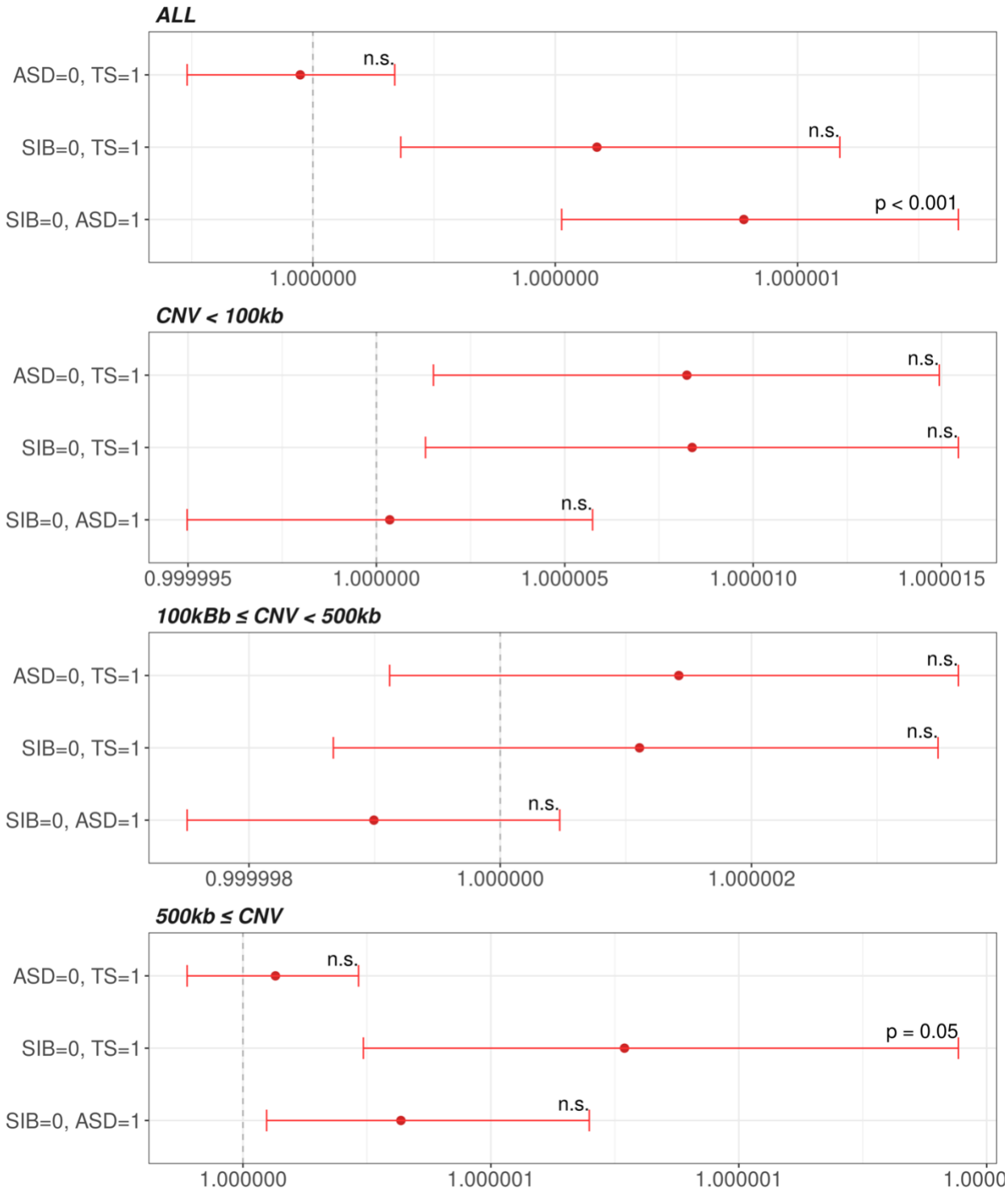

Figure 6-6. OR plots of average *de novo* CNV size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

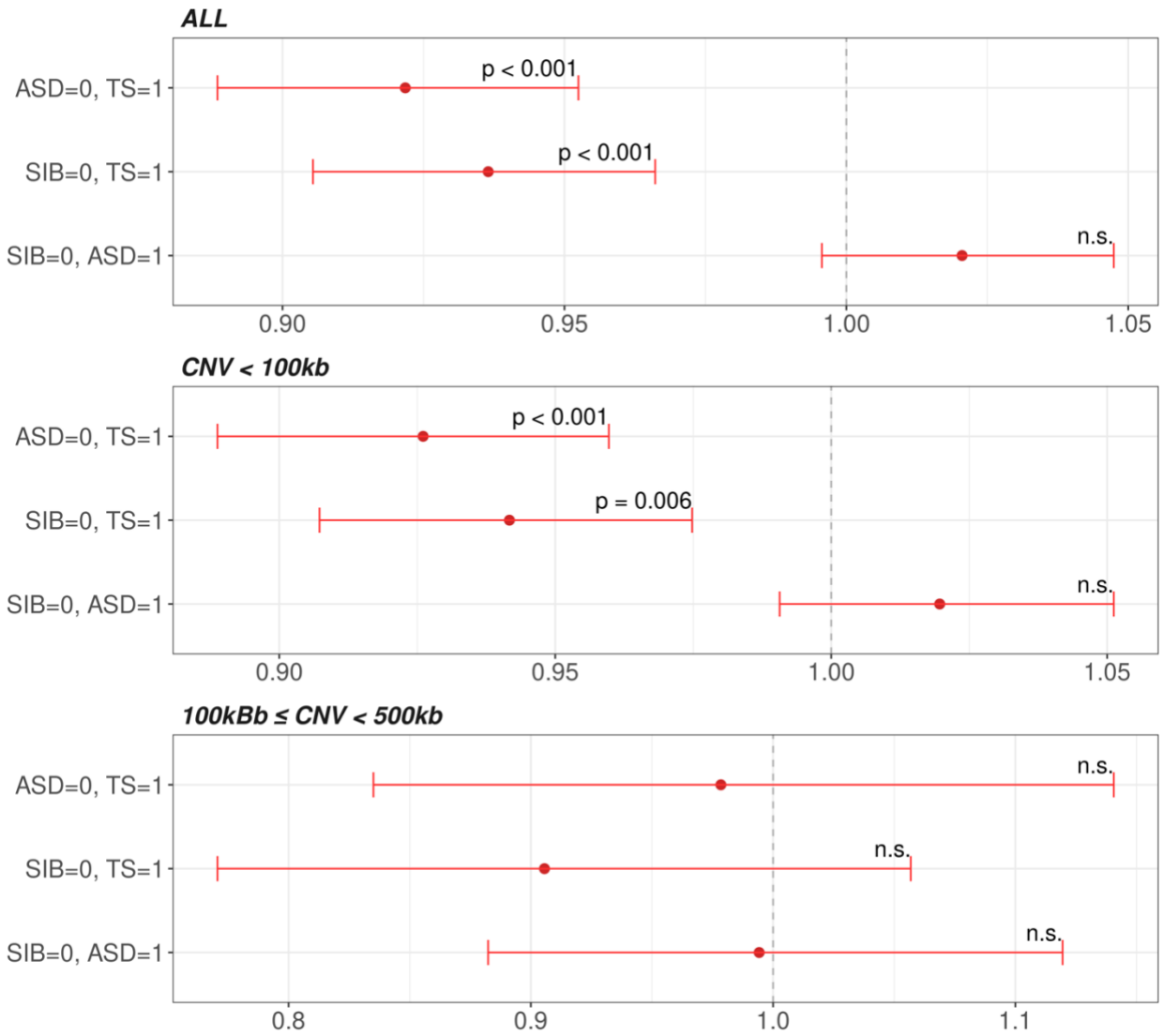

Figure 6-7. OR plots of CNV deletion count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

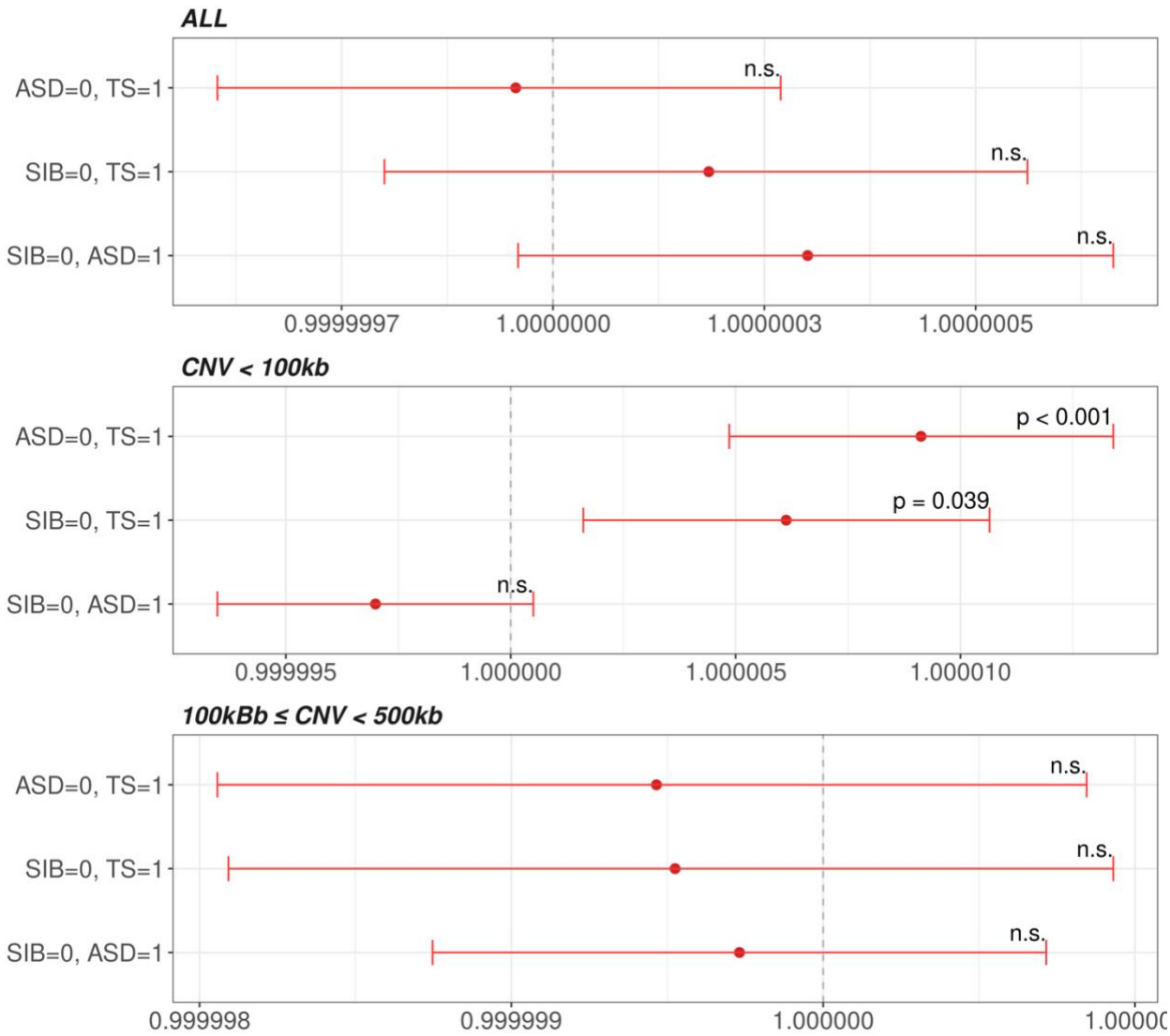

Figure 6-8. OR plots of average CNV deletion size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

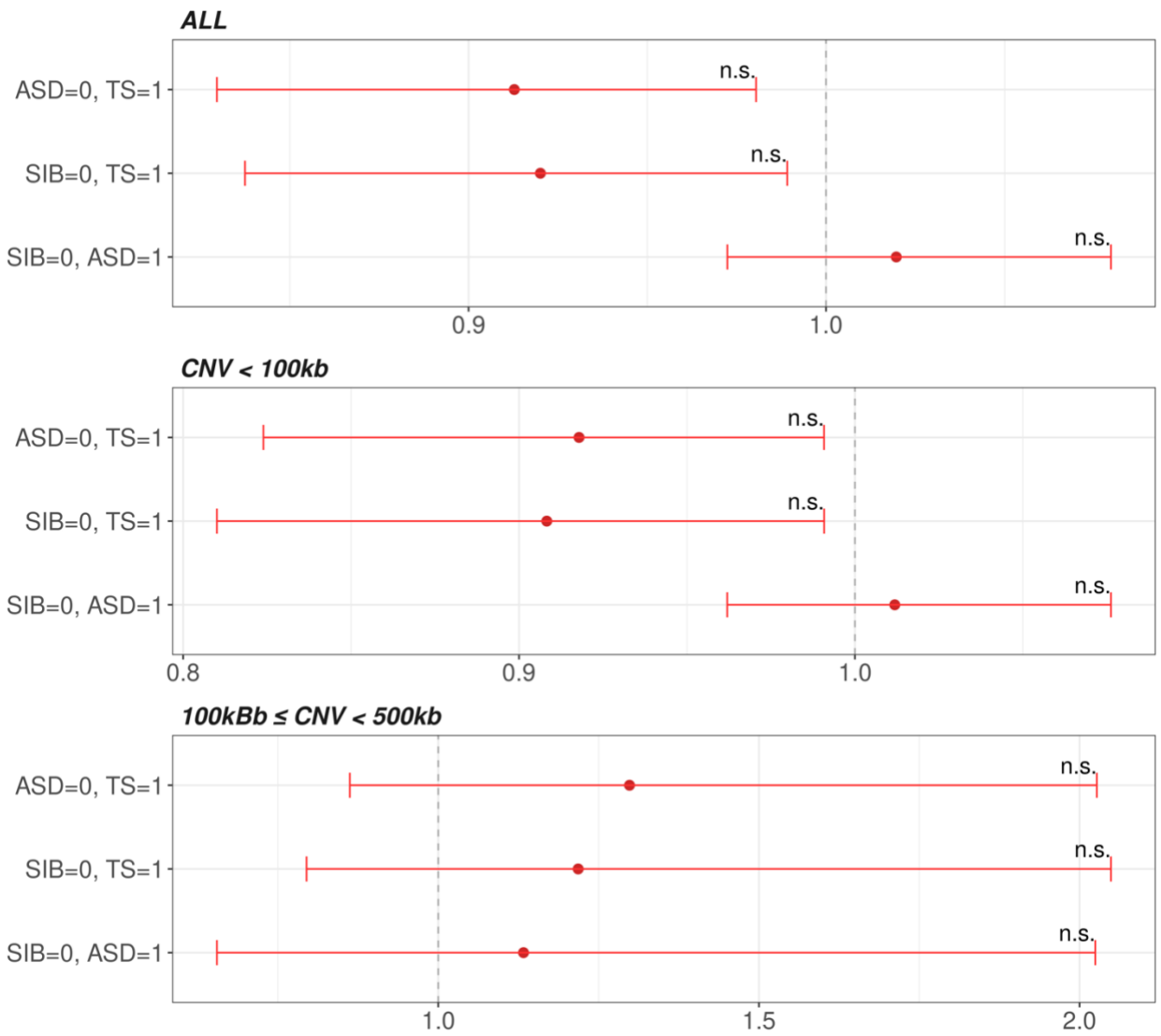

Figure 6-9. OR plots of *de novo* CNV deletion count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

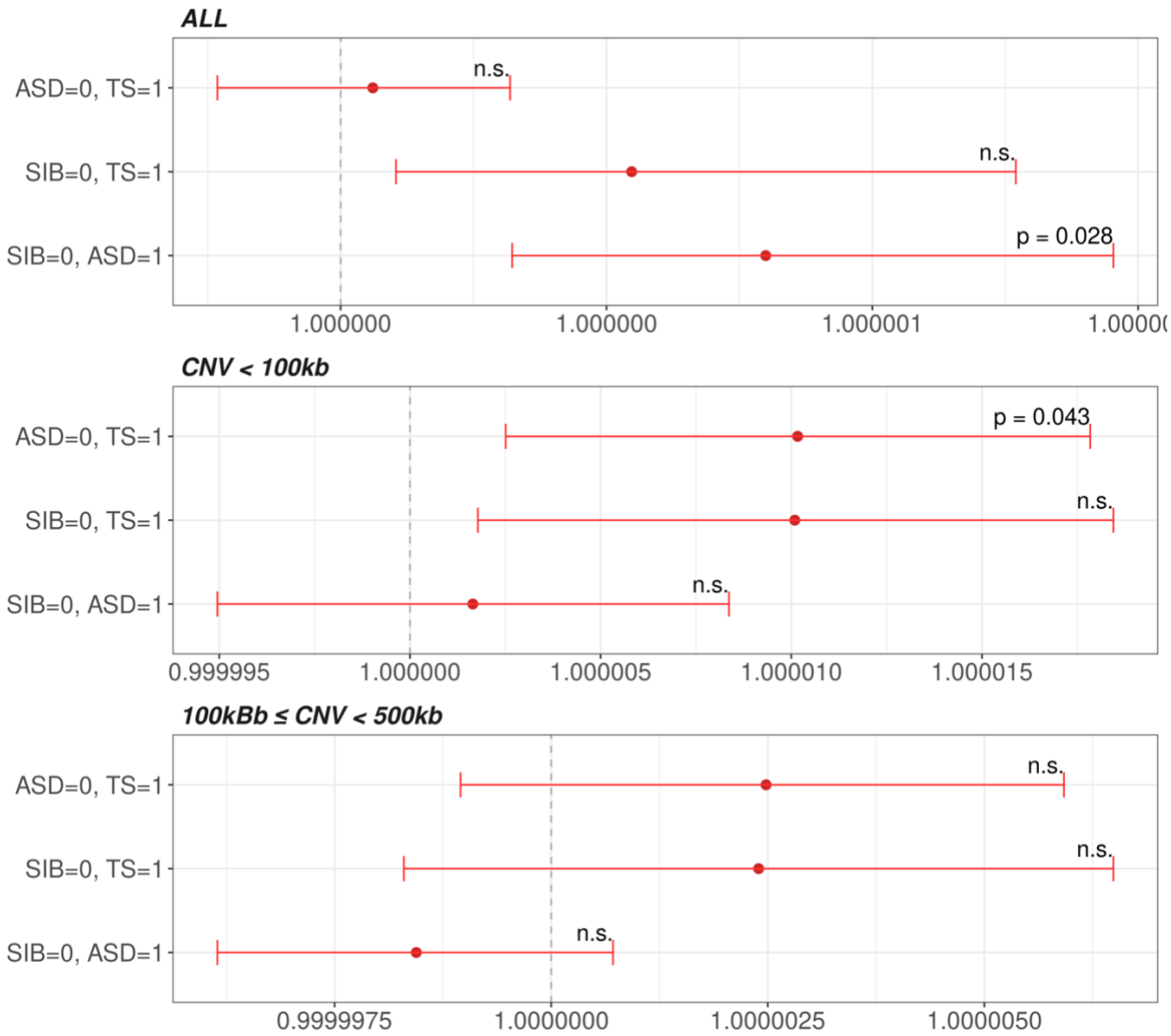

Figure 6-10. OR plots of average *de novo* CNV deletion size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

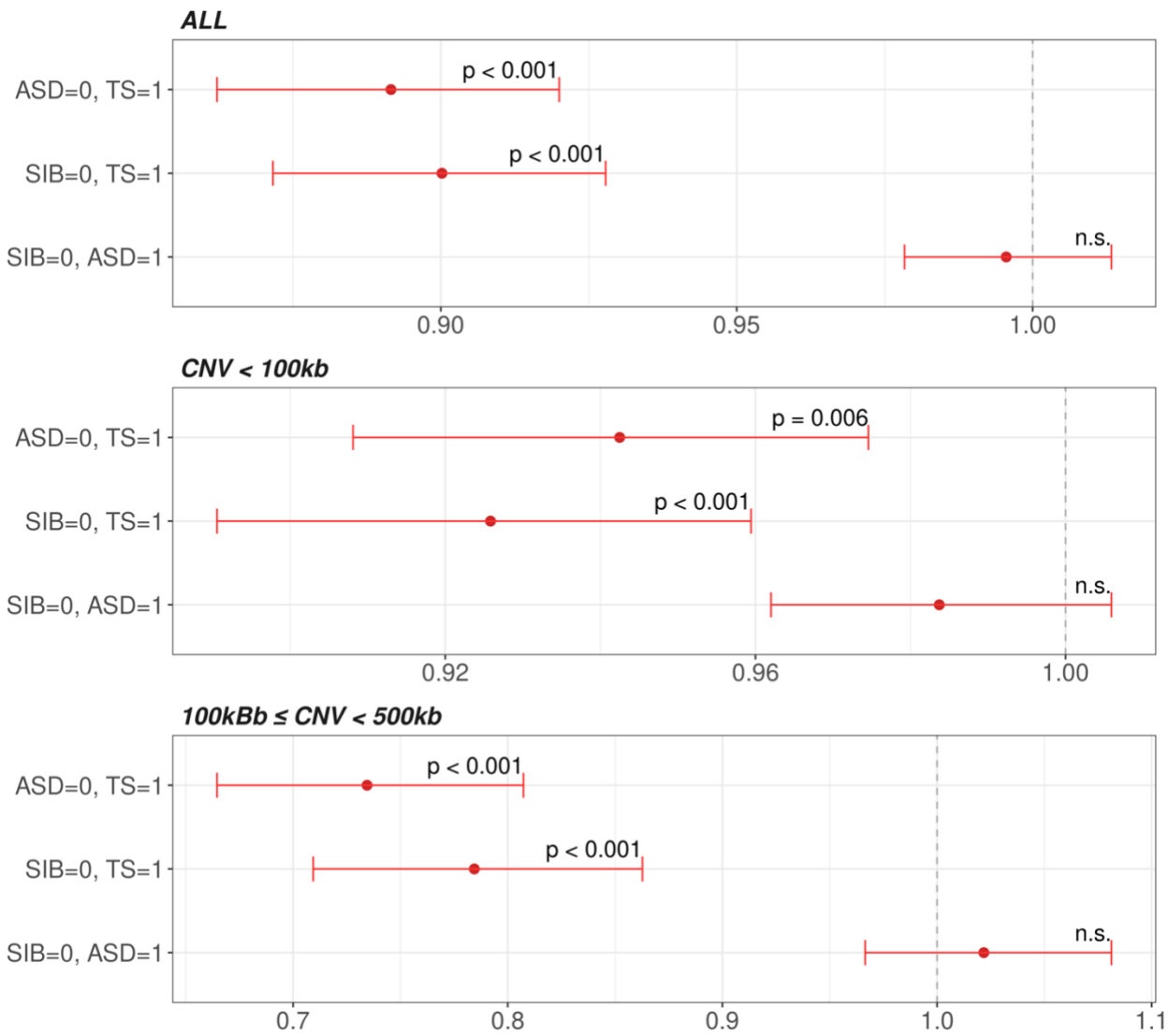

Figure 6-11. OR plots of CNV duplication count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

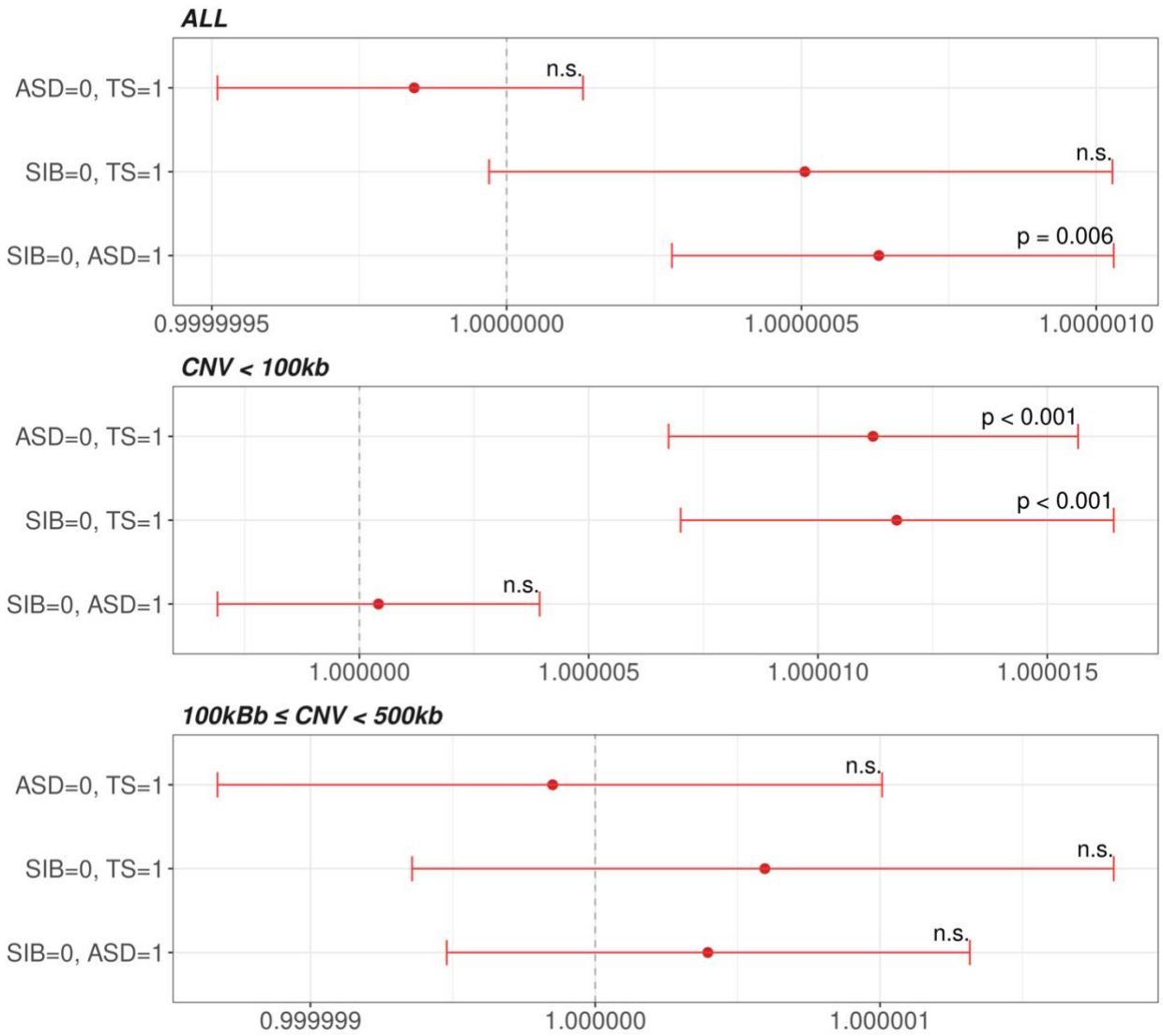

Figure 6-12. OR plots of average CNV duplication size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

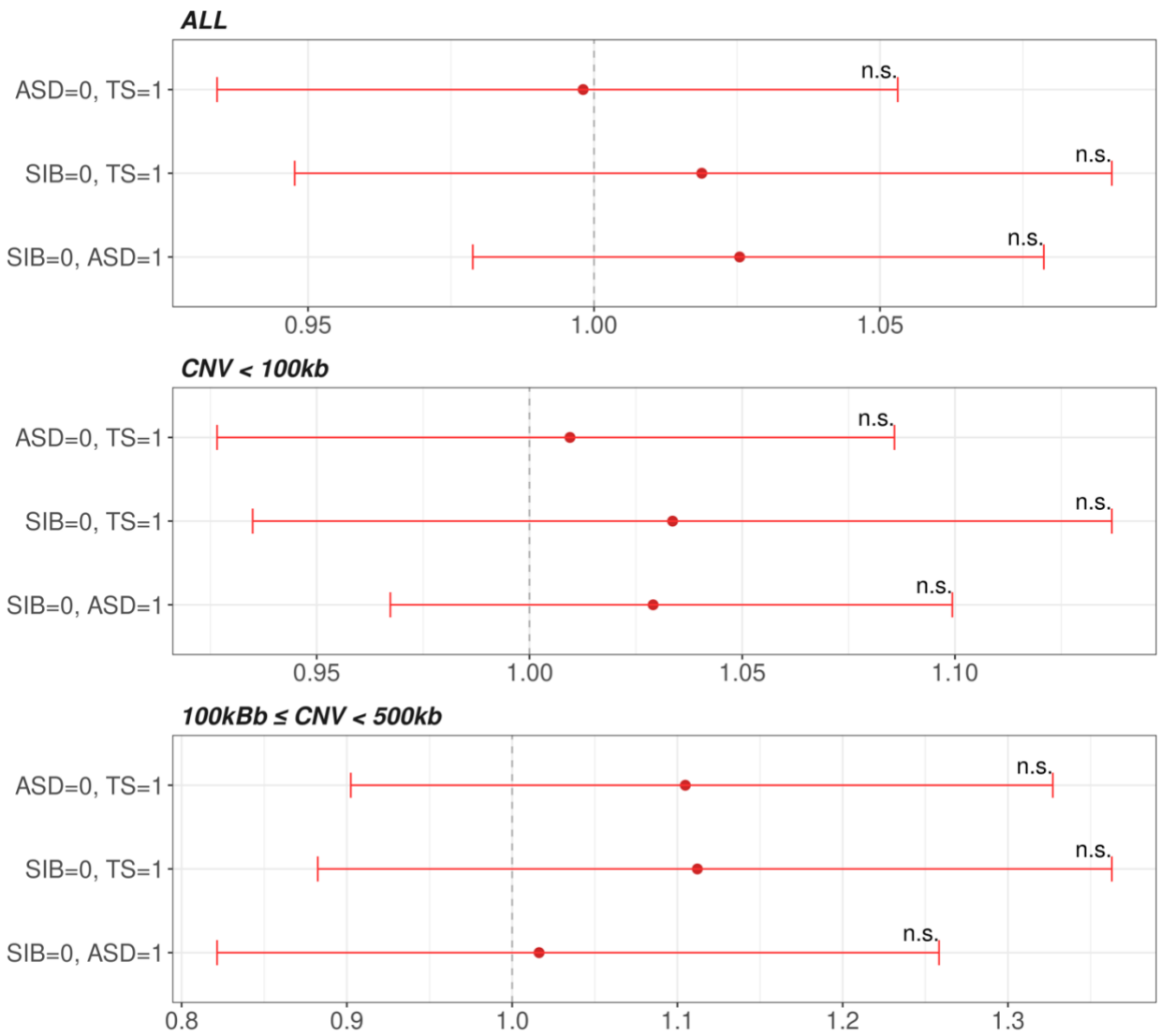

Figure 6-13. OR plots of *de novo* CNV duplication count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

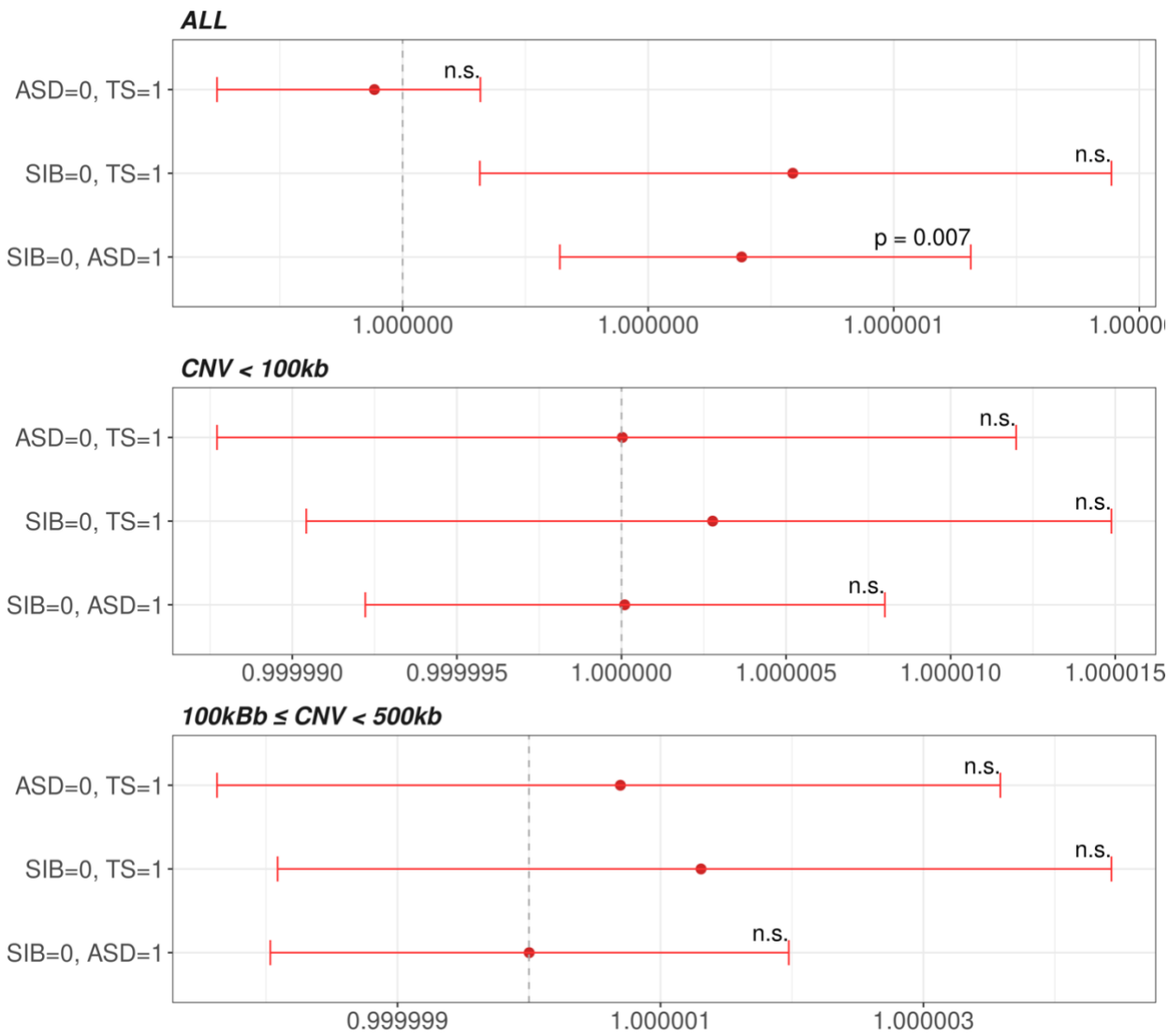

Figure 6-14. OR plots of average *de novo* CNV duplication size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

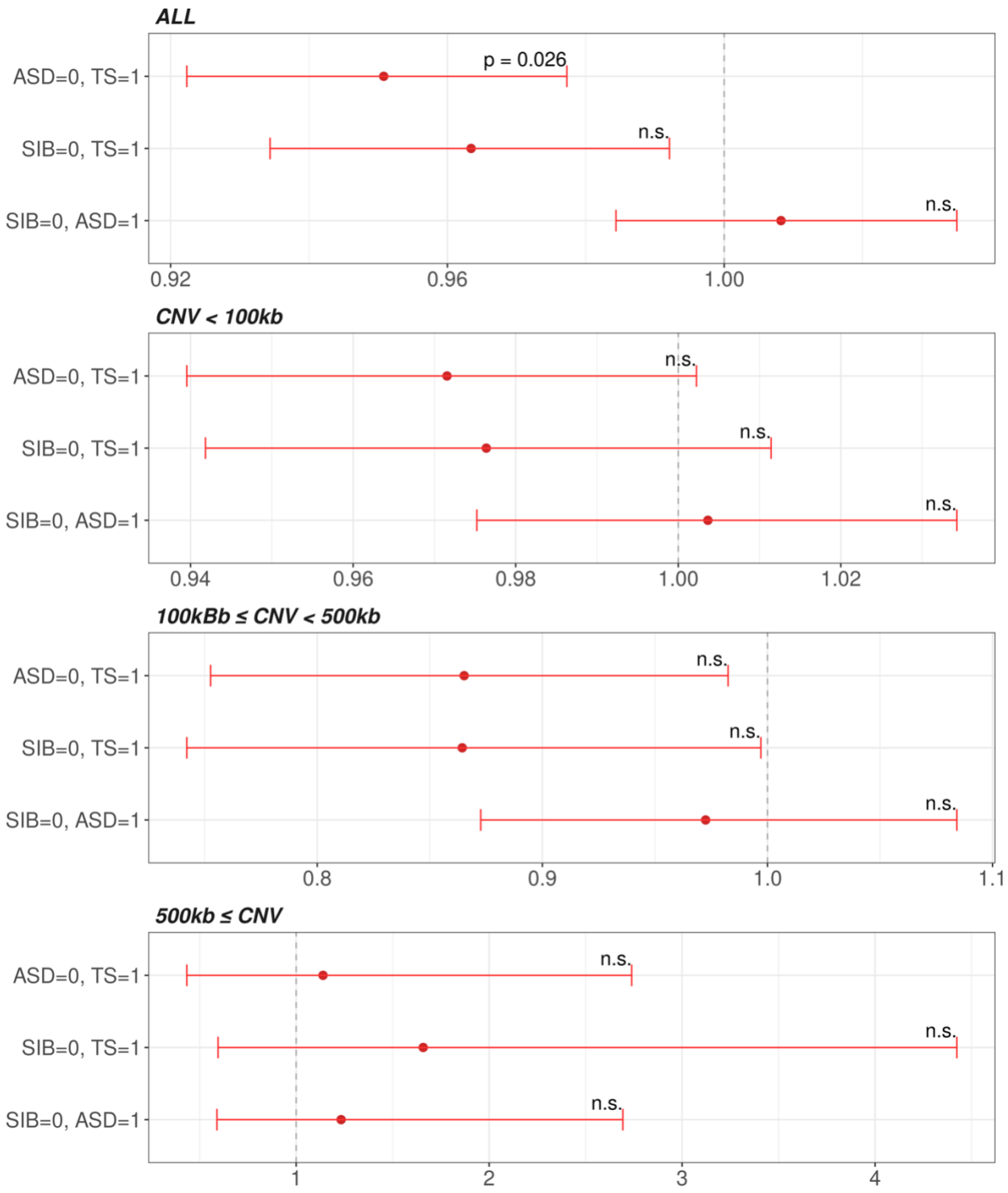

Figure 6-15. OR plots of rare genic CNV count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

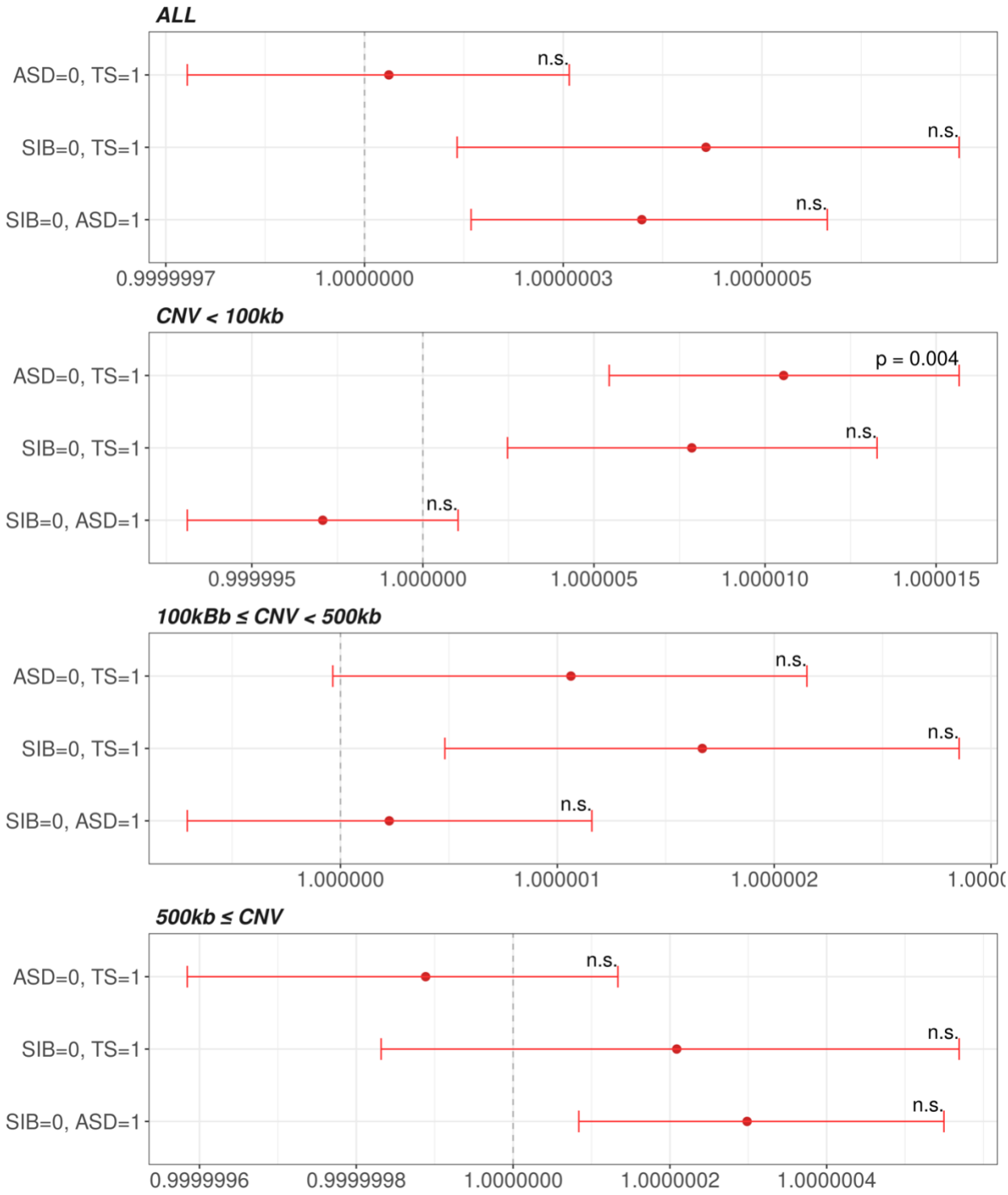

Figure 6-16. OR plots of average rare genic CNV size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

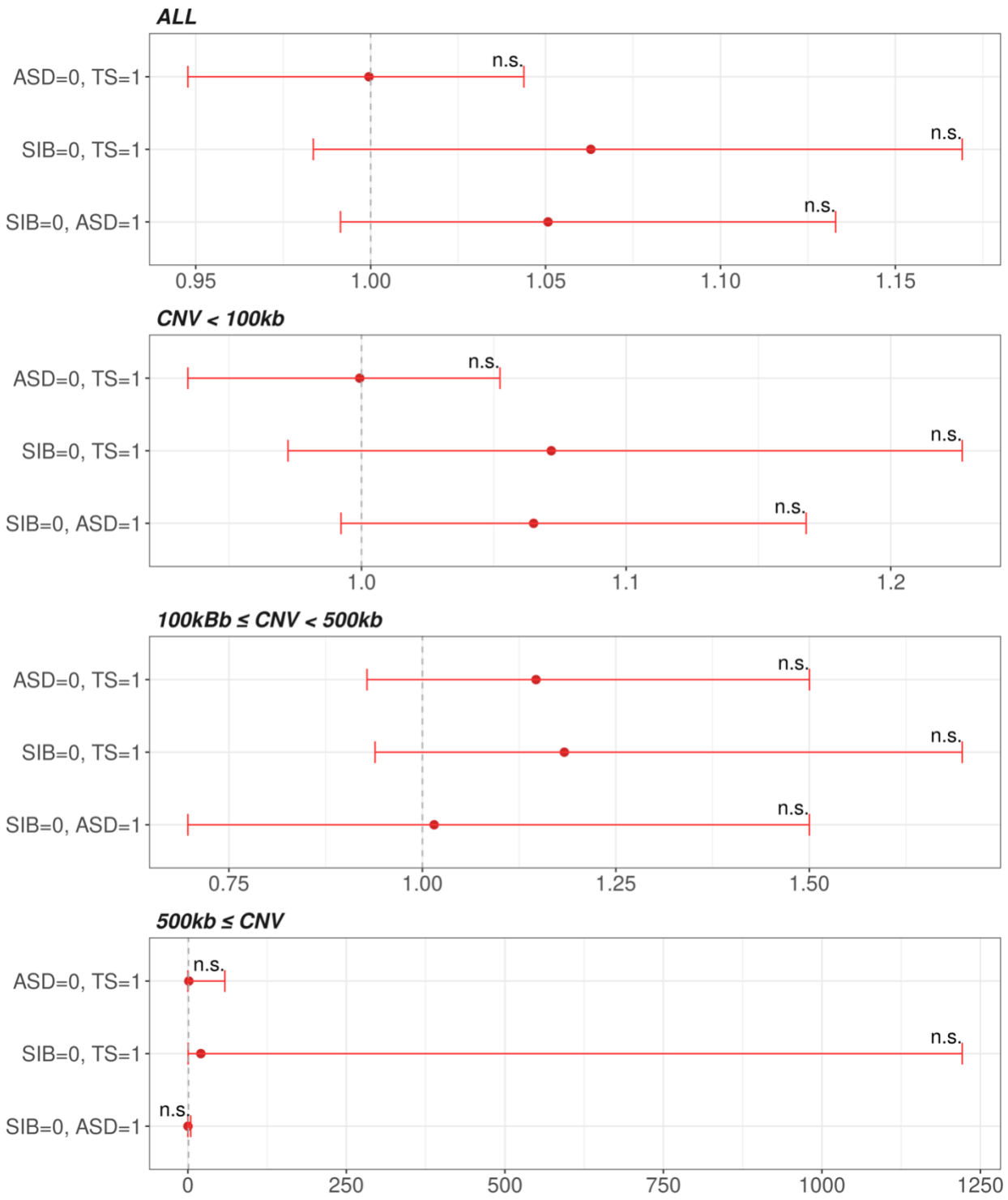

Figure 6-17. OR plots of *de novo* rare genic CNV count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

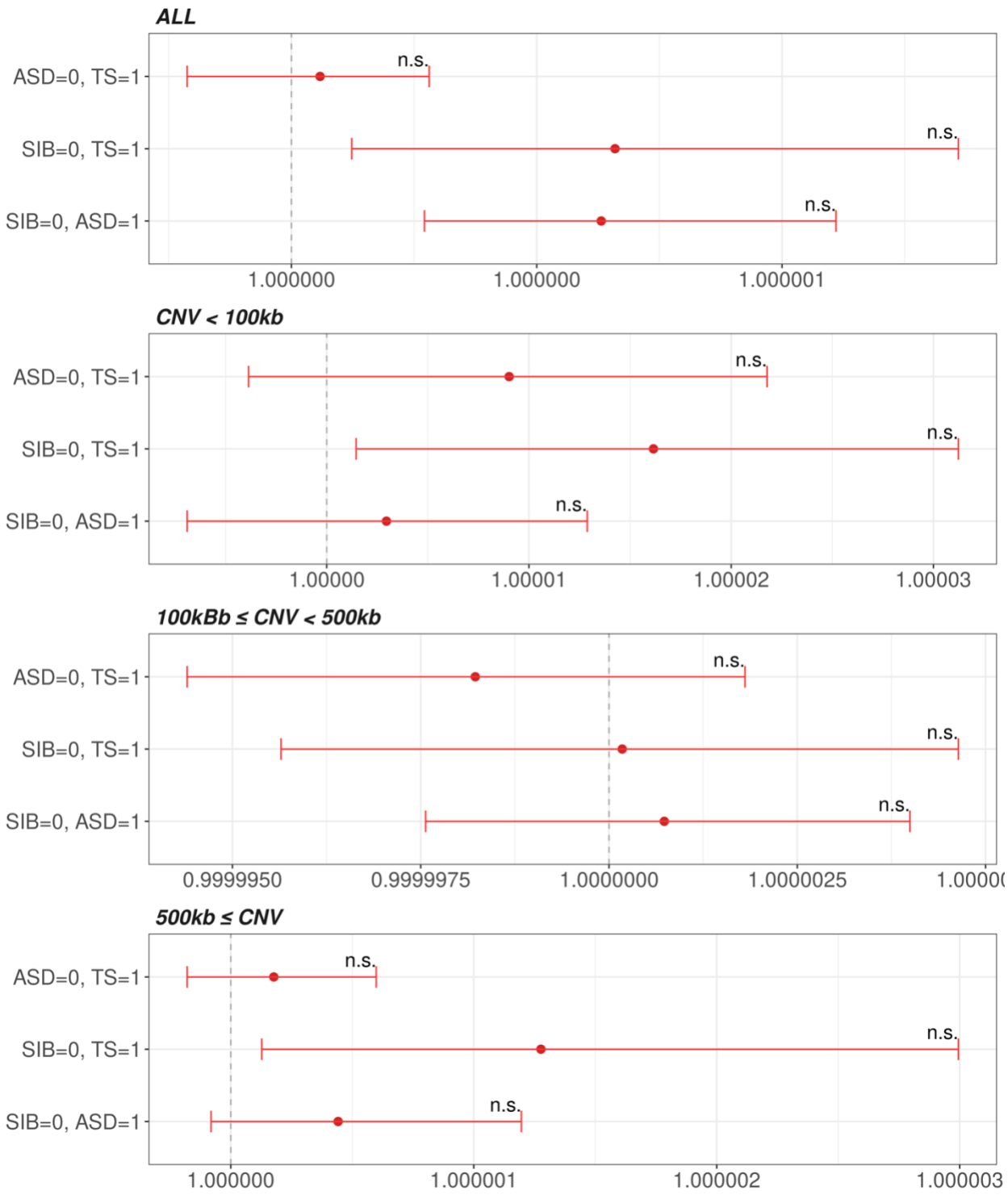

Figure 6-18. OR plots of average *de novo* rare genic CNV size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

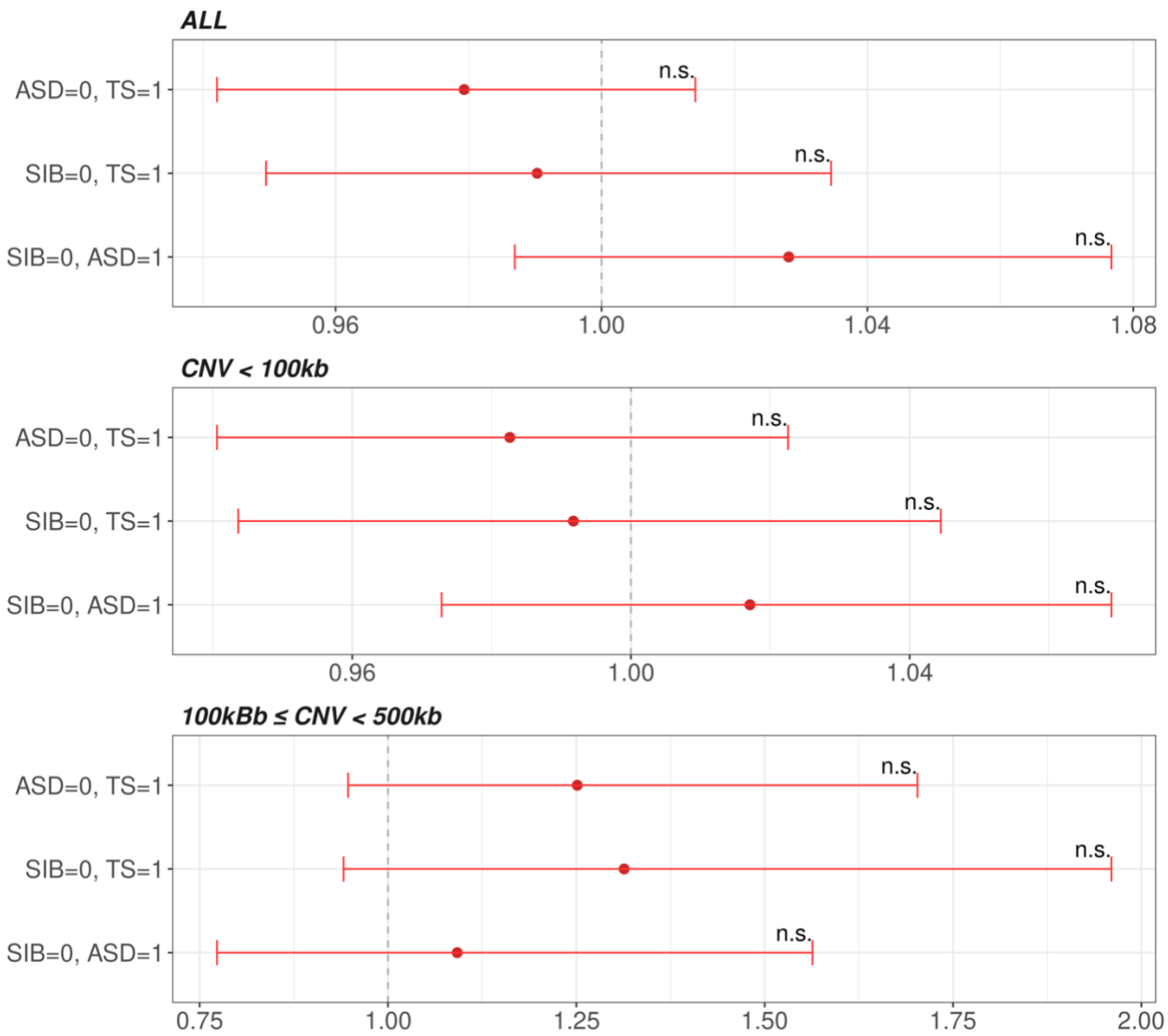

Figure 6-19. OR plots of rare genic CNV deletion count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

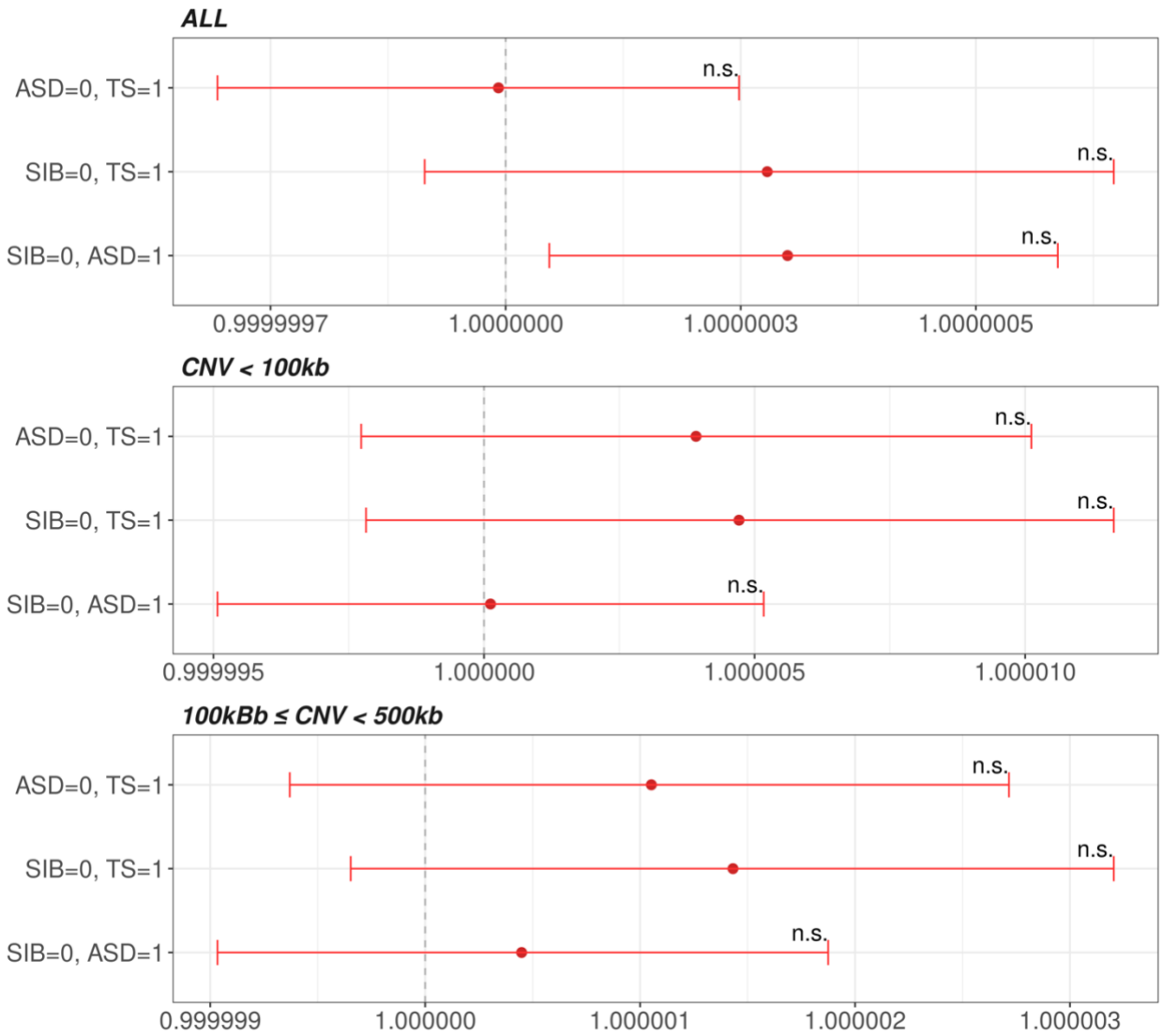

Figure 6-20. OR plots of average rare genic CNV deletion size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

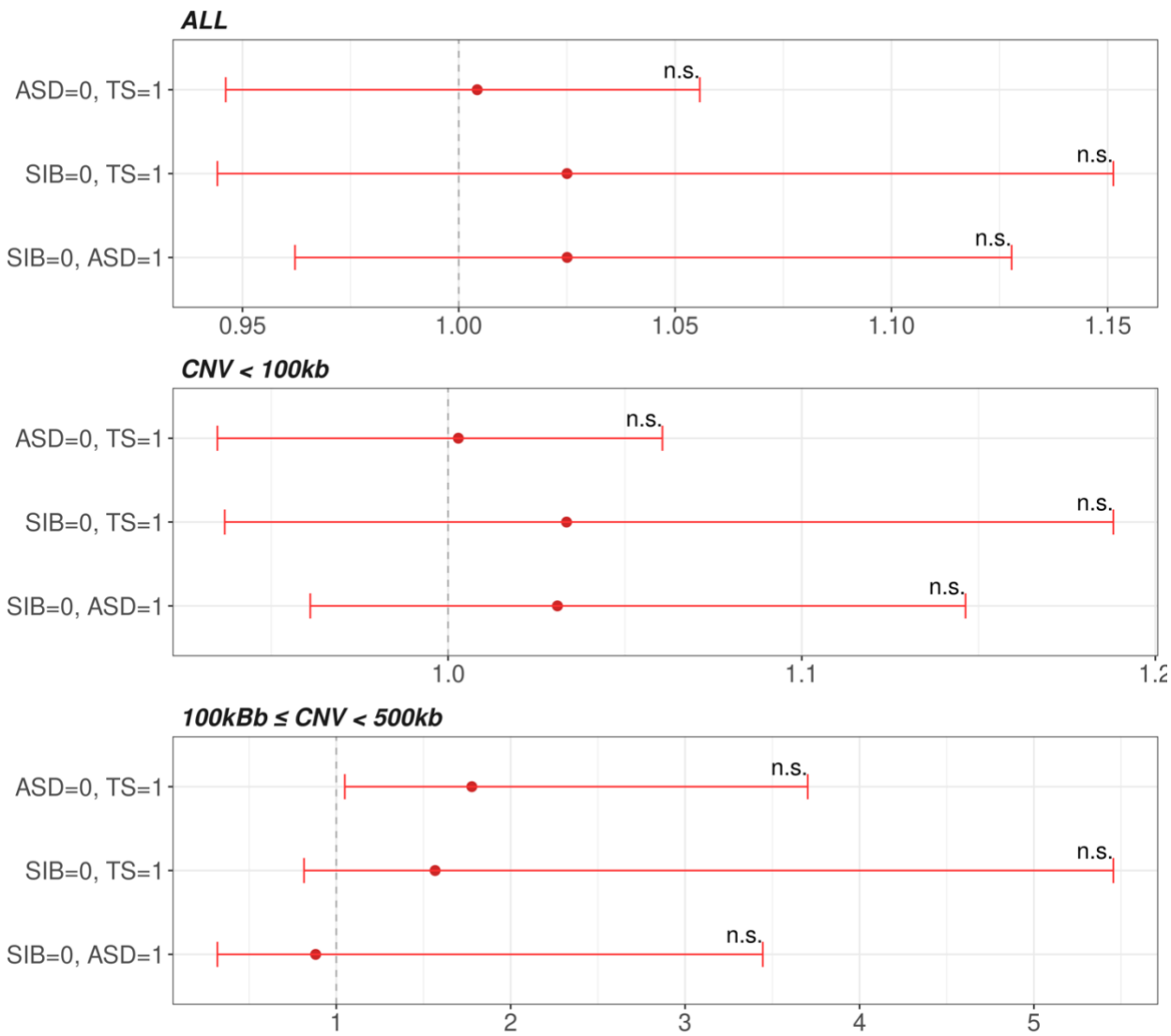

Figure 6-21. OR plots of *de novo* rare genic CNV deletion count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

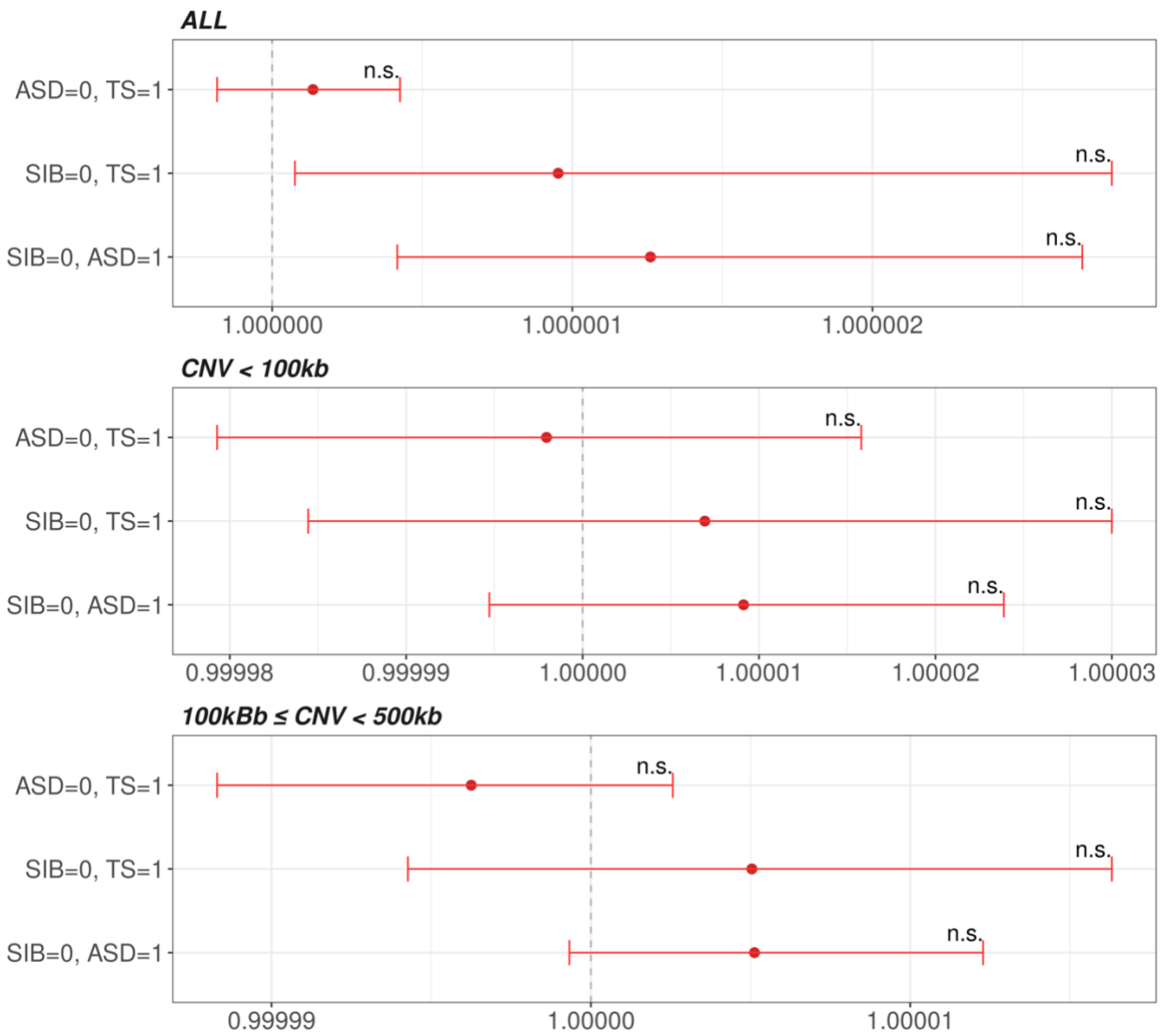

Figure 6-22. OR plots of average *de novo* rare genic CNV deletion size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

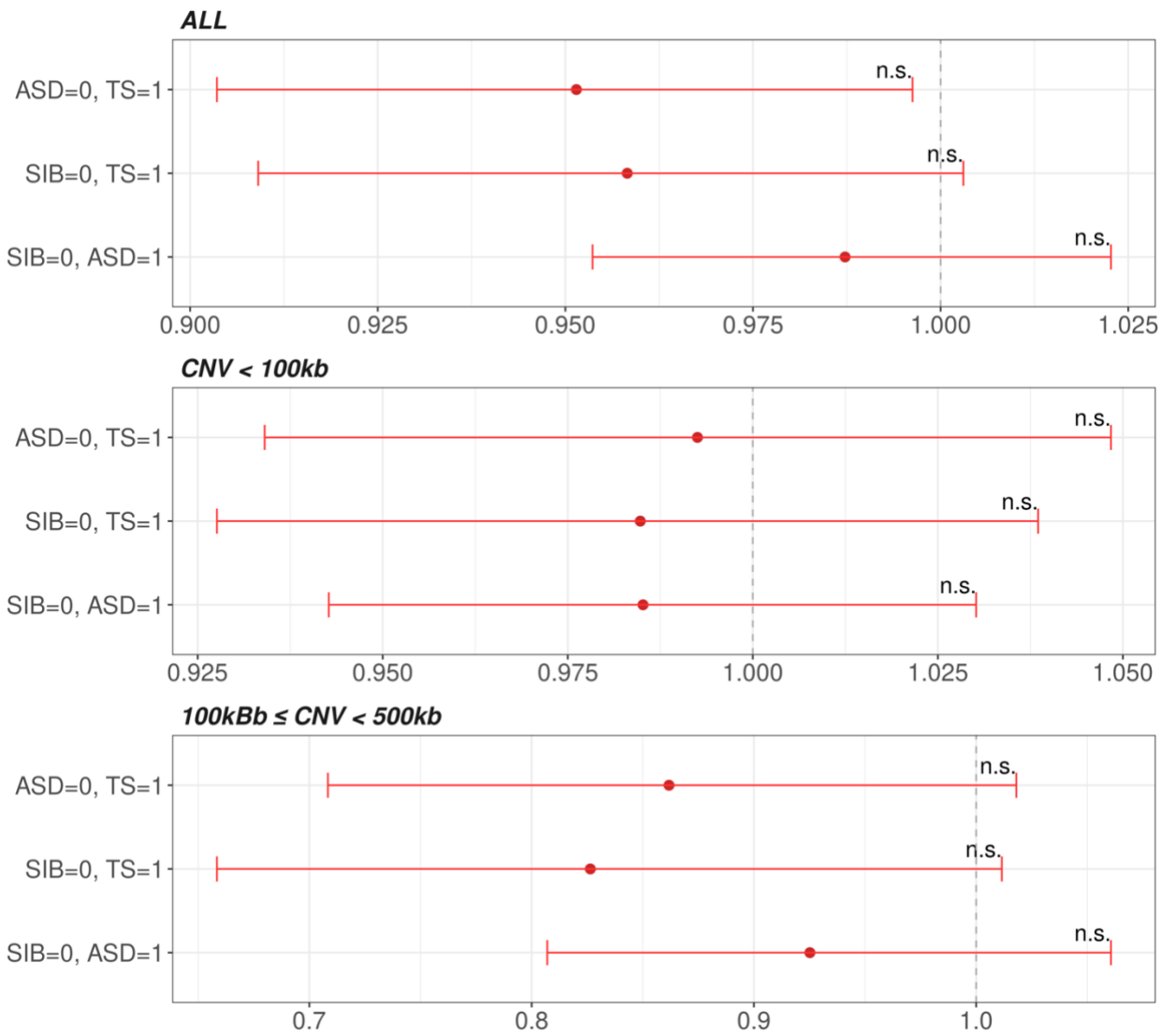

Figure 6-23. OR plots of rare genic CNV duplication count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

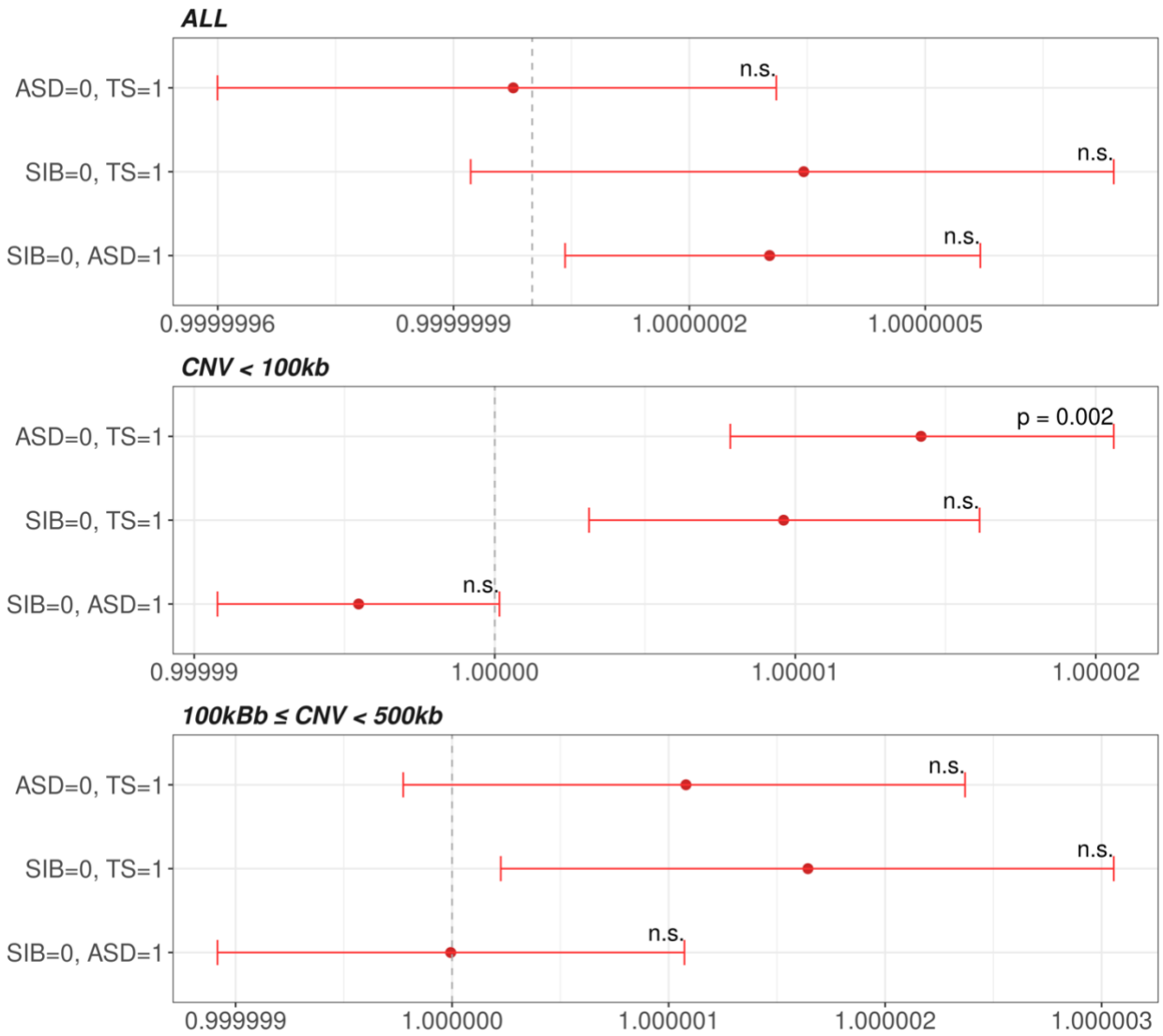

Figure 6-24. OR plots of average rare genic CNV duplication size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

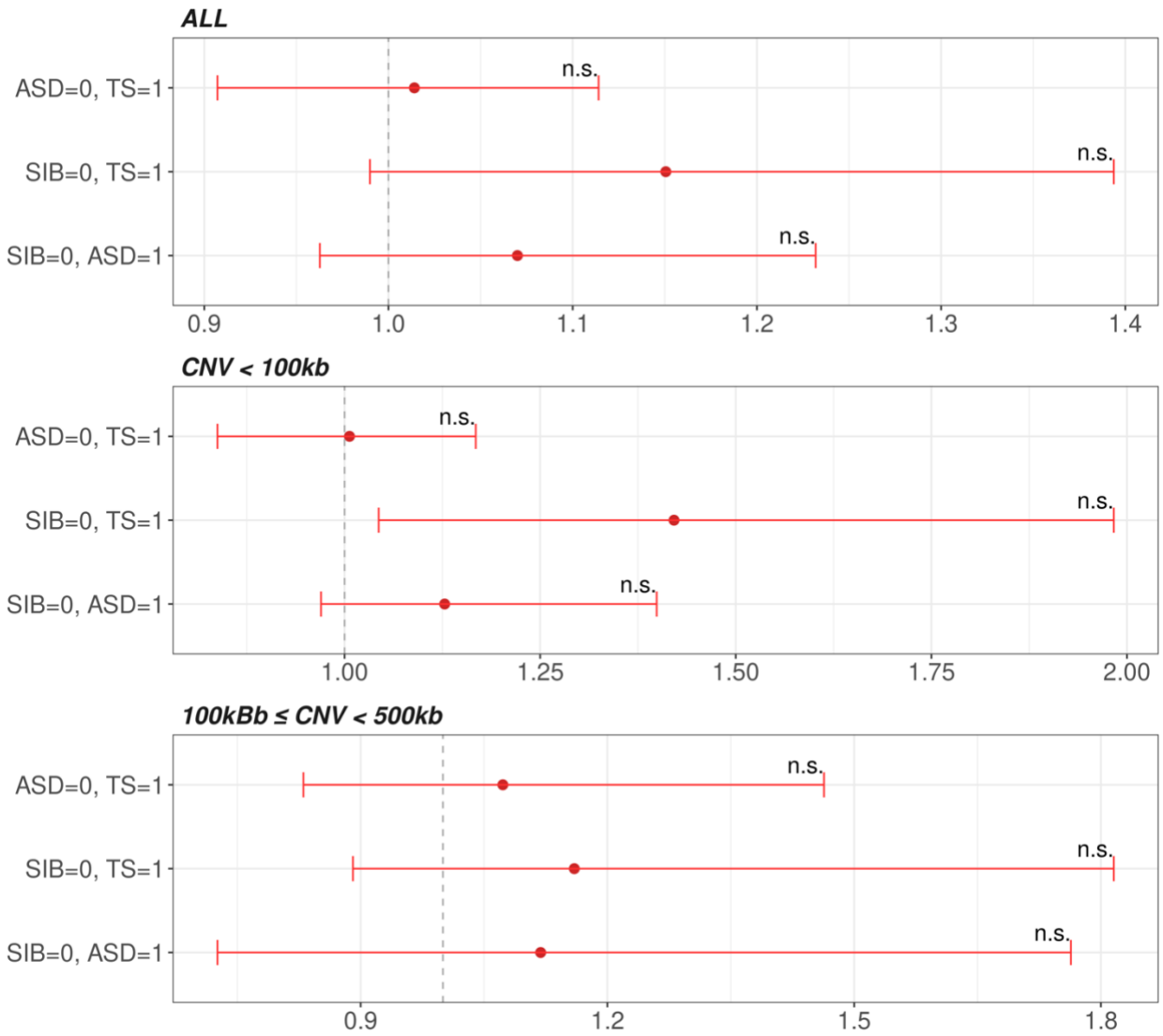

Figure 6-25. OR plots of *de novo* rare genic CNV duplication count burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

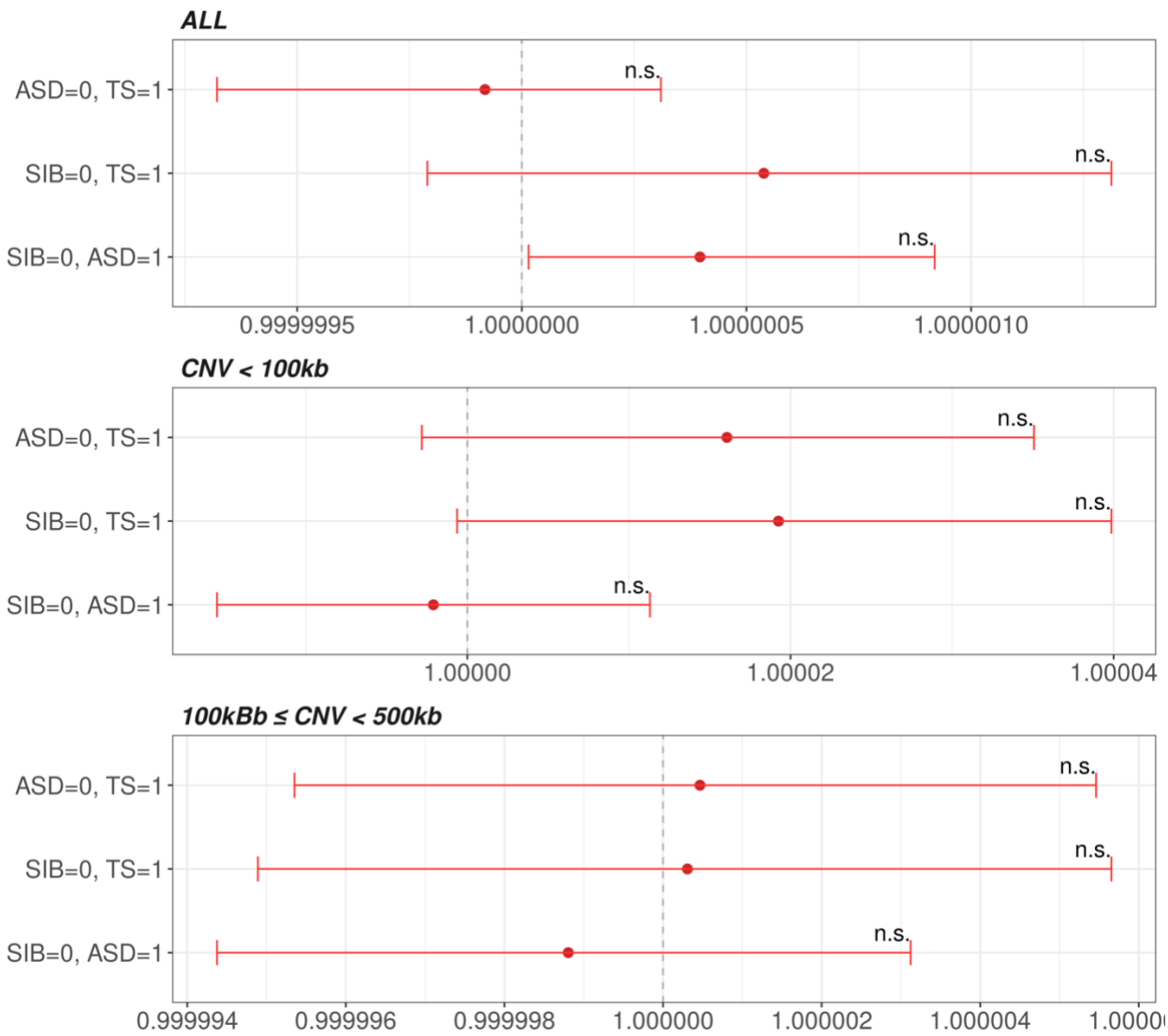

Figure 6-26. OR plots of average *de novo* rare genic CNV duplication size burden. Adjusted for LRR_{SD}, sex, and 4 PCs. Stratified by binned sizes.

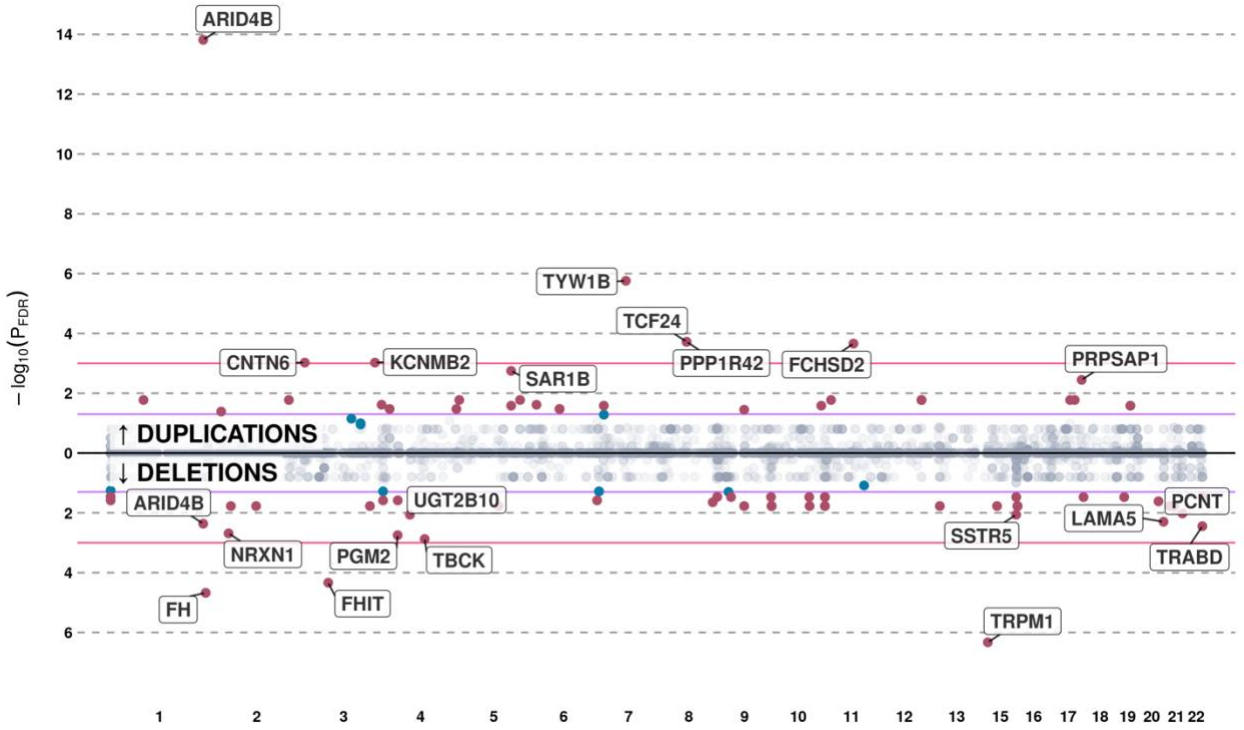

Figure 6-27. Miami plot of rare CNV gene-association tests between TS and unaffected siblings of ASD probands. $-\log_{10}(P_{FDR})$ values are shown for deletions (below 0) and duplications (above 0). Associations are ordered by genomic position (x-axis), colored by strength of association (grey – $p \geq 0.05$, blue – $P_{FDR} \geq 0.05$, and red – $P_{FDR} < 0.05$). Genes associating with $P_{FDR} < 0.01$ are labeled directly on the plot.

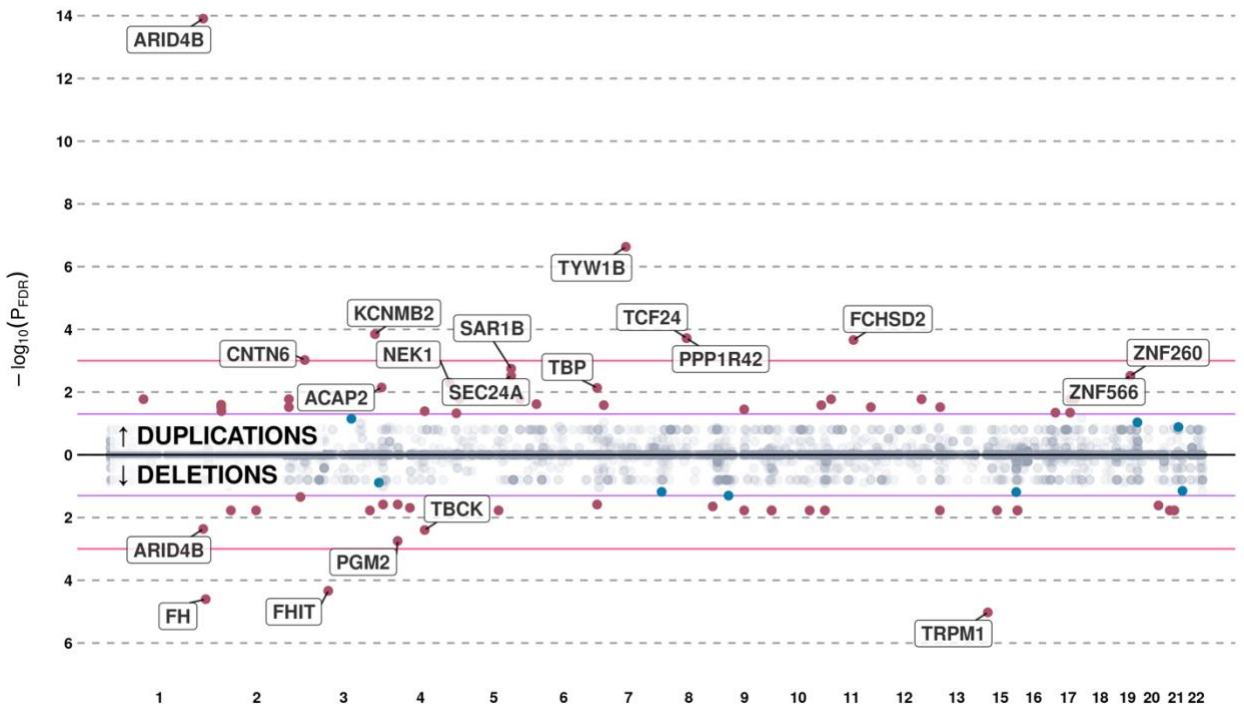

Figure 6-28. Miami plot of rare CNV gene-association tests between TS and ASD probands. -$\log_{10}(P_{FDR})$ values are shown for deletions (below 0) and duplications (above 0). Associations are ordered by genomic position (x-axis), colored by strength of association (grey – p ≥ 0.05, blue – P_{FDR} ≥ 0.05, and red – P_{FDR} <0.05). Genes associating with P_{FDR} < 0.01 are labeled directly on the plot.

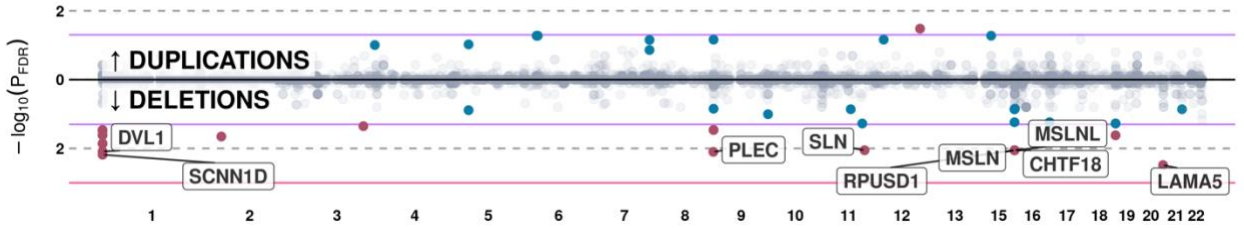

Figure 6-29. Miami plot of rare CNV gene-association tests between ASD probands and their unaffected siblings. -$\log_{10}(P_{FDR})$ values are shown for deletions (below 0) and duplications (above 0). Associations are ordered by genomic position (x-axis), colored by strength of association (grey – p ≥ 0.05, blue – P_{FDR} ≥ 0.05, and red – P_{FDR} <0.05). Genes associating with P_{FDR} < 0.01 are labeled directly on the plot.

Table 6-7. Summary of significant genic associations, $p_{FDR} < 0.01$.

Gene	Type	N_{TS}	N_{ASD}	N_{SIB}	OR_{TS-ASD}	OR_{TS-SIB}
ACAP2	Duplication	7 (0.57%)	2 (0.07%)	3 (0.11%)	**8.64	*4.15
ARID4B	Deletion	7 (0.57%)	2 (0.07%)	2 (0.07%)	**8.64	**8.30
ARID4B	Duplication	33 (2.60%)	2 (0.07%)	2 (0.07%)	****36.71	****35.26
CNTN6	Duplication	12 (0.96%)	3 (0.11%)	4 (0.14%)	***6.48	***6.22
FCHSD2	Duplication	18 (1.43%)	8 (0.29%)	8 (0.29%)	***4.86	***4.67
FH	Deletion	25 (1.97%)	13 (0.47%)	12 (0.43%)	****4.01	****4.50
FHIT	Deletion	23 (1.82%)	12 (0.43%)	11 (0.40%)	****4.32	****4.15
KCNMB2	Duplication	11 (0.88%)	2 (0.07%)	4 (0.14%)	***12.96	***6.22
LAMA5	Deletion	10 (0.80%)	19 (0.69%)	4 (0.14%)	1.08	**5.19
NEK1	Duplication	16 (1.27%)	10 (0.36%)	18 (0.65%)	**3.46	1.84
NRXN1	Deletion	14 (1.11%)	18 (0.65%)	6 (0.22%)	1.68	**4.84
PCNT	Deletion	11 (0.88%)	12 (0.43%)	6 (0.22%)	2.16	**4.15
PGM2	Deletion	5 (0.41%)	0 (0.00%)	0 (0.00%)	**Inf	**Inf
PPP1R42	Duplication	7 (0.57%)	0 (0.00%)	0 (0.00%)	***Inf	***Inf
PRPSAP1	Duplication	6 (0.49%)	3 (0.11%)	0 (0.00%)	3.24	**Inf
SAR1B	Duplication	6 (0.49%)	0 (0.00%)	0 (0.00%)	**Inf	**Inf
SEC24A	Duplication	6 (0.49%)	0 (0.00%)	2 (0.07%)	**Inf	*6.22
SSTR5	Deletion	8 (0.64%)	5 (0.18%)	1 (0.03%)	2.88	**8.30
TBCK	Deletion	9 (0.72%)	4 (0.14%)	2 (0.07%)	**5.40	**10.37
TBP	Duplication	11 (0.88%)	6 (0.22%)	11 (0.40%)	**4.32	2.08
TCF24	Duplication	7 (0.57%)	0 (0.00%)	0 (0.00%)	***Inf	***Inf
TRABD	Deletion	5 (0.41%)	3 (0.11%)	0 (0.00%)	3.24	**Inf
TRPM1	Deletion	13 (1.03%)	2 (0.07%)	0 (0.00%)	****15.12	****Inf
TYW1B	Duplication	23 (1.82%)	6 (0.22%)	7 (0.25%)	****8.64	****6.22
UGT2B10	Deletion	7 (0.57%)	4 (0.14%)	2 (0.07%)	*4.32	**8.30
ZNF260	Duplication	5 (0.41%)	0 (0.00%)	2 (0.07%)	**Inf	*6.22
ZNF566	Duplication	5 (0.41%)	0 (0.00%)	2 (0.07%)	**Inf	*6.22

Note: * $p_{FDR} < 0.05$, ** $p_{FDR} < 0.01$, *** $p_{FDR} < 0.001$, **** $p_{FDR} < 0.0001$.

CHAPTER 7
CONCLUSIONS AND FUTURE DIRECTIONS

NDDs are psychiatric that usually manifest early in childhood, and their severity can range from transient, mild impairments with minimal effect on everyday life, to severe disorders that drastically reduce quality of life and persist well into adulthood. TDs affect about 1% of the children and adolescents, whereas OCD affects about 2.3% of the children and adolescents (Zohar,1999; Scharf et al., 2012). As discussed in Chapter 2, these traits also have high heritability estimates based on family studies, yet little of that estimated genetic variation has been explored. Moreover, as discussed in Chapter 3, there is a substantial overlap between OCD and TD, as well as other disorders occurring in the childhood. Thus, OCRDs present a significant burden to children and adolescents, yet every little is known about their underlying genetic and biological mechanisms.

In this dissertation, I attempt to better understand these disorders utilizing statistical genomic approaches focusing on genome-wide, high-throughput analyses of genetic data in children affected by these disorders. Namely, I aim to explore phenotypic relationships between OCD and related disorders, including TD, utilizing phenotype data in the ABCD Study (Chapter 4), followed by similar exploration utilizing genetic data in the ABCD Study (Chapter 5), and finally exploring structural genomic variation of TD and OCD in TAAICG and SPARK datasets (Chapter 6).

As is the case with majority of projects investigating genomics of psychiatric disorder, sample sizes are crucial for successful pinpointing of associated variants and requirements for sample sizes usually stretch into tens to hundreds thousand. Nonetheless, while our modest sample size might not be sufficient for a powered

analysis to detect specific variants, they are sufficient for analysis of aggregate effects, like PRS or global burden.

Phenotypic analysis has revealed that large datasets of self-assessed psychiatric disorder phenotypes might be unreliable, however this issue can be ameliorated by assessing said psychiatric disorder phenotypes on a longitudinal basis. Furthermore, expanding administered batteries to include symptom-level data collection, even if as global as CBCL, can help further validate these phenotypes. Specifically with OCD, consistent reports of OCD are strongly associated to consistent reporting of both obsessions and compulsions in the CBCL. Analysis has shown that compulsions tend to associate stronger, however this effect can be a result of the fact that compulsions are externalized symptoms whereas obsessions are internalized symptoms, meaning compulsions are more easily observed by parents and caregivers. Narrow definition of diagnoses has globally reduced the prevalence of psychiatric disorders and comorbidity rates, however, such reduction resulted in the patterns of prevalences and comorbidity rates like those reported in the literature. Similar effects were observed between OCD-based and TD-based phenotypic analysis. There are several potential areas of improvements, including casting a wider net and examining additional datasets available in the ABCD Study, including medical history and medications used. Additionally, alternative, non-linear modelling, like deep learning or clustering-based methods could be used to leveraged in a machine learning approach to this project.

As anticipated, my GWAS studies were underpowered to detect effects of individual variants, however some interesting aggregate effects were observed. Namely, emergence of the developmentally significant structural proteins in the GO analysis,

which have been previously identified as important functional group of proteins in OCRDs (Huang et al., 2017; Wang et al., 2018). Furthermore, PRS analysis within ABCD Study shows higher rates of predictive ability between nOCD and OCS, compared to bOCD and either nOCD or OCS, indicating that nOCD might be representative of true OCD genetic risk, whereas bOCD might be more representative of general anxiety or compulsive childhood psychopathology. PGC PRS analysis shows that nOCD tends to have better cross-disorder association, specifically with other NDD disorders compared to bnOCD, further corroborating my hypothesis that nOCD is the optimal proxy for OCD in the ABCD Study. Some potential improvements to this study include local ancestry based GWASes like Tractor approach (Atkinson et al., 2021) which could account for extensive rates of admixture present in the ABCD Study sample. Additionally, non-LMM approaches might be better for PRS and heritability analyses. While LMM are generally more powered to detect individual variants, the resulting effect size estimates are not easily integrated with PRS and heritability analysis algorithms.

CNV analysis has shown a complicated relationship between ASD and TS. I have validated previous findings of risks conferred by *NRXN1* deletions and *CNTN6* duplications. Unexpected trends in associations between TS and unaffected siblings of ASD probands, as well as ASD probands and their unaffected siblings, indicate pervasive batch effects with serious impact to power of global CNV burden tests. There are several potential reasons for existence of these batch effects, as well as different avenues to address them – all of which are under present consideration. Finally, low samples sizes, phenotype misclassification and heterogeneity, and complex genome-

phenome relationships still represent the biggest obstacles in both TS and OCD realm, thus increased recruitment, rigorous phenotyping, and methodological innovation are likely to be the main driving forces behind OCDR genetics discoveries.

LIST OF REFERENCES

23andMe. (n.d.). *Understanding Personal Genetics*. 23andMe for Medical Professionals. Retrieved May 18, 2022, from https://medical.23andme.com/

Abdulkadir, M., Londono, D., Gordon, D., Fernandez, T. V., Brown, L. W., Cheon, K.-A., Coffey, B. J., Elzerman, L., Fremer, C., Fründt, O., Garcia-Delgar, B., Gilbert, D. L., Grice, D. E., Hedderly, T., Heyman, I., Hong, H. J., Huyser, C., Ibanez-Gomez, L., Jakubovski, E., … Dietrich, A. (2017). Investigation of previously implicated genetic variants in chronic tic disorders: a transmission disequilibrium test approach. *European Archives of Psychiatry and Clinical Neuroscience*, *268(3)*, 301–316. https://doi.org/10.1007/s00406-017-0808-8

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., Rašin, M.-R., Gunel, M., Davis, N. R., Ercan-Sencicek, A. G., Guez, D. H., Spertus, J. A., Leckman, J. F., Dure, L. S., Kurlan, R., Singer, H. S., Gilbert, D. L., … State, M. W. (2005). Sequence Variants in SLITRK1 Are Associated with Tourette's Syndrome. *Science*, *310(5746)*, 317–320. https://doi.org/10.1126/science.1116502

Abramowitz, J. S., Deacon, B. J., Olatunji, B. O., Wheaton, M. G., Berman, N. C., Losardo, D., Timpano, K. R., McGrath, P. B., Riemann, B. C., Adams, T., Björgvinsson, T., Storch, E. A., & Hale, L. R. (2010). Assessment of obsessive-compulsive symptom dimensions: Development and evaluation of the Dimensional Obsessive-Compulsive Scale. *Psychological Assessment*, *22(1)*, 180–198. https://doi.org/10.1037/a0018260

Alemany-Navarro, M., Cruz, R., Real, E., Segalàs, C., Bertolín, S., Baenas, I., Domènech, L., Rabionet, R., Carracedo, Á., Menchón, J. M., & Alonso, P. (2020). Exploring genetic variants in obsessive compulsive disorder severity: A GWAS approach. *Journal of Affective Disorders*, *267*, 23–32. https://doi.org/10.1016/j.jad.2020.01.161

Alemany-Navarro, M., Cruz, R., Real, E., Segalàs, C., Bertolín, S., Rabionet, R., Carracedo, Á., Menchón, J. M., & Alonso, P. (2020). Looking into the genetic bases of OCD dimensions: a pilot genome-wide association study. *Translational Psychiatry*, *10(1)*. https://doi.org/10.1038/s41398-020-0804-z

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19(9)*, 1655–1664. https://doi.org/10.1101/gr.094052.109

Allen, B., Bresnahan, J., Childers, L., Foster, I., Kandaswamy, G., Kettimuthu, R., Kordas, J., Link, M., Martin, S., Pickett, K., & Tuecke, S. (2012). Software as a service for data scientists. *Communications of the ACM*, *55(2)*, 81–88. https://doi.org/10.1145/2076450.2076468

Altemus, M., Murphy, D. L., Greenberg, B., & Lesch, K. P. (1996). Intact Coding Region of the Serotonin Transporter Gene in Obsessive-Compulsive Disorder. *American Journal of Medical Genetics*, *67(4)*, 409–411.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). American Psychiatric Publishing.

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). American Psychiatric Publishing.

Amersham Biosciences. (2002). *Microarray Handbook*. Amersham Biosciences.

Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G. J., Gormley, P., Malik, R., Patsopoulos, N. A., Ripke, S., Wei, Z., Yu, D., Lee, P. H., Turley, P., Grenier-Boley, B., Chouraki, V., Kamatani, Y., … Neale, B. M. (2018). Analysis of shared heritability in common disorders of the brain. *Science*, *360(6395)*, eaap8757. https://doi.org/10.1126/science.aap8757

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25(1)*, 25–29. https://doi.org/10.1038/75556

Atkinson, E. G., Maihofer, A. X., Kanai, M., Martin, A. R., Karczewski, K. J., Santoro, M. L., Ulirsch, J. C., Kamatani, Y., Okada, Y., Finucane, H. K., Koenen, K. C., Nievergelt, C. M., Daly, M. J., & Neale, B. M. (2021). Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nature Genetics*, *53(2)*, 195–204. https://doi.org/10.1038/s41588-020-00766-y

Barr, C. L., Wigg, K. G., & Sandor, P. (1999). Catechol-O-methyltransferase and Gilles de la Tourette syndrome. *Molecular Psychiatry*, *4(5)*, 492–495. https://doi.org/10.1038/sj.mp.4000549

Barr, C. L., Wigg, K. G., Zovko, E., Sandor, & Tsui, L. C. (1997). Linkage study of the dopamine D5 receptor gene and Gilles de la Tourette syndrome . *American Journal of Human Genetics*, *74(1)*, 58–61.

Baurley, J. W., Edlund, C. K., Pardamean, C. I., Conti, D. V., & Bergen, A. W. (2016). Smokescreen: a targeted genotyping array for addiction research. *BMC Genomics*, *17(1)*, 145. https://doi.org/10.1186/s12864-016-2495-7

Belloso, J. M., Bache, I., Guitart, M., Caballin, M. R., Halgren, C., Kirchhoff, M., Ropers, H.-H., Tommerup, N., & Tümer, Z. (2007). Disruption of the CNTNAP2 gene in a t(7;15) translocation family without symptoms of Gilles de la Tourette syndrome. *European Journal of Human Genetics*, *15(6)*, 711–713. https://doi.org/10.1038/sj.ejhg.5201824

Bertelsen, B., Melchior, L., Jensen, L. R., Groth, C., Glenthøj, B., Rizzo, R., Debes, N. M., Skov, L., Brøndum-Nielsen, K., Paschou, P., Silahtaroglu, A., & Tümer, Z. (2014). Intragenic deletions affecting two alternative transcripts of the IMMP2L gene in patients with Tourette syndrome. *European Journal of Human Genetics*, *22(11)*, 1283–1289. https://doi.org/10.1038/ejhg.2014.24

Bertelsen, B., Stefánsson, H., Riff Jensen, L., Melchior, L., Mol Debes, N., Groth, C., Skov, L., Werge, T., Karagiannidis, I., Tarnok, Z., Barta, C., Nagy, P., Farkas, L., Brøndum-Nielsen, K., Rizzo, R., Gulisano, M., Rujescu, D., Kiemeney, L. A., Tosato, S., … Tümer, Z. (2016). Association of AADAC Deletion and Gilles de la Tourette Syndrome in a Large European Cohort. *Biological Psychiatry*, *79(5)*, 383–391. https://doi.org/10.1016/j.biopsych.2015.08.027

Billett, E. A., Richter, M. A., King, N., Heils, A., Lesch, K. P., & Kennedy, J. L. (1997). Obsessive compulsive disorder, response to serotonin reuptake inhibitors and the serotonin transporter gene. *Molecular Psychiatry*, *2(5)*, 403–406. https://doi.org/10.1038/sj.mp.4000257

Billett, E. A., Richter, M. A., Sam, F., Swinson, R. P., Dai, X.-Y., King, N., Badri, F., Sasaki, T., Buchanan, J. A., & Kennedy, J. L. (1998). Investigation of dopamine system genes in obsessive-compulsive disorder. *Psychiatric Genetics*, *8(3)*, 163–170. https://doi.org/10.1097/00041444-199800830-00005

Bjork, J. M., Straub, L. K., Provost, R. G., & Neale, M. C. (2017). The ABCD Study of Neurodevelopment: Identifying Neurocircuit Targets for Prevention and Treatment of Adolescent Substance Abuse. *Current Treatment Options in Psychiatry*, *4(2)*, 196–209. https://doi.org/10.1007/s40501-017-0108-y

Bolton, D., Rijsdijk, F., O'Connor, T. G., Perrin, S., & Eley, T. C. (2006). Obsessive–compulsive disorder, tics and anxiety in 6-year-old twins. *Psychological Medicine*, *37(1)*, 39–48. https://doi.org/10.1017/s0033291706008816

Brett, P M, Curtis, D., Robertson, M. M., & Gurling, H. M. (1995). Exclusion of the 5-HT1A serotonin neuroreceptor and tryptophan oxygenase genes in a large British kindred multiply affected with Tourette's syndrome, chronic motor tics, and obsessive-compulsive behavior. *American Journal of Psychiatry*, *152(3)*, 437–440. https://doi.org/10.1176/ajp.152.3.437

Brett, Peter M., Curtis, D., Robertson, M. M., & Gurling, H. M. D. (1995). The genetic susceptibility to Gilles de la Tourette Syndrome in a large multiple affected british kindred: Linkage analysis excludes a role for the genes coding for dopamine D1, D2, D3, D4, D5 receptors, dopamine beta hydroxylase, tyrosinase, and tyrosine hydroxylase. *Biological Psychiatry*, *37(8)*, 533–540. https://doi.org/10.1016/0006-3223(94)00161-u

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., & Grice, D. E. (2015).

Familial Clustering of Tic Disorders and Obsessive-Compulsive Disorder. *JAMA Psychiatry*, *72(4)*, 359. https://doi.org/10.1001/jamapsychiatry.2014.2656

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., Daly, M. J., Price, A. L., & Neale, B. M. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, *47(11)*, 1236–1241. https://doi.org/10.1038/ng.3406

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., & Neale, B. M. (2015a). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47(3)*, 291–295. https://doi.org/10.1038/ng.3211

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., & Neale, B. M. (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, *47(3)*, 291–295. https://doi.org/10.1038/ng.3211

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2018). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47(D1)*, D1005–D1012. https://doi.org/10.1093/nar/gky1120

Burton, C. L., Lemire, M., Xiao, B., Corfield, E. C., Erdman, L., Bralten, J., Poelmans, G., Yu, D., Shaheen, S.-M., Goodale, T., Sinopoli, V. M., Soreni, N., Hanna, G. L., Fitzgerald, K. D., Rosenberg, D., Nestadt, G., Paterson, A. D., Strug, L. J., Schachar, R. J., … Zai, G. (2021). Genome-wide association study of pediatric obsessive-compulsive traits: shared genetic risk between traits and disorder. *Translational Psychiatry*, *11(1)*. https://doi.org/10.1038/s41398-020-01121-9

Cappi, C, Brentani, H., Lima, L., Sanders, S. J., Zai, G., Diniz, B. J., Reis, V. N. S., Hounie, A. G., Conceição do Rosário, M., Mariani, D., Requena, G. L., Puga, R., Souza-Duran, F. L., Shavitt, R. G., Pauls, D. L., Miguel, E. C., & Fernandez, T. V. (2016). Whole-exome sequencing in obsessive-compulsive disorder identifies rare mutations in immunological and neurodevelopmental pathways. *Translational Psychiatry*, *6(3)*, e764–e764. https://doi.org/10.1038/tp.2016.30

Cappi, Carolina, Oliphant, M. E., Péter, Z., Zai, G., Conceição do Rosário, M., Sullivan, C. A. W., Gupta, A. R., Hoffman, E. J., Virdee, M., Olfson, E., Abdallah, S. B., Willsey, A. J., Shavitt, R. G., Miguel, E. C., Kennedy, J. L., Richter, M. A., & Fernandez, T. V. (2020). De Novo Damaging DNA Coding Mutations Are Associated With Obsessive-Compulsive Disorder and Overlap With Tourette's

Disorder and Autism. *Biological Psychiatry*, *87(12)*, 1035–1044. https://doi.org/10.1016/j.biopsych.2019.09.029

Carey, G., & Gottesman, I. I. (1981). Twin and family studies of anxiety, phobic and obsessive disorder. In D. Klein & J. Rabkin, *Anxiety: New Research and Changing Concepts* (pp. 117–136). Raven Press.

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*, *39(S7)*, S16–S21. https://doi.org/10.1038/ng2028

Cavallini, M. C., Di Bella, D., Catalano, M., & Bellodi, L. (2000). An association study between 5-HTTLPR polymorphism, COMT polymorphism, and Tourette's syndrome. *Psychiatry Research*, *97(2-3)*, 93–100. https://doi.org/10.1016/s0165-1781(00)00220-1

Chabane, N., Delorme, R., Millet, B., Mouren, M.-C., Leboyer, M., & Pauls, D. (2005). Early-onset obsessive-compulsive disorder: a subgroup with a specific clinical and familial pattern? *Journal of Child Psychology and Psychiatry*, *46(8)*, 881–887. https://doi.org/10.1111/j.1469-7610.2004.00382.x

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4(1)*. https://doi.org/10.1186/s13742-015-0047-8

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., Szpiro, A. A., Chen, W., Brehm, J. M., Celedón, J. C., Redline, S., Papanicolaou, G. J., Thornton, T. A., Laurie, C. C., Rice, K., & Lin, X. (2016). Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *The American Journal of Human Genetics*, *98(4)*, 653–666. https://doi.org/10.1016/j.ajhg.2016.02.012

Chen, Z., Boehnke, M., Wen, X., & Mukherjee, B. (2021). Revisiting the genome-wide significance threshold for common variant GWAS. *G3 Genes|Genomes|Genetics*, *11(2)*. https://doi.org/10.1093/g3journal/jkaa056

Choi, S. W., Mak, T. S.-H., & O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*, *15(9)*, 2759–2772. https://doi.org/10.1038/s41596-020-0353-1

Chou, I.-C., Tsai, C.-H., Wan, L., Hsu, Y.-A., & Tsai, F.-J. (2007). Association study between Tourette's syndrome and polymorphisms of noradrenergic genes (ADRA2A, ADRA2C). *Psychiatric Genetics*, *17(6)*, 359. https://doi.org/10.1097/ypg.0b013e3281ac2358

Claudio-Campos, K., Stevens, D., Koo, S., Valko, A., Bienvenu, O. J., Budman, C. B., Cath, D. C., Darrow, S., Geller, D., Goes, F. S., Grados, M. A., Greenberg, B. D., Greenberg, E., Hirschtritt, M. E., Illmann, C., Ivankovic, F., King, R. A., Knowles,

J. A., Krasnow, J., … Mathews, C. A. (2021). Is Persistent Motor or Vocal Tic Disorder a Milder Form of Tourette Syndrome? *Movement Disorders*, *36(8)*, 1899–1910. https://doi.org/10.1002/mds.28593

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., Bassett, A. S., Seller, A., Holmes, C. C., & Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, *35(6)*, 2013–2025. https://doi.org/10.1093/nar/gkm076

Comings, D. E., Muhleman, D., Dietz, G., Dino, M., LeGro, R., & Gade, R. (1993). Association between Tourette's syndrome and homozygosity at the dopamine D3 receptor gene. *The Lancet*, *341(906)*.

Conomos, M. P., Miller, M. B., & Thornton, T. A. (2015). Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genetic Epidemiology*, *39(4)*, 276–293. https://doi.org/10.1002/gepi.21896

Conomos, M. P., Reiner, A. P., Weir, B. S., & Thornton, T. A. (2016). Model-free Estimation of Recent Genetic Relatedness. *The American Journal of Human Genetics*, *98(1)*, 127–148. https://doi.org/10.1016/j.ajhg.2015.11.022

Coombes, B. J., Ploner, A., Bergen, S. E., & Biernacka, J. M. (2020). A principal component approach to improve association testing with polygenic risk scores. *Genetic Epidemiology*, *44(7)*, 676–686. https://doi.org/10.1002/gepi.22339

Coughlin, C. R., Scharer, G. H., & Shaikh, T. H. (2012). Clinical impact of copy number variation analysis using high-resolution microarray technologies: advantages, limitations and concerns. *Genome Medicine*, *4(10)*, 80. https://doi.org/10.1186/gm381

Crawford, F. C., Ait-Ghezala, G., Morris, M., Sutcliffe, M. J., Hauser, R. A., Silver, A. A., & Mullan, M. J. (2003). Translocation breakpoint in two unrelated Tourette syndrome cases, within a region previously linked to the disorder. *Human Genetics*, *113(2)*, 154–161. https://doi.org/10.1007/s00439-003-0942-4

Cuker, A., State, M. W., King, R. A., Davis, N., & Ward, D. C. (2004). Candidate locus for Gilles de la Tourette syndrome/obsessive compulsive disorder/chronic tic disorder at 18q22. *American Journal of Medical Genetics*, *30A(1)*, 37–39. https://doi.org/10.1002/ajmg.a.30066

Cukier, H. N., Dueker, N. D., Slifer, S. H., Lee, J. M., Whitehead, P. L., Lalanne, E., Leyva, N., Konidari, I., Gentry, R. C., Hulme, W. F., Booven, D. V., Mayo, V., Hofmann, N. K., Schmidt, M. A., Martin, E. R., Haines, J. L., Cuccaro, M. L., Gilbert, J. R., & Pericak-Vance, M. A. (2014). Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and

neuropsychiatric disorders. *Molecular Autism*, *5(1)*. https://doi.org/10.1186/2040-2392-5-1

Darrow, S. M., Hirschtritt, M. E., Davis, L. K., Illmann, C., Osiecki, L., Grados, M., Sandor, P., Dion, Y., King, R., Pauls, D., Budman, C. L., Cath, D. C., Greenberg, E., Lyon, G. J., Yu, D., McGrath, L. M., McMahon, W. M., Lee, P. C., Delucchi, K. L., … Mathews, C. A. (2017). Identification of Two Heritable Cross-Disorder Endophenotypes for Tourette Syndrome. *American Journal of Psychiatry*, *174(4)*, 387–396. https://doi.org/10.1176/appi.ajp.2016.16020240

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., Berlin, C. M., Malaty, I. A., Woods, D. W., Scharf, J. M., & Mathews, C. A. (2015). Web-based phenotyping for Tourette Syndrome: Reliability of common co-morbid diagnoses. *Psychiatry Research*, *228(3)*, 816–825. https://doi.org/10.1016/j.psychres.2015.05.017

Davey Smith, G., & Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International Journal of Epidemiology*, *32(1)*, 1–22. https://doi.org/10.1093/ije/dyg070

de Bakker, P. I. W., Ferreira, M. A. R., Jia, X., Neale, B. M., Raychaudhuri, S., & Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human Molecular Genetics*, *17(R2)*, R122–R128. https://doi.org/10.1093/hmg/ddn288

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, *10(1)*, 5436. https://doi.org/10.1038/s41467-019-13225-y

Demontis, D., Walters, R. K., Martin, J., Mattheisen, M., Als, T. D., Agerbo, E., Baldursson, G., Belliveau, R., Bybjerg-Grauholm, J., Bækvad-Hansen, M., Cerrato, F., Chambert, K., Churchhouse, C., Dumont, A., Eriksson, N., Gandal, M., Goldstein, J. I., Grasby, K. L., Grove, J., … Neale, B. M. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature Genetics*, *51*, 63–75. https://doi.org/10.1038/s41588-018-0269-7

Dey, R., & Lee, S. (2019). Asymptotic properties of principal component analysis and shrinkage-bias adjustment under the generalized spiked population model. *Journal of Multivariate Analysis*, *173*, 145–164. https://doi.org/10.1016/j.jmva.2019.02.007

Díaz-Anzaldúa, A, Joober, R., Rivière, J.-B., Dion, Y., Lespérance, P., Richer, F., Chouinard, S., & Rouleau, G. A. (2004). Tourette syndrome and dopaminergic genes: a family-based association study in the French Canadian founder population. *Molecular Psychiatry*, *9(3)*, 272–277. https://doi.org/10.1038/sj.mp.4001411

Díaz-Anzaldúa, Adriana, Rivière, J.-B., Dubé, M.-P., Joober, R., Saint-Onge, J., Dion, Y., Lespérance, P., Richer, F., Chouinard, S., & Rouleau, G. A. (2005). Chromosome 11-q24 region in Tourette syndrome: Association and linkage disequilibrium study in the French Canadian population. *American Journal of Medical Genetics Part A*, *138A(3)*, 225–228. https://doi.org/10.1002/ajmg.a.30928

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M., & Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Research*, *36(19)*, e126–e126. https://doi.org/10.1093/nar/gkn556

do Rosario-Campos, M. C., Leckman, J. F., Curi, M., Quatrano, S., Katsovitch, L., Miguel, E. C., & Pauls, D. L. (2005). A family study of early-onset obsessive-compulsive disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *136B(1)*, 92–97. https://doi.org/10.1002/ajmg.b.30149

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, *21(16)*, 3439–3440. https://doi.org/10.1093/bioinformatics/bti525

Durinck, Steffen, Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4(8)*, 1184–1191. https://doi.org/10.1038/nprot.2009.97

Eapen, V., Pauls, D. L., & Robertson, M. M. (1993). Evidence for Autosomal Dominant Transmission in Tourette's Syndrome. *British Journal of Psychiatry*, *162(5)*, 593–596. https://doi.org/10.1192/bjp.162.5.593

Eley, T. C., Bolton, D., O'Connor, T. G., Perrin, S., Smith, P., & Plomin, R. (2003). A twin study of anxiety-related behaviours in pre-school children. *Journal of Child Psychology and Psychiatry*, *44(7)*, 945–960. https://doi.org/10.1111/1469-7610.00179

Ercan-Sencicek, A. G., Stillman, A. A., Ghosh, A. K., Bilguvar, K., O'Roak, B. J., Mason, C. E., Abbott, T., Gupta, A., King, R. A., Pauls, D. L., Tischfield, J. A., Heiman, G. A., Singer, H. S., Gilbert, D. L., Hoekstra, P. J., Morgan, T. M., Loring, E., Yasuno, K., Fernandez, T., … State, M. W. (2010). L-Histidine Decarboxylase and Tourette's Syndrome. *New England Journal of Medicine*, *362(20)*, 1901–1908. https://doi.org/10.1056/nejmoa0907006

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2014). PRSice: Polygenic Risk Score software. *Bioinformatics*, *31(9)*, 1466–1468. https://doi.org/10.1093/bioinformatics/btu848

Evans, L. M., Tahmasbi, R., Vrieze, S. I., Abecasis, G. R., Das, S., Gazal, S., Bjelland, D. W., de Candia, T. R., Goddard, M. E., Neale, B. M., Yang, J., Visscher, P. M.,

& Keller, M. C. (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics*, *50(5)*, 737–745. https://doi.org/10.1038/s41588-018-0108-x

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., Bilguvar, K., Glessner, J., Chu, S. H., Leckman, J. F., King, R. A., Gilbert, D. L., Heiman, G. A., Tischfield, J. A., Hoekstra, P. J., … State, M. W. (2012). Rare Copy Number Variants in Tourette Syndrome Disrupt Genes in Histaminergic Pathways and Overlap with Autism. *Biological Psychiatry*, *71(5)*, 392–402. https://doi.org/10.1016/j.biopsych.2011.09.034

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The Obsessive-Compulsive Inventory: Development and validation of a short version. *Psychological Assessment*, *14(4)*, 485–496. https://doi.org/10.1037/1040-3590.14.4.485

Foa, E. B., Kozak, M. J., Salkovskis, P. M., Coles, M. E., & Amir, N. (1998). The validation of a new obsessive–compulsive disorder scale: The Obsessive–Compulsive Inventory. *Psychological Assessment*, *10(3)*, 206–214. https://doi.org/10.1037/1040-3590.10.3.206

Forstner, A. J., Awasthi, S., Wolf, C., Maron, E., Erhardt, A., Czamara, D., Eriksson, E., Lavebratt, C., Allgulander, C., Friedrich, N., Becker, J., Hecker, J., Rambau, S., Conrad, R., Geiser, F., McMahon, F. J., Moebus, S., Hess, T., Buerfent, B. C., … Schumacher, J. (2021). Genome-wide association study of panic disorder reveals genetic overlap with neuroticism and depression. *Molecular Psychiatry*, *26(8)*, 4179–4190. https://doi.org/10.1038/s41380-019-0590-2

Foster, I. (2011). Globus Online: Accelerating and Democratizing Science through Cloud-Based Services. *IEEE Internet Computing*, *15(3)*, 70–73. https://doi.org/10.1109/mic.2011.64

Freedman, R., Lewis, D. A., Michels, R., Pine, D. S., Schultz, S. K., Tamminga, C. A., Gabbard, G. O., Gau, S. S.-F., Javitt, D. C., Oquendo, M. A., Shrout, P. E., Vieta, E., & Yager, J. (2013). The Initial Field Trials of DSM-5: New Blooms and Old Thorns. *American Journal of Psychiatry*, *170(1)*, 1–5. https://doi.org/10.1176/appi.ajp.2012.12091189

Friel, P. B. (1973). Familial Incidence of Gilles de la Tourette's Disease, with Observations on Aetiology and Treatment. *British Journal of Psychiatry*, *122(571)*, 655–658. https://doi.org/10.1192/bjp.122.6.655

Friis, R. H., & Sellers, T. (2020). *Epidemiology for Public Health Practice* (6th ed.). Jones & Bartlett Learning.

Gazzellone, M. J., Zarrei, M., Burton, C. L., Walker, S., Uddin, M., Shaheen, S. M., Coste, J., Rajendram, R., Schachter, R. J., Colasanto, M., Hanna, G. L.,

Rosenberg, D. R., Soreni, N., Fitzgerald, K. D., Marshall, C. R., Buchanan, J. A., Merico, D., Arnold, P. D., & Scherer, S. W. (2016). Uncovering obsessive-compulsive disorder risk genes in a pediatric cohort by high-resolution analysis of copy number variation. *Journal of Neurodevelopmental Disorders*, *8(1)*. https://doi.org/10.1186/s11689-016-9170-9

Gelernter, J., Kennedy, J. L., Grandy, D. K., Zhou, Q. Y., Civelli, Pauls, D. L., Pakstis, Kurlan, R., Sunahara, R. K., Niznik, H. B., O'Dowd, B., Seeman, P., & Kidd, K. K. (1993). Exclusion of close linkage of Tourette's syndrome to D1 dopamine receptor. *American Journal of Psychiatry*, *150(3)*, 449–453. https://doi.org/10.1176/ajp.150.3.449

Gene Ontology Consortium, Carbon, S., Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L.-P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., … Elser, J. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, *49(D1)*, D325–D334. https://doi.org/10.1093/nar/gkaa1113

GeneCards (n.d.). RBFOX1 Gene Disorders. GeneCards. Retrieved June 30, 2022, from https://www.genecards.org/cgi-bin/carddisp.pl?gene=RBFOX1#diseases

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. (2007). *Nature*, *447(7145)*, 661–678. https://doi.org/10.1038/nature05911

Glessner, J. T., Hou, X., Zhong, C., Zhang, J., Khan, M., Brand, F., Krawitz, P., Sleiman, P. M. A., Hakonarson, H., & Wei, Z. (2021). DeepCNV: a deep learning approach for authenticating copy number variations. *Briefings in Bioinformatics*, *22(5)*. https://doi.org/10.1093/bib/bbaa381

Gogarten, S. M., Bhangale, T., Conomos, M. P., Laurie, C. A., McHugh, C. P., Painter, I., Zheng, X., Crosslin, D. R., Levine, D., Lumley, T., Nelson, S. C., Rice, K., Shen, J., Swarnkar, R., Weir, B. S., & Laurie, C. C. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, *28(24)*, 3329–3331. https://doi.org/10.1093/bioinformatics/bts610

Gogarten, Stephanie M, Sofer, T., Chen, H., Yu, C., Brody, J. A., Thornton, T. A., Rice, K. M., & Conomos, M. P. (2019). Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*, *35(24)*, 5346–5348. https://doi.org/10.1093/bioinformatics/btz567

Goodman, W. K. (1989). The Yale-Brown Obsessive Compulsive Scale. *Archives of General Psychiatry*, *46(11)*, 1006. https://doi.org/10.1001/archpsyc.1989.01810110048007

Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O. A., Anney, R., Awashti, S., Belliveau, R., Bettella, F., Buxbaum, J. D., Bybjerg-Grauholm, J., Bækvad-Hansen, M., Cerrato, F., Chambert, K., Christensen, J. H., … Børglum, A. D. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics*, *51(3)*, 431–444. https://doi.org/10.1038/s41588-019-0344-8

Grünblatt, E., Oneda, B., Ekici, A. B., Ball, J., Geissler, J., Uebe, S., Romanos, M., Rauch, A., & Walitza, S. (2017). High resolution chromosomal microarray analysis in paediatric obsessive-compulsive disorder. *BMC Medical Genomics*, *10(1)*. https://doi.org/10.1186/s12920-017-0299-5

Guo, W., Samuels, J. F., Wang, Y., Cao, H., Ritter, M., Nestadt, P. S., Krasnow, J., Greenberg, B. D., Fyer, A. J., McCracken, J. T., Geller, D. A., Murphy, D. L., Knowles, J. A., Grados, M. A., Riddle, M. A., Rasmussen, S. A., McLaughlin, N. C., Nurmi, E. L., Askland, K. D., … Shugart, Y. Y. (2017). Polygenic risk score and heritability estimates reveals a genetic relationship between ASD and OCD. *European Neuropsychopharmacology*, *27(7)*, 657–666. https://doi.org/10.1016/j.euroneuro.2017.03.011

Hamza, T. H., Zabetian, C. P., Tenesa, A., Laederach, A., Montimurro, J., Yearout, D., Kay, D. M., Doheny, K. F., Paschall, J., Pugh, E., Kusel, V. I., Collura, R., Roberts, J., Griffith, A., Samii, A., Scott, W. K., Nutt, J., Factor, S. A., & Payami, H. (2010). Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nature Genetics*, *42(9)*, 781–785. https://doi.org/10.1038/ng.642

Hanna, G. L., Fingerlin, T. E., Himle, J. A., & Boehnke, M. (2005). Complex Segregation Analysis of Obsessive-Compulsive Disorder in Families with Pediatric Probands. *Human Heredity*, *60(1)*, 1–9. https://doi.org/10.1159/000087135

Hanna, G. L., Veenstra-VanderWeele, J., Cox, N. J., Boehnke, M., Himle, J. A., Curtis, G. C., Leventhal, B. L., & Cook, E. H. (2002). Genome-wide linkage analysis of families with obsessive-compulsive disorder ascertained through pediatric probands. *American Journal of Medical Genetics*, *114(5)*, 541–552. https://doi.org/10.1002/ajmg.10519

Hasstedt, S. J., Leppert, M., Filloux, F., van de Wetering, B. J., & McMahon, W. M. (1995). Intermediate inheritance of Tourette syndrome, assuming assortative mating. *American Journal of Human Genetics*, *57(3)*, 682–689.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, *10(8)*, 551–564. https://doi.org/10.1038/nrg2593

He, F., Zheng, Y., Huang, H.-H., Cheng, Y.-H., & Wang, C.-Y. (2015). Association between Tourette Syndrome and the Dopamine D3 Receptor Gene Rs6280.

Chinese Medical Journal, *128(5)*, 654–658. https://doi.org/10.4103/0366-6999.151665

He, X., Sanders, S. J., Liu, L., De Rubeis, S., Lim, E. T., Sutcliffe, J. S., Schellenberg, G. D., Gibbs, R. A., Daly, M. J., Buxbaum, J. D., State, M. W., Devlin, B., & Roeder, K. (2013). Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. *PLoS Genetics*, *9(8)*, e1003671. https://doi.org/10.1371/journal.pgen.1003671

Hebebrand, J., Nöthen, M., Lehmkuhl, G., Poustka, F., Schmidt, M., Propping, P., Remschmidt, H., Comings, D., Muhleman, D., Dietz, G., Dino, M., Legro, R., & Gade, R. (1993). Tourette's syndrome and homozygosity for the dopamine D3 receptor gene. *The Lancet*, *341(8858)*, 1483–1484. https://doi.org/10.1016/0140-6736(93)90931-6

Hill, T., & Unckless, R. L. (2019). A Deep Learning Approach for Detecting Copy Number Variation in Next-Generation Sequencing Data. *G3 Genes|Genomes|Genetics*, *9(11)*, 3575–3582. https://doi.org/10.1534/g3.119.400596

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., Cath, D. C., Greenberg, E., Lyon, G. J., Yu, D., McGrath, L. M., McMahon, W. M., Lee, P. C., Delucchi, K. L., Scharf, J. M., & Mathews, C. A. (2016). Social disinhibition is a heritable subphenotype of tics in Tourette syndrome. *Neurology*, *87(5)*, 497–504. https://doi.org/10.1212/wnl.0000000000002910

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., Cath, D. C., Kurlan, R., Robertson, M. M., Osiecki, L., Scharf, J. M., & Mathews, C. A. (2015). Lifetime Prevalence, Age of Risk, and Genetic Relationships of Comorbid Psychiatric Disorders in Tourette Syndrome. *JAMA Psychiatry*, *72(4)*, 325. https://doi.org/10.1001/jamapsychiatry.2014.2650

Hooper, S. D., Johansson, A. C., Tellgren-Roth, C., Stattin, E.-L., Dahl, N., Cavelier, L., & Feuk, L. (2012). Genome-wide sequencing for the identification of rearrangements associated with Tourette syndrome and obsessive-compulsive disorder. *BMC Medical Genetics*, *13(123)*. https://doi.org/10.1186/1471-2350-13-123

Howard, D. M., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., Shirali, M., Coleman, J. R. I., Hagenaars, S. P., Ward, J., Wigmore, E. M., Alloza, C., Shen, X., Barbu, M. C., Xu, E. Y., Whalley, H. C., Marioni, R. E., Porteous, D. J., Davies, G., Deary, I. J., … McIntosh, A. M. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature Neuroscience*, *22(3)*, 343–352. https://doi.org/10.1038/s41593-018-0326-7

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, *5(6)*, e1000529. https://doi.org/10.1371/journal.pgen.1000529

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., McGrath, L. M., Illmann, C., Sandor, P., Barr, C. L., Grados, M., Singer, H. S., Nöthen, M. M., Hebebrand, J., King, R. A., … Smit, J. (2017). Rare Copy Number Variants in NRXN1 and CNTN6 Increase Risk for Tourette Syndrome. *Neuron*, *94(6)*, 1101-1111.e7. https://doi.org/10.1016/j.neuron.2017.06.010

Hudziak, J. J., van Beijsterveldt, C. E. M., Althoff, R. R., Stanger, C., Rettew, D. C., Nelson, E. C., Todd, R. D., Bartels, M., & Boomsma, D. I. (2004). Genetic and Environmental Contributions to the Child Behavior ChecklistObsessive-Compulsive Scale. *Archives of General Psychiatry*, *61(6)*, 608. https://doi.org/10.1001/archpsyc.61.6.608

Huisman-van Dijk, H. M., Matthijssen, S. J. M. A., Stockmann, R. T. S., Fritz, A. V., & Cath, D. C. (2019). Effects of comorbidity on Tourette's tic severity and quality of life*. Acta Neurologica Scandinavica, 140(6)*, 390–398. https://doi.org/10.1111/ane.13155

Hyde, T. M., Aaronson, B. A., Randolph, C., Rickler, K. C., & Weinberger, D. R. (1992). Relationship of birth weight to the phenotypic expression of Gilles de la Tourette's syndrome in monozygotic twins. *Neurology*, *42(3)*, 652–652. https://doi.org/10.1212/wnl.42.3.652

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., & Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, *36(9)*, 949–951. https://doi.org/10.1038/ng1416

Illumina. (2014). *DNA Copy Number and Loss of Heterozygosity Analysis Algorithms*. Illumina. https://www.illumina.com/documents/products/technotes/technote_cnv_algorithms.pdf

Illumina. (2017). *Infinium® Global Screening Array-24 v1.0*. Illumina. https://grcf.jhmi.edu/wp-content/uploads/2017/12/infinium-commercial-gsa-data-sheet-370-2016-016.pdf

Illumina. (2020a). *GenomeStudio v2.0.5*. Illumina. https://www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html

Illumina. (2020b). *InfiniumTM Global Screening Array-24 v3.0 BeadChip*. Illumina. https://www.illumina.com/content/dam/illumina-

marketing/documents/products/datasheets/infinium-global-screening-array-data-sheet-370-2016-016.pdf

Illumna. (n.d.). *Infinium Omni5-4 Kit*. Illumina. Retrieved May 18, 2022, from https://www.illumina.com/products/by-type/microarray-kits/infinium-omni5-quad.html

Inouye, E. (1965). Similar and Dissimilar Manifestations of Obsessive-compulsive Neurosis in Monozygotic Twins. *American Journal of Psychiatry*, *121(12)*, 1171–1175. https://doi.org/10.1176/ajp.121.12.1171

IOCDF-GC & OCGAS. (2017). Revealing the complex genetic architecture of obsessive–compulsive disorder using meta-analysis. *Molecular Psychiatry*, *23(5)*, 1181–1188. https://doi.org/10.1038/mp.2017.154

Jankovic, J., & Deng, H. (2007). Candidate Locus for Chorea and Tic Disorders at 15q? *Pediatric Neurology*, *37(1)*, 70–73. https://doi.org/10.1016/j.pediatrneurol.2007.02.015

Karagiannidis, I., Dehning, S., Sandor, P., Tarnok, Z., Rizzo, R., Wolanczyk, T., Madruga-Garrido, M., Hebebrand, J., Nöthen, M. M., Lehmkuhl, G., Farkas, L., Nagy, P., Szymanska, U., Anastasiou, Z., Stathias, V., Androutsos, C., Tsironi, V., Koumoula, A., Barta, C., … Paschou, P. (2013). Support of the histaminergic hypothesis in Tourette Syndrome: association of the histamine decarboxylase gene in a large sample of families. *Journal of Medical Genetics*, *50(11)*, 760–764. https://doi.org/10.1136/jmedgenet-2013-101637

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581(7809)*, 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kaufman, J., Kobak, K., Birmaher, B., & de Lacy, N. (2021). KSADS-COMP Perspectives on Child Psychiatric Diagnostic Assessment and Treatment Planning. *Journal of the American Academy of Child & Adolescent Psychiatry*, *60(5)*, 540–542. https://doi.org/10.1016/j.jaac.2020.08.470

Kerbeshian, J., Severud, R., Burd, L., & Larson, L. (2000). Peek-a-boo fragile site at 16d associated with Tourette syndrome, bipolar disorder, autistic disorder, and mental retardation. *American Journal of Medical Genetics*, *96(1)*, 69–73. https://doi.org/10.1002/(sici)1096-8628(20000207)96:1<69::aid-ajmg14>3.0.co;2-5

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders

in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62(6)*, 593. https://doi.org/10.1001/archpsyc.62.6.593

Khramtsova, E. A., Heldman, R., Derks, E. M., Yu, D., Davis, L. K., & Stranger, B. E. (2019). Sex differences in the genetic architecture of obsessive-compulsive disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *180(6)*, 351–364. https://doi.org/10.1002/ajmg.b.32687

Kjeldbjerg, M. L., & Clausen, L. (2021). Prevalence of binge-eating disorder among children and adolescents: a systematic review and meta-analysis. *European Child & Adolescent Psychiatry*. https://doi.org/10.1007/s00787-021-01850-2

Lawson-Yuen, A., Saldivar, J.-S., Sommer, S., & Picker, J. (2008). Familial deletion within NLGN4 associated with autism and Tourette syndrome. *European Journal of Human Genetics*, *16(5)*, 614–618. https://doi.org/10.1038/sj.ejhg.5202006

Leckman, J. F., Sholomskas, D., Thompson, W. D., Belanger, A., & Weissman, M. M. (1982). Best Estimate of Lifetime Psychiatric Diagnosis. *Archives of General Psychiatry*, *39(8)*, 879. https://doi.org/10.1001/archpsyc.1982.04290080001001

Lee, P. H., Anttila, V., Won, H., Feng, Y.-C. A., Rosenthal, J., Zhu, Z., Tucker-Drob, E. M., Nivard, M. G., Grotzinger, A. D., Posthuma, D., Wang, M. M.-J., Yu, D., Stahl, E. A., Walters, R. K., Anney, R. J. L., Duncan, L. E., Ge, T., Adolfsson, R., Banaschewski, T., … Smoller, J. W. (2019). Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell*, *179(7)*, 1469-1482.e11. https://doi.org/10.1016/j.cell.2019.11.020

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., & Wray, N. R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, *28(19)*, 2540–2542. https://doi.org/10.1093/bioinformatics/bts474

Lenane, M. C., Swedo, S. E., Leonard, H., Pauls, D. L., Sceery, W., & Rapoport, J. L. (1990). Psychiatric Disorders in First Degree Relatives of Children and Adolescents with Obsessive Compulsive Disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *29(3)*, 407–412. https://doi.org/10.1097/00004583-199005000-00012

Lewien, C., Genuneit, J., Meigen, C., Kiess, W., & Poulain, T. (2021). Sleep-related difficulties in healthy children and adolescents. *BMC Pediatrics*, *21(1)*. https://doi.org/10.1186/s12887-021-02529-y

Lewis. (1936). Problems of Obsessional Illness. *Proceedings of the Royal Society of Medicine*, *29(4)*, 325–336.

Li, W., & Olivier, M. (2013). Current analysis platforms and methods for detecting copy number variation. *Physiological Genomics*, *45(1)*, 1–16. https://doi.org/10.1152/physiolgenomics.00082.2012

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype Imputation. *Annual Review of Genomics and Human Genetics*, *10(1)*, 387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242

Liu, S., Tian, M., He, F., Li, J., Xie, H., Liu, W., Zhang, Y., Zhang, R., Yi, M., Che, F., Ma, X., Zheng, Y., Deng, H., Wang, G., Chen, L., Sun, X., Xu, Y., Wang, J., Zang, Y., … Guan, J.-S. (2019). Mutations in ASH1L confer susceptibility to Tourette syndrome. *Molecular Psychiatry*, *25(2)*, 476–490. https://doi.org/10.1038/s41380-019-0560-8

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45(6)*, 580–585. https://doi.org/10.1038/ng.2653

Macé, A., Tuke, M. A., Beckmann, J. S., Lin, L., Jacquemont, S., Weedon, M. N., Reymond, A., & Kutalik, Z. (2016). New quality measure for SNP array based CNV detection. *Bioinformatics*, *32(21)*, 3298–3305. https://doi.org/10.1093/bioinformatics/btw477

Malhotra, D., & Sebat, J. (2012). CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics. *Cell*, *148(6)*, 1223–1241. https://doi.org/10.1016/j.cell.2012.02.039

Manchia, M., Cullis, J., Turecki, G., Rouleau, G. A., Uher, R., & Alda, M. (2013). The Impact of Phenotypic and Genetic Heterogeneity on Results of Genome Wide Association Studies of Complex Diseases. *PLoS ONE*, *8(10)*, e76295. https://doi.org/10.1371/journal.pone.0076295

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26(22)*, 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

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., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461(7265)*, 747–753. https://doi.org/10.1038/nature08494

Mansueto, C. S., & Keuler, D. J. (2007). Tic or Compulsion? It's Tourettic OCD. *FOCUS*, *5(3)*, 361–367. https://doi.org/10.1176/foc.5.3.foc361

Marchini, J. (2019). Haplotype Estimation and Genotype Imputation. In D. J. Balding, I. Moltke, & J. Marioni, *Handbook of Statistical Genomics* (pp. 87–113). John Wiley & Sons Ltd.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11(7)*, 499–511. https://doi.org/10.1038/nrg2796

Marshall, C. R., Howrigan, D. P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D. S., Antaki, D., Shetty, A., Holmans, P. A., Pinto, D., Gujral, M., Brandler, W. M., Malhotra, D., Wang, Z., Fajarado, K. V. F., Maile, M. S., Ripke, S., Agartz, I., Albus, M., … Sebat, J. (2016). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics*, *49(1)*, 27–35. https://doi.org/10.1038/ng.3725

Mataix-Cols, D., Isomura, K., Pérez-Vigil, A., Chang, Z., Rück, C., Larsson, K. J., Leckman, J. F., Serlachius, E., Larsson, H., & Lichtenstein, P. (2015). Familial Risks of Tourette Syndrome and Chronic Tic Disorders. *JAMA Psychiatry*, *72(8)*, 787. https://doi.org/10.1001/jamapsychiatry.2015.0627

Mathews, C. A., Badner, J. A., Andresen, J. M., Sheppard, B., Himle, J. A., Grant, J. E., Williams, K. A., Chavira, D. A., Azzam, A., Schwartz, M., Reus, V. I., Kim, S. W., Cook, E. H., & Hanna, G. L. (2012). Genome-Wide Linkage Analysis of Obsessive-Compulsive Disorder Implicates Chromosome 1p36. *Biological Psychiatry*, *72(8)*, 629–636. https://doi.org/10.1016/j.biopsych.2012.03.037

Mathews, C. A., & Grados, M. A. (2011). Familiality of Tourette Syndrome, Obsessive-Compulsive Disorder, and Attention-Deficit/Hyperactivity Disorder: Heritability Analysis in a Large Sib-Pair Sample. *Journal of the American Academy of Child & Adolescent Psychiatry*, *50(1)*, 46–54. https://doi.org/10.1016/j.jaac.2010.10.004

Matsumoto, N., David, D. E., Johnson, E. W., Konecki, D., Burmester, J. K., Ledbetter, D. H., & Weber, J. L. (2000). Breakpoint sequences of an 1;8 translocation in a family with Gilles de la Tourette syndrome. *European Journal of Human Genetics*, *8(11)*, 875–883. https://doi.org/10.1038/sj.ejhg.5200549

Mattheisen, M., Samuels, J. F., Wang, Y., Greenberg, B. D., Fyer, A. J., McCracken, J. T., Geller, D. A., Murphy, D. L., Knowles, J. A., Grados, M. A., Riddle, M. A., Rasmussen, S. A., McLaughlin, N. C., Nurmi, E. L., Askland, K. D., Qin, H.-D., Cullen, B. A., Piacentini, J., Pauls, D. L., … Nestadt, G. (2014). Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS. *Molecular Psychiatry*, *20(3)*, 337–344. https://doi.org/10.1038/mp.2014.43

McGrath, L. M., Yu, D., Marshall, C., Davis, L. K., Thiruvahindrapuram, B., Li, B., Cappi, C., Gerber, G., Wolf, A., Schroeder, F. A., Osiecki, L., O'Dushlaine, C., Kirby, A., Illmann, C., Haddad, S., Gallagher, P., Fagerness, J. A., Barr, C. L., Bellodi, L., … Scharf, J. M. (2014). Copy Number Variation in Obsessive-Compulsive Disorder and Tourette Syndrome: A Cross-Disorder Study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53(8)*, 910–919. https://doi.org/10.1016/j.jaac.2014.04.022

Meisner, A., & Chatterjee, N. (2019). Disease Risk Models. In D. J. Balding, I. Moltke, & J. Marioni, *Handbook of Statistical Genomics* (pp. 815–841). John Wiley & Sons Ltd.

Melchior, L., Bertelsen, B., Debes, N. M., Groth, C., Skov, L., Mikkelsen, J. D., Brøndum-Nielsen, K., & Tümer, Z. (2013). Microduplication of 15q13.3 and Xq21.31 in a family with tourette syndrome and comorbidities. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *162B(8)*, 825–831. https://doi.org/10.1002/ajmg.b.32186

Mérette, C., Brassard, A., Potvin, A., Bouvier, H., Rousseau, F., Émond, C., Bissonnette, L., Roy, M.-A., Maziade, M., Ott, J., & Caron, C. (2000). Significant Linkage for Tourette Syndrome in a Large French Canadian Family. *The American Journal of Human Genetics*, *67(4)*, 1008–1013. https://doi.org/10.1086/303093

Merikangas, K. R., He, J., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., Benjet, C., Georgiades, K., & Swendsen, J. (2010). Lifetime Prevalence of Mental Disorders in U.S. Adolescents: Results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry*, *49(10)*, 980–989. https://doi.org/10.1016/j.jaac.2010.05.017

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2018). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, *47(D1)*, D419–D426. https://doi.org/10.1093/nar/gky1038

Mullins, N., Forstner, A. J., O'Connell, K. S., Coombes, B., Coleman, J. R. I., Qiao, Z., Als, T. D., Bigdeli, T. B., Børte, S., Bryois, J., Charney, A. W., Drange, O. K., Gandal, M. J., Hagenaars, S. P., Ikeda, M., Kamitaki, N., Kim, M., Krebs, K., Panagiotaropoulou, G., … Andreassen, O. A. (2021). Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nature Genetics*, *53(6)*, 817–829. https://doi.org/10.1038/s41588-021-00857-4

Murphy, D. L., Timpano, K. R., Wheaton, M. G., Greenberg, B. D., & Miguel, E. C. (2010). Obsessive-compulsive disorder and its related disorders: a reappraisal of obsessive-compulsive spectrum concepts. *Dialogues in Clinical Neuroscience*, *12(2)*, 131–148. https://doi.org/10.31887/dcns.2010.12.2/dmurphy

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., Garcia, J., Bedoya, G., Cornejo, W., Herrera, L. D., Romero, R., Fournier, E., Reus, V. I., Lowe, T. L., Farooqi, I. S., … Ruiz-Linares, A. (2013). CNV Analysis in Tourette Syndrome

Implicates Large Genomic Rearrangements in COL8A1 and NRXN1. *PLoS ONE*, *8(3)*, e59061. https://doi.org/10.1371/journal.pone.0059061

Nestadt, G., Lan, T., Samuels, J., Riddle, M., Bienvenu, O. J., Liang, K. Y., Hoehn-Saric, R., Cullen, B., Grados, M., Beaty, T. H., & Shugart, Y. Y. (2000). Complex Segregation Analysis Provides Compelling Evidence for a Major Gene Underlying Obsessive-Compulsive Disorder and for Heterogeneity by Sex. *The American Journal of Human Genetics*, *67(6)*, 1611–1616. https://doi.org/10.1086/316898

Nestadt, G., Wang, Y., Grados, M. A., Riddle, M. A., Greenberg, B. D., Knowles, J. A., Fyer, A. J., McCracken, J. T., Rauch, S. L., Murphy, D. L., Rasmussen, S. A., Cullen, B., Piacentini, J., Geller, D., Pauls, D., Bienvenu, O. J., Chen, Y., Liang, K. Y., Goes, F. S., … Chang, Y. C. (2011). Homeobox genes in obsessive-compulsive disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *159B(1)*, 53–60. https://doi.org/10.1002/ajmg.b.32001

Nestadt, Gerald, Samuels, J., Riddle, M., Bienvenu, O. J., Liang, K.-Y., LaBuda, M., Walkup, J., Grados, M., & Hoehn-Saric, R. (2000). A Family Study of Obsessive-compulsive Disorder. *Archives of General Psychiatry*, *57(4)*, 358. https://doi.org/10.1001/archpsyc.57.4.358

NHGRI. (2020). *Genome-Wide Association Studies Fact Sheet*. Genome.Gov; National Human Genome Research Institute. https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet

Nicolini, H., Cruz, C., Camarena, B., Orozco, B., Kennedy, J. L., King, N., Weissbecker, K., de la Fuente, J. R., & Sidenberg, D. (1996). DRD2, DRD3 and 5HT2A receptor genes polymorphisms in obsessive-compulsive disorder. *Molecular Psychiatry*, *1(6)*, 461–465.

Nievergelt, C. M., Maihofer, A. X., Klengel, T., Atkinson, E. G., Chen, C.-Y., Choi, K. W., Coleman, J. R. I., Dalvie, S., Duncan, L. E., Gelernter, J., Levey, D. F., Logue, M. W., Polimanti, R., Provost, A. C., Ratanatharathorn, A., Stein, M. B., Torres, K., Aiello, A. E., Almli, L. M., … Koenen, K. C. (2019). International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci. *Nature Communications*, *10(1)*. https://doi.org/10.1038/s41467-019-12576-w

Nussbaum, R. L., McInnes, R. R., & Willard, H. F. (2016). *Thompson & Thompson Genetics in Medicine*. Elsevier Health Sciences.

O'Donovan, M. C. (2015). What have we learned from the Psychiatric Genomics Consortium. *World Psychiatry*, *14(3)*, 291–293. https://doi.org/10.1002/wps.20270

Osborn, I. (1999). *Tormenting Thoughts and Secret Rituals: The Hidden Epidemic of Obsessive-Compulsive Disorder* (1st ed.). Dell.

Otowa, T., Hek, K., Lee, M., Byrne, E. M., Mirza, S. S., Nivard, M. G., Bigdeli, T., Aggen, S. H., Adkins, D., Wolen, A., Fanous, A., Keller, M. C., Castelao, E., Kutalik, Z., der Auwera, S. V., Homuth, G., Nauck, M., Teumer, A., Milaneschi, Y., … Hettema, J. M. (2016). Meta-analysis of genome-wide association studies of anxiety disorders. *Molecular Psychiatry*, *21(10)*, 1391–1399. https://doi.org/10.1038/mp.2015.197

Pardiñas, A. F., Holmans, P., Pocklington, A. J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S. E., Bishop, S., Cameron, D., Hamshere, M. L., Han, J., Hubbard, L., Lynham, A., Mantripragada, K., Rees, E., MacCabe, J. H., McCarroll, S. A., Baune, B. T., Breen, G., … Walters, J. T. R. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics*, *50(3)*, 381–389. https://doi.org/10.1038/s41588-018-0059-2

Park, L. S., Burton, C. L., Dupuis, A., Shan, J., Storch, E. A., Crosbie, J., Schachar, R. J., & Arnold, P. D. (2016). The Toronto Obsessive-Compulsive Scale: Psychometrics of a Dimensional Measure of Obsessive-Compulsive Traits. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55(4)*, 310-318.e4. https://doi.org/10.1016/j.jaac.2016.01.008

Paschou, P., Feng, Y., Pakstis, A. J., Speed, W. C., DeMille, M. M., Kidd, J. R., Jaghori, B., Kurlan, R., Pauls, D. L., Sandor, P., Barr, C. L., & Kidd, K. K. (2004). Indications of Linkage and Association of Gilles de la Tourette Syndrome in Two Independent Family Samples: 17q25 Is a Putative Susceptibility Region. *The American Journal of Human Genetics*, *75(4)*, 545–560. https://doi.org/10.1086/424389

Patel, C., Cooper-Charles, L., McMullan, D. J., Walker, J. M., Davison, V., & Morton, J. (2011). Translocation breakpoint at 7q31 associated with tics: further evidence for IMMP2L as a candidate gene for Tourette syndrome. *European Journal of Human Genetics*, *19(6)*, 634–639. https://doi.org/10.1038/ejhg.2010.238

Pauls, D L, Alsobrook 2nd, J. P., Goodman, W., Rasmussen, S., & Leckman, J. F. (1995). A family study of obsessive-compulsive disorder. *American Journal of Psychiatry*, *152(1)*, 76–84. https://doi.org/10.1176/ajp.152.1.76

Pauls, David L. (2003). An update on the genetics of Gilles de la Tourette syndrome. *Journal of Psychosomatic Research*, *55(1)*, 7–12. https://doi.org/10.1016/s0022-3999(02)00586-x

Pauls, David L. (2010). The genetics of obsessive-compulsive disorder: a review. *Dialogues in Clinical Neuroscience*, *12(2)*, 149–163. https://doi.org/10.31887/dcns.2010.12.2/dpauls

Pauls, David L., & Leckman, J. F. (1986). The Inheritance of Gilles de la Tourette's Syndrome and Associated Behaviors. *New England Journal of Medicine*, *315(16)*, 993–997. https://doi.org/10.1056/nejm198610163151604

Pauls, David L., Raymond, C. L., Stevenson, J. M., & Leckman, J. F. (1991). A Family Study of Gilles de la Tourette Syndrome. *American Journal of Human Genetics*, *48(1)*, 154–163.

Petek, E., Windpassinger, C., Vincent, J. B., Cheung, J., Boright, A. P., Scherer, S. W., Kroisel, P. M., & Wagner, K. (2001). Disruption of a Novel Gene (IMMP2L) by a Breakpoint in 7q31 Associated with Tourette Syndrome. *The American Journal of Human Genetics*, *68(4)*, 848–858. https://doi.org/10.1086/319523

PGC. (n.d.). *Psychiatric Genomics Consortium | Psychiatric Genomics Consortium*. Psychiatric Genomics Consortium; https://www.facebook.com/PGCgenetics/. Retrieved May 20, 2022, from https://www.med.unc.edu/pgc/

Pies, R. (2008). Maimonides and Depression. *American Journal of Psychiatry*, *165(8)*, 1050–1051. https://doi.org/10.1176/appi.ajp.2008.08040502

Pounraja, V. K., Jayakar, G., Jensen, M., Kelkar, N., & Girirajan, S. (2019). A machine-learning approach for accurate detection of copy number variants from exome sequencing. *Genome Research*, *29(7)*, 1134–1143. https://doi.org/10.1101/gr.245928.118

Price, R. A. (1985). A Twin Study of Tourette Syndrome. *Archives of General Psychiatry*, *42(8)*, 815. https://doi.org/10.1001/archpsyc.1985.01790310077011

Qi, Y., Zheng, Y., Li, Z., Liu, Z., & Xiong, L. (2019). Genetic Studies of Tic Disorders and Tourette Syndrome. *Methods in Molecular Biology*, *2011*, 547–571. https://doi.org/10.1007/978-1-4939-9554-7_32

Qi, Y., Zheng, Y., Li, Z., & Xiong, L. (2017). Progress in Genetic Studies of Tourette's Syndrome. *Brain Sciences*, *7(10)*, 134. https://doi.org/10.3390/brainsci7100134

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Reese, H. E., Scahill, L., Peterson, A. L., Crowe, K., Woods, D. W., Piacentini, J., Walkup, J. T., & Wilhelm, S. (2014). The Premonitory Urge to Tic: Measurement, Characteristics, and Correlates in Older Adolescents and Adults. *Behavior Therapy*, *45(2)*, 177–186. https://doi.org/10.1016/j.beth.2013.09.002

Reynolds, T., Johnson, E. C., Huggett, S. B., Bubier, J. A., Palmer, R. H. C., Agrawal, A., Baker, E. J., & Chesler, E. J. (2020). Interpretation of psychiatric genome-wide association studies with multispecies heterogeneous functional genomic data integration. *Neuropsychopharmacology*, *46(1)*, 86–97. https://doi.org/10.1038/s41386-020-00795-5

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D. I., South, S. T., Thorland, E. C., Pineda-Alvarez, D., Aradhya, S., & Martin, C. L. (2020). Technical standards for the interpretation and reporting of

constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genetics in Medicine*, *22(2)*, 245–257. https://doi.org/10.1038/s41436-019-0686-8

Risch, N., & Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. *Science*, *273(5281)*, 1516–1517. https://doi.org/10.1126/science.273.5281.1516

Ritter, M. L., Guo, W., Samuels, J. F., Wang, Y., Nestadt, P. S., Krasnow, J., Greenberg, B. D., Fyer, A. J., McCracken, J. T., Geller, D. A., Murphy, D. L., Knowles, J. A., Grados, M. A., Riddle, M. A., Rasmussen, S. A., McLaughlin, N. C., Nurmi, E. L., Askland, K. D., Cullen, B., … Shugart, Y. Y. (2017). Genome Wide Association Study (GWAS) between Attention Deficit Hyperactivity Disorder (ADHD) and Obsessive Compulsive Disorder (OCD). *Frontiers in Molecular Neuroscience*, *10(83)*. https://doi.org/10.3389/fnmol.2017.00083

Roach, J. C., Glusman, G., Hubley, R., Montsaroff, S. Z., Holloway, A. K., Mauldin, D. E., Srivastava, D., Garg, V., Pollard, K. S., Galas, D. J., Hood, L., & Smit, A. F. A. (2011). Chromosomal Haplotypes by Genetic Phasing of Human Families. *The American Journal of Human Genetics*, *89(3)*, 382–397. https://doi.org/10.1016/j.ajhg.2011.07.023

Robertson, M. M., Shelley, B. P., Dalwai, S., Brewer, C., & Critchley, H. D. (2006). A patient with both Gilles de la Tourette's syndrome and chromosome 22q11 deletion syndrome: clue to the genetics of Gilles de la Tourette's syndrome? *Journal of Psychosomatic Research*, *61(3)*, 365–368. https://doi.org/10.1016/j.jpsychores.2006.06.011

Rosario-Campos, M. C., Miguel, E. C., Quatrano, S., Chacon, P., Ferrao, Y., Findley, D., Katsovich, L., Scahill, L., King, R. A., Woody, S. R., Tolin, D., Hollander, E., Kano, Y., & Leckman, J. F. (2006). The Dimensional Yale–Brown Obsessive–Compulsive Scale (DY-BOCS): an instrument for assessing obsessive–compulsive symptom dimensions. *Molecular Psychiatry*, *11(5)*, 495–504. https://doi.org/10.1038/sj.mp.4001798

Rubinacci, S., Delaneau, O., & Marchini, J. (2020). Genotype imputation using the Positional Burrows Wheeler Transform. *PLOS Genetics*, *16(11)*, e1009049. https://doi.org/10.1371/journal.pgen.1009049

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., Tikhomirov, A., Osiecki, L., Illmann, C., Pluzhnikov, A., Konkashbaev, A., Davis, L. K., Han, B., Crane, J., Moorjani, P., … Pauls, D. L. (2013). Genome-wide association study of Tourette's syndrome. *Molecular Psychiatry*, *18(6)*, 721–728. https://doi.org/10.1038/mp.2012.69

Scharf, Jeremiah M., Miller, L. L., Mathews, C. A., & Ben-Shlomo, Y. (2012). Prevalence of Tourette Syndrome and Chronic Tics in the Population-Based Avon Longitudinal Study of Parents and Children Cohort. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51(2)*, 192-201.e5. https://doi.org/10.1016/j.jaac.2011.11.004

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, *270(5235)*, 467–470. https://doi.org/10.1126/science.270.5235.467

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., … Wigler, M. (2004). Large-Scale Copy Number Polymorphism in the Human Genome. *Science*, *305(5683)*, 525–528. https://doi.org/10.1126/science.1098918

Seiser, E. L., & Innocenti, F. (2014). Hidden Markov Model-Based CNV Detection Algorithms for Illumina Genotyping Microarrays. *Cancer Informatics*, *13(7)*, 77–83. https://doi.org/10.4137/cin.s16345

Shaikh, T. H. (2017). Copy Number Variation Disorders. *Current Genetic Medicine Reports*, *5(4)*, 183–190. https://doi.org/10.1007/s40142-017-0129-2

Sharma, E., Sharma, L. P., Balachander, S., Lin, B., Manohar, H., Khanna, P., Lu, C., Garg, K., Thomas, T. L., Au, A. C. L., Selles, R. R., Højgaard, D. R. M. A., Skarphedinsson, G., & Stewart, S. E. (2021). Comorbidities in Obsessive-Compulsive Disorder Across the Lifespan: A Systematic Review and Meta-Analysis. *Frontiers in Psychiatry, 12*. https://doi.org/10.3389/fpsyt.2021.703701

Shelley, B. P., Robertson, M. M., & Turk, J. (2007). An individual with Gilles de la Tourette syndrome and Smith?Magenis microdeletion syndrome: is chromosome 17p11.2 a candidate region for Tourette syndrome putative susceptibility genes? *Journal of Intellectual Disability Research*, *51(8)*, 620–624. https://doi.org/10.1111/j.1365-2788.2006.00943.x

Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., Wu, J., & Xiao, J. (2018). Comprehensive Assessment of Genotype Imputation Performance. *Human Heredity*, *83(3)*, 107–116. https://doi.org/10.1159/000489758

Shugart, Y. Y., Samuels, J., Willour, V. L., Grados, M. A., Greenberg, B. D., Knowles, J. A., McCracken, J. T., Rauch, S. L., Murphy, D. L., Wang, Y., Pinto, A., Fyer, A. J., Piacentini, J., Pauls, D. L., Cullen, B., Page, J., Rasmussen, S. A., Bienvenu, O. J., Hoehn-Saric, R., … Nestadt, G. (2006). Genomewide linkage scan for obsessive-compulsive disorder: evidence for susceptibility loci on chromosomes 3q, 7p, 1q, 15q, and 6q. *Molecular Psychiatry*, *11(8)*, 763–770. https://doi.org/10.1038/sj.mp.4001847

Simonic, I., Gericke, G. S., Ott, J., & Weber, J. L. (1998). Identification of Genetic Markers Associated with Gilles de la Tourette Syndrome in an Afrikaner Population. *The American Journal of Human Genetics*, *63(3)*, 839–846. https://doi.org/10.1086/302002

Singer, H. S. (2000). Current issues in Tourette syndrome. *Movement Disorders*, *15(6)*, 1051–1063.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). BioMart – biological queries made easy. *BMC Genomics*, *10(1)*, 22. https://doi.org/10.1186/1471-2164-10-22

Sordo Vieira, L., Nguyen, B., Nutley, S. K., Bertolace, L., Ordway, A., Simpson, H., Zakrzewski, J., Jean Gilles, M. E., Nosheny, R., Weiner, M., Mackin, R. S., & Mathews, C. A. (2022). Self-reporting of psychiatric illness in an online patient registry is a good indicator of the existence of psychiatric illness. *Journal of Psychiatric Research*, *151*, 34–41. https://doi.org/10.1016/j.jpsychires.2022.03.022

SPARK Consortium. (2018). SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron*, *97(3)*, 488–493. https://doi.org/10.1016/j.neuron.2018.01.015

State, M. W., Greally, J. M., Cuker, A., Bowers, P. N., Henegariu, O., Morgan, T. M., Gunel, M., DiLuna, M., King, R. A., Nelson, C., Donovan, A., Anderson, G. M., Leckman, J. F., Hawkins, T., Pauls, D. L., Lifton, R. P., & Ward, D. C. (2003). Epigenetic abnormalities associated with a chromosome 18(q21-q22) inversion and a Gilles de la Tourette syndrome phenotype. *Proceedings of the National Academy of Sciences*, *100(8)*, 4684–4689. https://doi.org/10.1073/pnas.0730775100

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *PLOS Biology*, *13(7)*, e1002195. https://doi.org/10.1371/journal.pbio.1002195

Stewart, S. E., Yu, D., Scharf, J. M., Neale, B. M., Fagerness, J. A., Mathews, C. A., Arnold, P. D., Evans, P. D., Gamazon, E. R., Osiecki, L., McGrath, L., Haddad, S., Crane, J., Hezel, D., Illman, C., Mayerfeld, C., Konkashbaev, A., Liu, C., Pluzhnikov, A., … Pauls, D. L. (2013). Genome-wide association study of obsessive-compulsive disorder. *Molecular Psychiatry*, *18(7)*, 788–798. https://doi.org/10.1038/mp.2012.85

Storch, E. A., Murphy, T. K., Bagner, D. M., Johns, N. B., Baumeister, A. L., Goodman, W. K., & Geffken, G. R. (2006). Reliability and validity of the Child Behavior Checklist Obsessive-Compulsive Scale. *Journal of Anxiety Disorders*, *20(4)*, 473–485. https://doi.org/10.1016/j.janxdis.2005.06.002

Streiner, D. L., & Norman, G. R. (2011). Correction for Multiple Testing. *CHEST*, *140(1)*, 16–18. https://doi.org/10.1378/chest.11-0523

Strom, N. I., Soda, T., Mathews, C. A., & Davis, L. K. (2021). A dimensional perspective on the genetics of obsessive-compulsive disorder. *Translational Psychiatry*, *11(1)*. https://doi.org/10.1038/s41398-021-01519-z

Sullivan, L. M. (2017). *Essentials of Biostatistics in Public Health (Essential Public Health)* (3rd ed.). Jones & Bartlett Learning.

Sullivan, P. F., Agrawal, A., Bulik, C. M., Andreassen, O. A., Børglum, A. D., Breen, G., Cichon, S., Edenberg, H. J., Faraone, S. V., Gelernter, J., Mathews, C. A., Nievergelt, C. M., Smoller, J. W., & O'Donovan, M. C. (2018). Psychiatric Genomics: An Update and an Agenda. *American Journal of Psychiatry*, *175(1)*, 15–27. https://doi.org/10.1176/appi.ajp.2017.17030283

Sun, N., Nasello, C., Deng, L., Wang, N., Zhang, Y., Xu, Z., Song, Z., Kwan, K., King, R. A., Pang, Z. P., Xing, J., Heiman, G. A., & Tischfield, J. A. (2017). The PNKD gene is associated with Tourette Disorder or Tic disorder in a multiplex family. *Molecular Psychiatry*, *23(6)*, 1487–1495. https://doi.org/10.1038/mp.2017.179

Sundaram, S. K., Huq, A. M., Wilson, B. J., & Chugani, H. T. (2010). Tourette syndrome is associated with recurrent exonic copy number variants. *Neurology*, *74(20)*, 1583–1590. https://doi.org/10.1212/wnl.0b013e3181e0f147

Sundaram, Senthil K., Huq, A. M., Sun, Z., Yu, W., Bennett, L., Wilson, B. J., Behen, M. E., & Chugani, H. T. (2011). Exome sequencing of a pedigree with tourette syndrome or chronic tic disorder. *Annals of Neurology*, *69(5)*, 901–904. https://doi.org/10.1002/ana.22398

TAAICG. (2007). Genome Scan for Tourette Disorder in Affected-Sibling-Pair and Multigenerational Families. *The American Journal of Human Genetics*, *80(2)*, 265–272. https://doi.org/10.1086/511052

Tarnok, Z., Ronai, Z., Gervai, J., Kereszturi, E., Gadoros, J., Sasvari-Szekely, M., & Nemoda, Z. (2007). Dopaminergic candidate genes in Tourette syndrome: Association between tic severity and 3′ UTR polymorphism of the dopamine transporter gene. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *144B(7)*, 900–905. https://doi.org/10.1002/ajmg.b.30517

Terra. (n.d.). *Terra*. Terra. Retrieved June 9, 2022, from https://app.terra.bio

The 1000 Genomes Project Consortium, Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526(7571)*, 68–74. https://doi.org/10.1038/nature15393

Thompson, M., Comings, D. E., Feder, L., George, S. R., & O'Dowd, B. F. (1998). Mutation screening of the dopamine D1 receptor gene in Tourette's syndrome and alcohol dependent patients. *American Journal of Human Genetics*, *81(3)*, 241–244.

Thygesen, J. H., Presman, A., Harju-Seppänen, J., Irizar, H., Jones, R., Kuchenbaecker, K., Lin, K., Alizadeh, B. Z., Austin-Zimmerman, I., Bartels-Velthuis, A., Bhat, A., Bruggeman, R., Cahn, W., Calafato, S., Crespo-Facorro, B., de Haan, L., de Zwarte, S. M. C., Di Forti, M., Díez-Revuelta, Á., … Bramon, E. (2021). Genetic copy number variants, cognition and psychosis: a meta-analysis and a family study. *Molecular Psychiatry*, *26(9)*, 5307–5319. https://doi.org/10.1038/s41380-020-0820-7

Tienari, P. (1963). Psychiatrc illnesses in identical twins. *Acta Psychiatrica Scandinavica*, *39(171)*, 1–195.

Townsend, L., Kobak, K., Kearney, C., Milham, M., Andreotti, C., Escalera, J., Alexander, L., Gill, M. K., Birmaher, B., Sylvester, R., Rice, D., Deep, A., & Kaufman, J. (2020). Development of Three Web-Based Computerized Versions of the Kiddie Schedule for Affective Disorders and Schizophrenia Child Psychiatric Diagnostic Interview: Preliminary Validity Data. *Journal of the American Academy of Child & Adolescent Psychiatry*, *59(2)*, 309–325. https://doi.org/10.1016/j.jaac.2019.05.009

Tsetsos, F., Yu, D., Sul, J. H., Huang, A. Y., Illmann, C., Osiecki, L., Darrow, S. M., Hirschtritt, M. E., Greenberg, E., Muller-Vahl, K. R., Stuhrmann, M., Dion, Y., Rouleau, G. A., Aschauer, H., Stamenkovic, M., Schlögelhofer, M., Sandor, P., Barr, C. L., Grados, M. A., … Zinner, S. (2021). Synaptic processes and immune-related pathways implicated in Tourette syndrome. *Translational Psychiatry*, *11(1)*, 56. https://doi.org/10.1038/s41398-020-01082-z

Tsuang, M. T., Tohen, M., & Joones, P. B. (2011). *Textbook of Psychiatric Epidemiology* (3rd ed.). Wiley-Blackwell.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1(59)*. https://doi.org/10.1038/s43586-021-00056-9

Vadgama, N., Pittman, A., Simpson, M., Nirmalananthan, N., Murray, R., Yoshikawa, T., De Rijk, P., Rees, E., Kirov, G., Hughes, D., Fitzgerald, T., Kristiansen, M., Pearce, K., Cerveira, E., Zhu, Q., Zhang, C., Lee, C., Hardy, J., & Nasir, J. (2019). De novo single-nucleotide and copy number variation in discordant monozygotic twins reveals disease-related genes. *European Journal of Human Genetics*, *27(7)*, 1121–1133. https://doi.org/10.1038/s41431-019-0376-7

Verbanck, M., Chen, C.-Y., Neale, B., & Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian

randomization between complex traits and diseases. *Nature Genetics*, *50(5)*, 693–698. https://doi.org/10.1038/s41588-018-0099-7

Verkerk, A. J. M. H., Mathews, C. A., Joosse, M., Eussen, B. H. J., Heutink, P., & Oostra, B. A. (2003). Cntnap2 is disrupted in a family with gilles de la tourette syndrome and obsessive compulsive disorder. *Genomics*, *82(1)*, 1–9. https://doi.org/10.1016/s0888-7543(03)00097-1

Visscher, P. M., Hemani, G., Vinkhuyzen, A. A. E., Chen, G.-B., Lee, S. H., Wray, N. R., Goddard, M. E., & Yang, J. (2014). Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples. *PLoS Genetics*, *10(4)*, e1004269. https://doi.org/10.1371/journal.pgen.1004269

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics*, *101(1)*, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Walkup, J. T., LaBuda, M. C., Singer, H. S., Brown, J., Riddle, M. A., & Hurko, O. (1996). Family study and segregation analysis of Tourette syndrome: evidence for a mixed model of inheritance. *American Journal of Human Genetics*, *59(3)*, 684–693.

Wand, H., Lambert, S. A., Tamburro, C., Iacocca, M. A., O'Sullivan, J. W., Sillari, C., Kullo, I. J., Rowley, R., Dron, J. S., Brockman, D., Venner, E., McCarthy, M. I., Antoniou, A. C., Easton, D. F., Hegele, R. A., Khera, A. V., Chatterjee, N., Kooperberg, C., Edwards, K., … Wojcik, G. L. (2021). Improving reporting standards for polygenic scores in risk prediction studies. *Nature*, *591(7849)*, 211–219. https://doi.org/10.1038/s41586-021-03243-6

Wang, K., Chen, Z., Tadesse, M. G., Glessner, J., Grant, S. F. A., Hakonarson, H., Bucan, M., & Li, M. (2008). Modeling genetic inheritance of copy number variations. *Nucleic Acids Research*, *36(21)*, e138–e138. https://doi.org/10.1093/nar/gkn641

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H., & Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, *17(11)*, 1665–1674. https://doi.org/10.1101/gr.6861907

Wang, S., Mandell, J. D., Kumar, Y., Sun, N., Morris, M. T., Arbelaez, J., Nasello, C., Dong, S., Duhn, C., Zhao, X., Yang, Z., Padmanabhuni, S. S., Yu, D., King, R. A., Dietrich, A., Khalifa, N., Dahl, N., Huang, A. Y., Neale, B. M., … Yu, D. (2018). De Novo Sequence and Copy Number Variants Are Strongly Associated with Tourette Disorder and Implicate Cell Polarity in Pathogenesis. *Cell Reports*, *24(13)*, 3441-3454.e12. https://doi.org/10.1016/j.celrep.2018.08.082

Watson, H. J., Yilmaz, Z., Thornton, L. M., Hübel, C., Coleman, J. R. I., Gaspar, H. A., Bryois, J., Hinney, A., Leppä, V. M., Mattheisen, M., Medland, S. E., Ripke, S., Yao, S., Giusti-Rodríguez, P., Hanscombe, K. B., Purves, K. L., Adan, R. A. H., Alfredsson, L., Ando, T., … Bulik, C. M. (2019). Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nature Genetics*, *51(8)*, 1207–1214. https://doi.org/10.1038/s41588-019-0439-2

Weissbecker, K., Baxter, L., Schwartz, J., Sparkes, R. S., & Spence, M. A. (1989). Linkage analysis of obsessive compulsive disorder. *Cytogenetics and Cell Genetics*, *51(1105)*.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4(43)*, 1686. https://doi.org/10.21105/joss.01686

Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A Note on Exact Tests of Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics*, *76(5)*, 887–893. https://doi.org/10.1086/429864

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., Smith, L., Dong, S., Samocha, K. E., Neale, B. M., Coppola, G., Mathews, C. A., Tischfield, J. A., Scharf, J. M., State, M. W., … Yu, D. (2017). De Novo Coding Variants Are Strongly Associated with Tourette Disorder. *Neuron*, *94(3)*, 486-499.e9. https://doi.org/10.1016/j.neuron.2017.04.024

Woody, S. R., Steketee, G., & Chambless, D. L. (1995). Reliability and validity of the Yale-Brown Obsessive-Compulsive Scale. *Behaviour Research and Therapy*, *33(5)*, 597–605. https://doi.org/10.1016/0005-7967(94)00076-v

WTCCC. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447(7145)*, 661–678. https://doi.org/10.1038/nature05911

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, *88(1)*, 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., & Visscher, P. M. (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nature Genetics*, *49(9)*, 1304–1310. https://doi.org/10.1038/ng.3941

Yang, Z., Wu, H., Lee, P. H., Tsetsos, F., Davis, L. K., Yu, D., Lee, S. H., Dalsgaard, S., Haavik, J., Barta, C., Zayats, T., Eapen, V., Wray, N. R., Devlin, B., Daly, M., Neale, B., Børglum, A. D., Crowley, J. J., Scharf, J., … Paschou, P. (2021).

Investigating Shared Genetic Basis Across Tourette Syndrome and Comorbid Neurodevelopmental Disorders Along the Impulsivity-Compulsivity Spectrum. *Biological Psychiatry*, *90(5)*, 317–327. https://doi.org/10.1016/j.biopsych.2020.12.028

Yilmaz, Z., Halvorsen, M., Bryois, J., Yu, D., Thornton, L. M., Zerwas, S., Micali, N., Moessner, R., Burton, C. L., Zai, G., Erdman, L., Kas, M. J., Arnold, P. D., Davis, L. K., Knowles, J. A., Breen, G., Scharf, J. M., Nestadt, G., Mathews, C. A., … Crowley, J. J. (2018). Examination of the shared genetic basis of anorexia nervosa and obsessive–compulsive disorder. *Molecular Psychiatry*, *25(9)*, 2036–2046. https://doi.org/10.1038/s41380-018-0115-4

Yorston, G., & Hindley, N. (1998). Study of a nervous disorder characterized by motor incoordination with echolalia and coprolalia. *History of Psychiatry*, *19(33)*, 097–101. https://doi.org/10.1177/0957154x9800903307

Young, M. E. (2010). *Comparison of Diagnostic Interviews for Children Accessing Outpatient Mental Health Services*. (Publication No. osu1274748739) [Doctoral dissertation, Ohio State University].

Yu, D., Mathews, C. A., Scharf, J. M., Neale, B. M., Davis, L. K., Gamazon, E. R., Derks, E. M., Evans, P., Edlund, C. K., Crane, J., Fagerness, J. A., Osiecki, L., Gallagher, P., Gerber, G., Haddad, S., Illmann, C., McGrath, L. M., Mayerfeld, C., Arepalli, S., … Pauls, D. L. (2015). Cross-Disorder Genome-Wide Analyses Suggest a Complex Genetic Relationship Between Tourette's Syndrome and OCD. *American Journal of Psychiatry*, *172(1)*, 82–93. https://doi.org/10.1176/appi.ajp.2014.13101306

Yu, D., Sul, J. H., Tsetsos, F., Nawaz, M. S., Huang, A. Y., Zelaya, I., Illmann, C., Osiecki, L., Darrow, S. M., Hirschtritt, M. E., Greenberg, E., Muller-Vahl, K. R., Stuhrmann, M., Dion, Y., Rouleau, G., Aschauer, H., Stamenkovic, M., Schlögelhofer, M., Sandor, P., … Scharf, J. M. (2019). Interrogating the Genetic Determinants of Tourette's Syndrome and Other Tic Disorders Through Genome-Wide Association Studies. *American Journal of Psychiatry*, *176(3)*, 217–227. https://doi.org/10.1176/appi.ajp.2018.18070857

Yuan, A., Wang, Z., Xu, W., Ding, Q., Zhao, Y., Han, J., & Sun, J. (2020). A Rare Novel CLCN2 Variation and Risk of Gilles de la Tourette Syndrome: Whole-Exome Sequencing in a Multiplex Family and a Follow-Up Study in a Chinese Population. *Frontiers in Psychiatry*, *11*, 543911. https://doi.org/10.3389/fpsyt.2020.543911

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*, *16(3)*, 172–183. https://doi.org/10.1038/nrg3871

Zhang, D., Dey, R., & Lee, S. (2020). Fast and robust ancestry prediction using principal component analysis. *Bioinformatics*, *36(11)*, 3439–3446. https://doi.org/10.1093/bioinformatics/btaa152

Zhang, L., Bai, W., Yuan, N., & Du, Z. (2019). Comprehensively benchmarking applications for detecting copy number variation. *PLOS Computational Biology*, *15(5)*, e1007069. https://doi.org/10.1371/journal.pcbi.1007069

Zhao, X., Wang, S., Hao, J., Zhu, P., Zhang, X., & Wu, M. (2020). A Whole-Exome Sequencing Study of Tourette Disorder in a Chinese Population. *DNA and Cell Biology*, *39(1)*, 63–68. https://doi.org/10.1089/dna.2019.4746

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W.-Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, *50(9)*, 1335–1341. https://doi.org/10.1038/s41588-018-0184-y

Zhuang, X., Ye, R., So, M.-T., Lam, W.-Y., Karim, A., Yu, M., Ngo, N. D., Cherny, S. S., Tam, P. K.-H., Garcia-Barcelo, M.-M., Tang, C. S., & Sham, P. C. (2020). A random forest-based framework for genotyping and accuracy assessment of copy number variations. *NAR Genomics and Bioinformatics*, *2(3)*. https://doi.org/10.1093/nargab/lqaa071

Zohar, A. H. (1999). The Epidemiology of Obsessive-Compulsive Disorder in Children and Adolescents. *Child and Adolescent Psychiatric Clinics of North America*, *8(3)*, 445–460. https://doi.org/10.1016/s1056-4993(18)30163-9

Zohar, J. (1987). Obsessive-Compulsive Disorders: Theory and Management. *Psychosomatics*, *28*.

Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, *109(4)*, 1193–1198. https://doi.org/10.1073/pnas.1119675109

BIOGRAPHICAL SKETCH

Franjo Ivankovic was born in Mostar and raised in Donji Hamzići, Bosnia and Herzegovina. Following graduation from Elementary School Čerin, in 2009 they enrolled in the High Electrical Engineering School Ruđer Bošković in Mostar. In 2011, they transferred to the United World College in Mostar, from where they ultimately graduated in 2013. In 2013, Franjo moved to the United States of America where they matriculated to the University of Florida.

They graduated from the University of Florida in 2017 with a Bachelor of Science in integrative biology and Bachelor of Arts in anthropology, earning a certificate in medical anthropology. They remained at the University of Florida to pursue Ph.D. program in genetics and genomics. Initially, Franjo spent time studying molecular and RNA biology of neuromuscular diseases under the mentorship of Dr. Maurice S. Swanson. However, to fully pursue their passion, Franjo transitioned into the Department of Psychiatry to pursue research in psychiatric genomics under the mentorship of Dr. Carol A. Mathews, focusing on psychiatric disorders of childhood, particularly Tourette syndrome and obsessive-compulsive disorder.

In the laboratory of Dr. Mathews, Franjo worked on exploring and better understanding psychiatric phenotypes in childhood and adolescence, delineating genetic underpinnings of childhood psychiatric disorders, and deconvoluting complex phenotype-genotype relationships within and between neurodevelopmental psychiatric disorders.

Franjo graduated in the summer semester of 2022. They continued their work on psychiatric genomics as a post-doctoral research fellow.