Title
How copy number variant gene expression contributes to neurobehavioral traits

Author
Vysotskiy, Mikhail

Publication Date
2022

Peer reviewed|Thesis/dissertation

How copy number variant gene expression contributes to neurobehavioral traits

by
Mikhail Vysotskiy

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Marina Sirota
925B61AB9C41499...

Marina Sirota
Chair

DocuSigned by:
Lauren Weiss

Lauren Weiss

DocuSigned by:
Elliott Sherr
5C5FEA6D64994FC...

Elliott Sherr

Committee Members

This work is dedicated to all individuals who – knowingly or unknowingly – provided the data that I analyzed. I acknowledge that you are humans and not just values in a file.

It is difficult – especially with the time crunch under which I write this dissertation – to find all the people to acknowledge for this doctorate. A few are on my mind.

I have been in the Weiss lab for many years and I have intersected with dozens of full time and part time researchers. When I started, we were a bunch that kept to ourselves, and I didn't know until much later how great it could feel to experience science as a team. I acknowledge Erika Yeh for her kind presence in my life after she left the lab, with her experienced advice and for involving me in guest lecturing, one of my favorite pastimes. Aleksa ("Dr Alex") Obradovic and Ian Finnegan brought a lot of excitement into the lab and I am honored to be on an upcoming publication with you. I felt particularly supported in my last year by Bhavika Mam, Zoe Bell, Andrew Gonzalez, and Clémence Fournié. Thank you for letting me express myself to you and for pushing me to wrap things up and leave. Thank you to Michela Traglia as well for her bioinformatics and scientific mentoring.

I acknowledge my family here in San Francisco. While I am not "first-gen" in the technical sense of the term, it was always difficult for me to explain what I was doing in an American-style university. My grandfather, in particular, gave me constant motivation and always aimed to show me how strong I really was. My mother – who could relate in some ways to my experience but

not others – tried to be patient with me no matter what I was going through. (And she took care of me in a great amount of ways). My father had his own way of being patient with me, and though I always feel like I disappoint him for not getting an MD instead, I believe him when he says he is proud of me.

Of course, mentorship is one of the most important things for a doctoral student. My advisor Laurie gave me, most of all, patience (I suppose that is a major theme here – yet I realize that family and advisors aren't all that different). Taking on a grad student for the first time is very challenging, and I acknowledge the effort it took to get me to the end. Somehow, it appears she wants another student after me; I'm surprised I didn't turn her off from that. My committee including Marina Sirota and Elliott Sherr (as well as Noah Zaitlen who gave me a lot of technical support though is not "officially" credited in this thesis) provided a great deal of support as well as grounding, reminding me that I really have accomplished something, even when it didn't feel like a lot. The PSPG graduate program has some great people working there. I acknowledge Rebecca Dawon and Nicole Flowers who were essential to my admissions "side project" as well as Deanna Kroetz, the program chair during my "formative years" who had more dedication to the students than anyone expected.

Fourth, I acknowledge my friends near and far who have not yet been mentioned. From San Francisco, Vika Smilansky (and recently her husband too!) plus many of the members of the Russian-speaking community I grew up with have seen me through this time (and I got to see them pursue their higher education as well). Also I acknowledge my Summerbridge crew – Hewett Yip, Patrick Gabon, Anni Wen and others – whom I've known for 10+ years and still

feel valued and remembered by. From college, there were several of us who tried the grad school things and had differing paths. Suzy Beeler, Madi Hansen, and I all had struggles (and occasional successes) that both overlapped and felt entirely different. I am glad to be able to speak to them non-stop about my school experience and feel understood. Being understood, I guess, is what we all want, and the more educated we get, the harder being understood becomes. To close philosophically, I suppose I will state that understanding is what got me to this point – that and patience.

Chapter two of this thesis has been previously published in:

Vysotskiy, M., Zhong, X., Miller-Fleming, T.W. et al. Integration of genetic, transcriptomic, and clinical data provides insight into 16p11.2 and 22q11.2 CNV genes. Genome Med 13, 172 (2021). https://doi.org/10.1186/s13073-021-00972-1

Chapter three will be published in the future, though not necessarily in the same form as it appears here.

“Always remember: You’re going to die soon enough anyway; even if it’s a hundred years from now, that’s still the blink of a cosmic eye. In the meantime, live like a scientist—even a controversial one with only an ally or two in all the world—and treat life as a grand experiment, blood, sweat, tears and all. Bear in mind that there's no such thing as a failed experiment—only data.”

- Jesse Bering

How copy number variant gene expression contributes to neurobehavioral traits

Mikhail Vysotskiy

Abstract

Problem

Many neurobehavioral traits are highly heritable, yet specific genes underlying them are difficult to identify. Copy number variants – large, often multi-gene, deletions and duplications - are one category of mutation that drives traits including autism spectrum disorder (ASD), bipolar disorder, schizophrenia, obesity, and intellectual disability. A 600kb 30-gene region on chromosome 16p11.2, for example, is strongly associated with ASD when duplicated or deleted, and associated with schizophrenia when only duplicated, while deletions of a 3mb 60-gene region on chromosome 22q11.2 are a well-known genetic cause of schizophrenia. Under the assumption that a subset of these 30 or 60 genes contribute to associated traits, either individually or in combination, we investigate the genic architecture of CNV association with neurobehavioral traits.

Methods

We think about CNVs' impact on traits in the context of gene expression: individuals that have genetic duplications have increased gene expression (1.5x of normal), while those with deletions have decreased gene expression (0.5x of normal). Among non-carriers, expression levels vary as well, although typically to a smaller extent. This gives rise to a hypothesis: expression variation of the genes in a CNV region may be associated with the CNV-associated (or related) traits in non-carriers. Under this hypothesis, we can separate out the association of the expression of

specific CNV genes with specific traits, which cannot be done in a CNV carrier where all genes are upregulated or downregulated.

We studied large populations of non-carriers with genetic data (and corresponding controls, where applicable) for five neurobehavioral traits: ASD, bipolar disorder, schizophrenia, BMI (as a proxy for obesity), and IQ (as a proxy for intellectual disability). We took an expression prediction approach which allowed us to convert GWAS-style data into imputed expression-level data which can be used for association analyses. We studied the association of individual CNV genes, pairs of CNV genes, and all genes in the region to these five traits. This study design was also used to assess association between CNV gene expression variation and clinical traits within a large biobank with genotypic and clinical information.

Findings

Following a brief introduction, the second chapter of this dissertation focuses on individual genes within GWAS datasets and clinical biobanks, and the third focuses on the extension of this approach to combinations of genes and the entire region. In support of our hypothesis, we were able to detect individual genes at 16p11.2 associated with neurobehavioral traits, most notably *INO80E*, significantly associated with schizophrenia and BMI, and nominally associated with bipolar disorder. Using the biobank, we found additional genes associated with related clinical traits including psychosis and mood disorders, with an overall over-representation of mental disorders in CNV gene associated phenotypes. We then found that variance in traits was better explained by pairs of CNV genes in nearly all instances, including those where we had identified single-gene associations. The regionwide prediction was associated with BMI and IQ at both

16p11.2 and 22q11.2, but not with any neuropsychiatric trait. The importance of the combinatorial contributions of genes did not extend to matched control regions for all 16p11.2 traits as well as ASD at 22q11.2. In sum, our studies provide insight into the transcription-based action of CNV genes, identify potential candidate genes for further study, describe combinatorial patterns of CNV gene impacts on neurobehavior, and demonstrate the utility of integrating genetic, clinical, and transcriptomic data for *in silico* analyses.

Table of Contents

List of Figures

List of Tables

Chapter one: Introduction

Although the pathophysiologies of many neurobehavioral traits are uncertain, it is clear that they have a significant genetic component. Estimates of heritability are 0.8 for autism spectrum disorder (ASD), 0.75 for bipolar disorder, and 0.81 for schizophrenia [1]. A major aim of research in the field of psychiatric and behavioral human genetics is the identification of risk factors for these disorders as well as interpreting how these risk factors might be biologically relevant, as this understanding could expedite the identification of improved traditional and gene-based therapeutics.

One starting point for identifying relevant biomarkers is by analyzing common population variation. This is typically done by collecting genetic samples from large case-control or population cohorts, and then using methods such as genome wide association studies (GWAS) in order to screen common variation (single nucleotide polymorphisms, SNPs) throughout the entire genome and find variants associated with traits. This method is very dependent on sample size, but large consortia including the Psychiatric Genomics Consortium (PGC) and the Genetic Investigation of Anthropometric Traits Consortium (GIANT) have been established to pool information for these studies. GWAS has so far pointed to 5 loci for ASD, 64 for bipolar disorder, 270 for schizophrenia, and 205 for IQ [2–5]. In this dissertation on neurobehavioral traits, I additionally include obesity (BMI). This is due to the finding that genes associated with BMI through GWAS have unusually high brain expression [6]. Using common genetic data, it is also possible to measure genetic correlation and find relationships between traits. In particular,

there is a high correlation between schizophrenia and bipolar disorder, ASD and IQ, and a small correlation between schizophrenia and ASD [7,8].

Other causes for neurobehavioral traits can come from much rarer mutations, more of which have been detected with the advent of exome and genome sequencing. The simplest type of rare mutation is a single nucleotide variant (SNV). These mutations vary in terms of their deleteriousness and effects on protein products. SNVs, both inherited and de novo, are frequently over-represented in ASD cases [9]. More complex mutations that are implicated in psychiatric traits are structural variants, such as translocations, insertions, deletions, and duplications. Within exons, structural variation can have a deleterious impact on function; in intergenic regions they can impact chromatin folding. Larger copy number variants (CNVs) spanning at least 1kb are deletions and duplications that often include multiple genes and have been detected by both traditional cytogenetic and contemporary microarray and sequencing methods. In ASD, about 5-10% of affected individuals have a pathogenic CNV [10]. The rates of pathogenic CNVs in schizophrenia are similar, with a notable overlap in ASD/schizophrenia CNVs [11].

Multigenic CNVs provide an opportunity to have a localized view at a genomic region that contains at least one neurobehavior-modifying gene, and yet fine-mapping a neurobehavioral trait to one or more of these genes has remained difficult. If we can identify specific genes that are responsible for phenotype, we can form further hypotheses about mechanisms – including understanding what goes awry during neurodevelopment – and lead to targeted therapies. Additional questions of interest involve interactions between genes within and outside of the CNV as well as pleiotropic effects of one gene controlling multiple traits.

One major CNV region that has been identified in the 1990s is located on 22q11.2 (also known as the locus of velo-cardio-facial syndrome and DiGeorge syndrome; the syndrome has been known from the 1970s). The most common CNVs at this locus span either a 3MB region encoding approximately 60 genes or a 1.5MB region with 35 genes, and deletions in the region are commonly detected (1 in 3000 births) [12,13]. The CNV is associated with multiple physical, developmental, and behavioral conditions, including ASD, intellectual disability , schizophrenia, bipolar disorder, and obesity [14–16]. Intriguingly, while overall risk for psychiatric disease increases with both deletions and duplications, it has been noted that duplication carriers are less likely to have schizophrenia [17]. A common variant in the 22q11.2 *COMT* gene has been associated with neurobehavioral traits and schizophrenia, though it is becoming appreciated that having the variant on its own is not enough to cause schizophrenia [18–21]. Mutations in the 22q11.2 gene *PRODH*, and general decreases in proline metabolism have been found in schizophrenic patients [19,22,23]; neuronal effects due to *Prodh* have been confirmed in mice, although the overall influence of the gene on neurobehavior is debated [24,25]. Furthermore, nearly all studies have been done in the context of deletions; as such, the phenotypes of duplications and why duplications have a protective effect on schizophrenia remain unsolved.

More recently, duplications and deletions of a 600kb region on 16p11.2 (spanning approximately 30 genes) were identified as ASD-associated mutations. The prevalence of this CNV in the autism population is about 0.4-1%, as opposed to 0.05% in controls [26–29]. Aside from ASD, several related traits have also been associated with the CNV: intellectual disability brain size (microcephaly with duplications, macrocephaly with deletions), schizophrenia, bipolar disorder,

and obesity, similar to the 22q11.2 region [30–32]. While several studies have attempted to dissect the region further, most have had limited success. A smaller deletion in this region, containing five genes – *MVP, CDIPT1*, *SEZ6L2, ASPHD1, and KCTD13* has been observed [33]. However, this small deletion is neither necessary nor sufficient to cause ASD-related symptoms: in the same family, there were individuals without the deletion that had ASD and individuals with the deletion but not the phenotype. Another study looked for genetic variations within the region that was associated with ASD: a rare exonic genetic variant was found in the *SEZ6L2* gene, but the association was not replicated in a different dataset [34]. Two different zebrafish knockdown models found that most genes in the region were necessary for brain development and identified dosage-sensitive (dependent on copy number) genes for brain size but with no direct implication for human behavior [35,36]. Notably, the two studies did not point to the same genes – *aldoa* and *kif22* in one, *kctd13* in the other. Transcriptional studies have confirmed that the genes in the region are transcribed according to their copy number in human LCLs, mouse brain, and human iPSCs, suggesting that transcription levels may be why the CNV is associated with human disease, but does not narrow down candidate genes [31,37,38]. To date, no conclusive link exists between a single gene in this region and a neurobehavioral trait.

Although deletions and duplications of genetic material at these locations include 30+ genes, it is not necessarily true that all genes are important for driving each phenotype. Disorders may be driven by one, all, or a subset of these genes, acting independently or together. In general, part of the difficulty in understanding complex traits in humans is that going from single variant causes (Mendelian or simple genetic traits) to multi-gene based causes is difficult due to a genome of tens of thousands of genes as well as many poorly understood noncoding regions. In yeast, a

single-cellular organism with an order of magnitude fewer genes, however, interaction mapping has revealed large patterns of pairwise effects [39].CNV regions lend themselves well to being studied in this way because it is reasonable to believe that the trait-relevant genes are in this locus, and the number of combinations, at least pairwise, are of a more reasonable scale. Studies in zebrafish and drosophila suggest additional complexity in the function of 16p11.2 genes, with interaction effects driving body and eye phenotypes [40,41]. In Chapter 2 of this dissertation, I analyze genes independently, and in Chapter 3 I include more complex genetic relationships.

The assumption underlying this work is that the link between CNV genes and disease is due to extreme changes in transcription levels based on copy number (0.5x in deletions, 1.5x in duplications). However, variation in transcription levels of these genes (some of which is due to genetics) is present among all individuals. Given that large expression changes due to CNVs are known to affect disease risk, we hypothesize that smaller changes in the expression of genes in the region (due to the presence of eQTLs) will also affect disease risk, albeit in a more modest way (**Figure 1.1**). If the mechanism of action for the genes is related to their transcription, studying the genes' transcriptional modifiers and regulators can not only provide further insight into the link between genes and disorders, but also extend the relevance of this knowledge to non-CNV carriers. In the following studies, I use population genetic data to determine how genetically-determined gene expression at the population level affects risk of neurobehavioral traits: ASD, schizophrenia, bipolar disorder, obesity (as measured by BMI), and intellectual disability (as measured by IQ). Two advantages of this approach are that it is purely driven by analyzing common genetic variation in the general population (which enables us to use large powerful previously-collected datasets and makes the result generalizable to individuals without

copy number variation in the regions) and that it enables us to efficiently test multiple genes in a combinatorial way. While this approach could theoretically be done by comparing measured expression levels between cases and controls, (1) expression data have much smaller sample sizes than genetic GWAS-style data; this is especially true in brain tissues that can only be collected post-mortem, (2) gene expression as measured at a given timepoint can be affected by disease-related (including medication), lifestyle, or perimortem factors; using genetically-regulated gene expression sidesteps this problem.

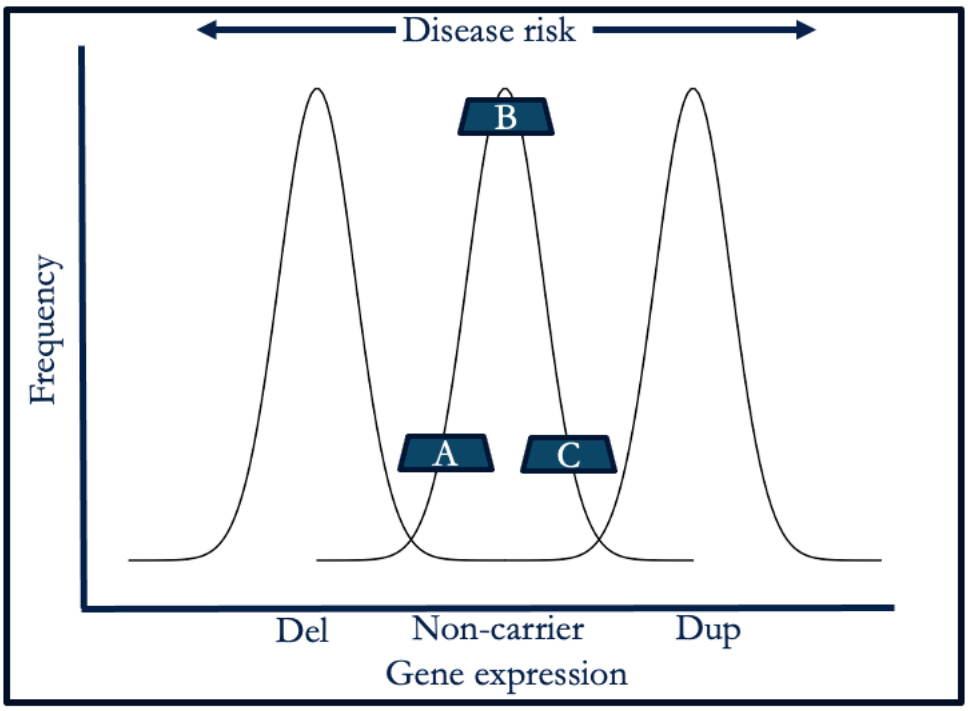


Figure 1.1: Variations in expression are hypothesized to increase disease risk in carriers as well as non-carriers.

Here, individuals A and C are more "deletion-like" and "duplication-like", respectively, in their gene expression than individual B, and we hypothesize that A and C would be more likely to develop a neurobehavioral trait than B.

The articles that make up the following chapters in the dissertation aim to provide insight into the impact of 16p11.2 and 22q11.2 copy number variants on the same five traits (ASD, schizophrenia, bipolar disorder, BMI, and IQ) in three ways: the contribution of individual genes, the contribution of pairs of genes, and the contribution of the entire region. As previously mentioned, these well-defined genetic regions associated with neurobehavioral traits allow themselves for a targeted fine-mapping approach and a deeper look into their genetic architecture. Successfully understanding genetic patterns at these CNV regions can first be applied to any other CNV-trait pairs but also tell us about what we might expect more broadly in understanding complex genetics, such as whether individual genes, pairs, or large combinatorial groups are likely to affect our traits of interest. Furthermore, understanding any specific genes or subsets of genes that drive one or more traits has therapeutic applications, as we can target specific genes and upregulate or downregulate them to counteract CNV impact. Because my study of CNV genes is primarily done in non-carriers, treatments based on any genes I find should be generalizable to the entire population of people affected by one of these five disorders.

References

1. Sullivan PF, Daly MJ, O'Donovan M. Genetic Architectures of Psychiatric Disorders: The Emerging Picture and Its Implications. Nat Rev Genet. 2012;13:537.

2. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51:431–44.

3. Consortium TSWG of the PG, Ripke S, Walters JT, O'Donovan MC. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. medRxiv. 2020;2020.09.12.20192922.

4. Savage JE, Jansen PR, Stringer S, Watanabe K, Bryois J, De Leeuw CA, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat Genet. 2018;50:912–9.

5. Mullins N, Forstner AJ, O'Connell KS, Coombes B, Coleman JRI, Qiao Z, et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat Genet. 2021;53:817–29.

6. Ndiaye FK, Huyvaert M, Ortalli A, Canouil M, Lecoeur C, Verbanck M, et al. The expression of genes in top obesity-associated loci is enriched in insula and substantia nigra brain regions involved in addiction and reward. Int J Obes (Lond). 2019;44:539.

7. Lee SH, Ripke S, Neale BM, Faraone S V., Purcell SM, Perlis RH, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat Genet. 2013;45:984.

8. Nishiyama T, Taniai H, Miyachi T, Ozaki K, Tomita M, Sumi S. Genetic correlation between autistic traits and IQ in a population-based sample of twins with autism spectrum disorders (ASDs). J Hum Genet. 2009;54:56–61.

9. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, et al. Excess of rare,

inherited truncating mutations in autism. Nat Genet. 2015;47:582–8.

10. Bacchelli E, Cameli C, Viggiano M, Igliozzi R, Mancini A, Tancredi R, et al. An integrated analysis of rare CNV and exome variation in Autism Spectrum Disorder using the Infinium PsychArray. Sci Rep. 2020;10:3198.

11. Kushima I, Aleksic B, Nakatochi M, Shimamura T, Okada T, Uno Y, et al. Comparative Analyses of Copy-Number Variation in Autism Spectrum Disorder and Schizophrenia Reveal Etiological Overlap and Biological Insights. Cell Rep. 2018;24:2838–56.

12. Karayiorgou M, Simon TJ, Gogos JA. 22q11.2 microdeletions: linking DNA structural variation to brain dysfunction and schizophrenia. Nat Rev Neurosci. 2010;11:402–16.

13. Kobrynski LJ, Sullivan KE. Velocardiofacial syndrome, DiGeorge syndrome: the chromosome 22q11.2 deletion syndromes. Lancet. 2007;370:1443–52.

14. Bassett AS, Marshall CR, Lionel AC, Chow EWC, Scherer SW. Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. Hum Mol Genet. 2008;17:4045–53.

15. Portnoï M-F. Microduplication 22q11.2: A new chromosomal syndrome. Eur J Med Genet. 2009;52:88–93.

16. Hoeffding LK, Trabjerg BB, Olsen L, Mazin W, Sparsø T, Vangkilde A, et al. Risk of Psychiatric Disorders Among Individuals With the 22q11.2 Deletion or Duplication. JAMA Psychiatry. 2017;79:348–51.

17. Rees E, Kirov G, Sanders A, Walters JTR, Chambert KD, Shi J, et al. Evidence that duplications of 22q11.2 protect against schizophrenia. Mol Psychiatry. 2014;19:37–40.

18. Egan MF, Goldberg TE, Kolachana BS, Callicott JH, Mazzanti CM, Straub RE, et al. Effect of COMT Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. Proc Natl Acad Sci U S A. 2001;98:6917–22.

19. Prasad SE, Howley S, Murphy KC. Candidate genes and the behavioral phenotype in 22q11.2 deletion syndrome. Dev Disabil Res Rev. 2008;14:26–34.

20. Williams HJ, Owen MJ, O'Donovan MC. Is COMT a Susceptibility Gene for Schizophrenia? Schizophr Bull. 2007;33:635–41.

21. Gothelf D, Eliez S, Thompson T, Hinard C, Penniman L, Feinstein C, et al. COMT genotype predicts longitudinal cognitive decline and psychosis in 22q11.2 deletion syndrome. Nat Neurosci. 2005;8:1500–2.

22. Jacquet H, Raux G, Thibaut F, Hecketsweiler B, Houy E, Demilly C, et al. PRODH mutations and hyperprolinemia in a subset of schizophrenic patients. Hum Mol Genet. 2002;11:2243–9.

23. Liu H, Heath SC, Sobin C, Roos JL, Galke BL, Blundell ML, et al. Genetic variation at the 22q11 PRODH2/DGCR6 locus presents an unusual pattern and increases susceptibility to schizophrenia. Proc Natl Acad Sci U S A. 2002;99:3717–22.

24. Devaraju P, Zakharenko SS. Mitochondria in complex psychiatric disorders: Lessons from mouse models of 22q11.2 deletion syndrome. BioEssays. 2017;39:1600177.

25. Willis A, Bender HU, Steel G, Valle D. PRODH variants and risk for schizophrenia. Amino Acids. 2008;35:673–9.

26. Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA, et al. Recurrent 16p11.2 microdeletions in autism. Hum Mol Genet. 2007;17:628–38.

27. Hanson E, Bernier R, Porche K, Jackson FI, Goin-Kochel RP, Snyder LG, et al. The Cognitive and Behavioral Phenotype of the 16p11.2 Deletion in a Clinically Ascertained Population. Biol Psychiatry. 2015;77:785–93.

28. Walsh KM, Bracken MB. Copy number variation in the dosage-sensitive 16p11.2 interval

accounts for only a small proportion of autism incidence: A systematic review and meta-analysis. Genet Med. 2011;13:377–84.

29. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, et al. Association between Microdeletion and Microduplication at 16p11.2 and Autism. N Engl J Med. 2008;358:667–75.

30. Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature. 2010;463:671–5.

31. Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, et al. Genome-wide Transcriptome Profiling Reveals the Functional Impact of Rare De Novo and Recurrent CNVs in Autism Spectrum Disorders. Am J Hum Genet. 2012;91:38–55.

32. Qureshi AY, Mueller S, Snyder AZ, Mukherjee P, Berman JI, Roberts TPL, et al. Opposing brain differences in 16p11.2 deletion and duplication carriers. J Neurosci. 2014;34:11199–211.

33. Crepel A, Steyaert J, De la Marche W, De Wolf V, Fryns J-P, Noens I, et al. Narrowing the critical deletion region for autism spectrum disorders on 16p11.2. Am J Med Genet Part B Neuropsychiatr Genet. 2011;156:243–5.

34. Kumar RA, Marshall CR, Badner JA, Babatz TD, Mukamel Z, Aldinger KA, et al. Association and Mutation Analyses of 16p11.2 Autism Candidate Genes. Reif A, editor. PLoS One. 2009;4:e4582.

35. Blaker-Lee A, Gupta S, McCammon JM, De Rienzo G, Sive H. Zebrafish homologs of genes within 16p11.2, a genomic region associated with brain disorders, are active during brain development, and include two deletion dosage sensor genes. Dis Model Mech. 2012;5.

36. Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S, et al. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant.

Nature. 2012;485:363–7.

37. Blumenthal I, Ragavendran A, Erdin S, Klei L, Sugathan A, Guide JR, et al. Transcriptional Consequences of 16p11.2 Deletion and Duplication in Mouse Cortex and Multiplex Autism Families. Am J Hum Genet. 2014;94:870–83.

38. Tai DJC, Razaz P, Erdin S, Gao D, Wang J, Nuttle X, et al. Tissue and cell-type specific molecular and functional signatures of 16p11.2 reciprocal genomic disorder across mouse brain and human neuronal models. bioRxiv. 2022;2022.05.12.491670.

39. Costanzo M, Hou J, Messier V, Nelson J, Rahman M, VanderSluis B, et al. Environmental robustness of the global yeast genetic interaction network. Science (80-). 2021;372.

40. Iyer J, Singh MD, Jensen M, Patel P, Pizzo L, Huber E, et al. Pervasive genetic interactions modulate neurodevelopmental defects of the autism-Associated 16p11.2 deletion in Drosophila melanogaster. Nat Commun. 2018;9:1–19.

41. McCammon JM, Blaker-Lee A, Chen X, Sive H. The 16p11.2 homologs fam57ba and doc2a generate certain brain and body phenotypes. Hum Mol Genet. 2017;26:3699–712.

Chapter two: Integration of genetic, transcriptomic, and clinical data provides insight into 16p11.2 and 22q11.2 CNV genes

Abstract

Background:

Deletions and duplications of the multigenic 16p11.2 and 22q11.2 copy number variant (CNV) regions are associated with brain-related disorders including schizophrenia, intellectual disability, obesity, bipolar disorder, and autism spectrum disorder (ASD). The contribution of individual CNV genes to each of these identified phenotypes is unknown, as well as the contribution of these CNV genes to other potentially subtler health implications for carriers. Hypothesizing that DNA copy number exerts most effects via impacts on RNA expression, we attempted a novel *in silico* fine-mapping approach in non-CNV carriers using both GWAS and biobank data.

Methods:

We first asked whether gene expression level in any individual gene in the CNV region alters risk for a known CNV-associated behavioral phenotype(s). Using transcriptomic imputation, we performed association testing for CNV genes within large genotyped cohorts for schizophrenia, IQ, BMI, bipolar disorder, and ASD. Second, we used a biobank containing electronic health data to compare the medical phenome of CNV carriers to controls within 700,000 individuals in order to investigate the full spectrum of health effects of the CNVs. Third, we used genotypes for over 48,000 individuals within the biobank to perform phenome-wide association studies between imputed expressions of individual 16p11.2 and 22q11.2 genes and over 1,500 health traits.

Results:

Using large genotyped cohorts, we found individual genes within 16p11.2 associated with schizophrenia (*TMEM219*, *INO80E*, *YPEL3*), BMI (*TMEM219*, *SPN*, *TAOK2*, *INO80E*), and IQ (*SPN*), using conditional analysis to identify upregulation of *INO80E* as the driver of schizophrenia, and downregulation of *SPN* and *INO80E* as increasing BMI. We identified both novel and previously observed overrepresented traits within the electronic health records of 16p11.2 and 22q11.2 CNV carriers. In the phenome-wide association study, we found seventeen significant gene-trait pairs, including psychosis (*NPIPB11*, *SLX1B*) and mood disorders (*SCARF2*), and overall enrichment of mental traits.

Conclusions:

Our results demonstrate how integration of genetic and clinical data aids in understanding CNV gene function and implicates pleiotropy and multigenicity in CNV biology.

BACKGROUND

Multi-gene copy number variants (CNVs), including a 600kb region at 16p11.2 and a 3Mb region at 22q11.2, are known causes of multiple brain-related disorders. The 16p11.2 CNV, originally identified as a risk factor for autism spectrum disorder (ASD), has also been associated with schizophrenia, bipolar disorder, intellectual disability, and obesity[1–5]. The 22q11.2 CNV, identified as the cause of DiGeorge (velocardiofacial) syndrome, is associated with schizophrenia, intellectual disability, obesity, bipolar disorder, and ASD, as well [6–11]. The effects of these two CNVs can be further subdivided into the effects of deletions vs. duplications. Some disorders are shared among carriers of deletions and duplications of the same region, and others show opposite associations. For instance, ASD and intellectual disability are observed in

both deletion and duplication carriers in both 16p11.2 and 22q11.2 [3–8,12–14]. Other traits are specific to one direction of the copy number change: schizophrenia and bipolar disorder are observed in 16p11.2 duplication carriers, but not deletion carriers [15]. A third category of 16p11.2 and 22q11.2-associated traits are "mirrored". 16p11.2 deletion carriers show increased rates of obesity, while duplication carriers tend to be underweight. 22q11.2 duplication carriers show reduced rates of schizophrenia, as opposed to increased rates in deletion carriers [1,16,17]. The question of which specific genes drive which brain-related traits associated with 16p11.2 or 22q11.2 CNVs remains unanswered. Likewise, what else these genes might be doing has been difficult to assess in small numbers of identified CNV carriers, who are primarily children. Identifying the role of specific gene(s) in behavioral and medical traits will clarify the biological processes that go awry as a result of these CNV mutations and the mechanisms by which they do so. Knowledge of the genes and mechanisms involved would, in turn, provide opportunities to develop targeted treatments.

Three of the traditional ways to map CNV genes to disorders are identifying loss-of-function mutations in these genes, analyzing smaller subsets of the entire region, and finding mutations in animal models that are sufficient to recapitulate the phenotype. The loss-of-function mutation method was used to fine-map the 17p11.2 CNV, another CNV associated with behavioral and non-behavioral traits [18,19]. Most of the features of the deletion syndrome, including intellectual disability, are represented in individuals who carry a defective copy of the *RAI1* gene due to point mutation [20]. Duplications of *Rai1* appear to explain body weight and behavior abnormalities in mouse models of 17p11.2 duplications [21]. Another example is the Williams syndrome CNV at 7q11.23 [22,23]. The cardiac traits associated with this syndrome are present

in individuals with only one functional copy of the *ELN* gene, but this gene does not explain the behavioral traits [24,25]. The second method, of finding a smaller "critical region" was used to fine-map the 17q21.31 CNV [26,27]. By comparing patients who had similar symptoms with overlapping cytogenetic profiles, the common breakpoints of the CNV region were refined to a region containing only six genes [27]. Later, Koolen *et al* identified patients showing intellectual disability and facial dysmorphisms characteristic of this CNV with disruptive mutations in one of the six genes, *KANSL1* [28]. The third method of recapitulating similar phenotypes in animal models was successful in identifying *TBX1* as a gene important for some of the physical traits involved with 22q11.2 deletions. Mice with heterozygous mutations in the *TBX1* gene show cardiac outflow tract anomalies, similar to human 22q11.2 deletion carriers [29–31]. However, it is unclear that *TBX1* is sufficient to explain brain-related disorders in 22q11.2 carriers [32,33].

The 16p11.2 and 22q11.2 CNVs have been resistant to these traditional approaches for fine-mapping brain-related traits. To date, no highly penetrant point mutations in 16p11.2 or 22q11.2 genes have been shown to be sufficient for a brain-related disorder. The most recent schizophrenia GWAS from the Psychiatric Genomics Consortium discovered a common SNP association near the 16p11.2 region, however the specific genes underlying GWAS signals are often unknown[34]. No small subsets of 16p11.2 or 22q11.2 genes have been proven necessary and sufficient to cause a brain-related disorder. A subregion of 22q11.2 has been proposed to explain ASD associated with deletions [35]. As this subset of 22q11.2 contains approximately 20 genes, it is likely that further fine-mapping within this subset is possible. At 16p11.2, a subset of five deleted genes was isolated in a family with a history of ASD [36]. However, this mutation neither caused ASD in all deletion carriers, nor was responsible for ASD in some non-carrier

family members. Non-human models for the 16p11.2 and 22q11.2 CNVs, as well as knockouts for individual genes are available in mouse, zebrafish, and fruit flies [37–42], but have not successfully mapped individual genes in these CNVs to brain-related traits [29–31]. Different zebrafish studies of 16p11.2 homologs have implicated different genes as phenotype drivers, as well as shown that most were involved in nervous system development [38,39,43]. The complex brain-related traits associated with these CNVs are unlikely to be fully captured in model organisms. Hallucinations, a common symptom of schizophrenia, can be identified only in humans. There may be other aspects of 16p11.2 and 22q11.2 CNV biology that are human-specific. For example, mice carrying 16p11.2 duplications are obese, while obesity is associated with deletions in humans [44]. Given the insufficiency of previous approaches, new approaches for fine-mapping genes in these regions for brain-related traits are necessary.

The motivation behind our approach is that in 16p11.2 and 22q11.2 CNV carriers, variation in gene copy number is expected to lead to variation in RNA expression level (with downstream effects on protein product). Expression measurements in mouse or human cell lines carrying 16p11.2 and 22q11.2 deletions and duplications confirm that for nearly all genes, duplication carriers have increased expression of individual CNV genes compared to controls, and deletion carriers have reduced expression compared to controls [45–50]. As the breakpoints of these CNVs are unlikely to cause gain-of-function, we believe that the variation in expression of one or more of the genes in/near the CNV is the cause of pathogenicity. While these CNVs significantly disrupt gene expression levels, most genes' expression levels vary among the general population, sometimes by a factor of two or more, as studies such as the Genotype-Tissue Expression Consortium (GTEx) have shown [51–54]. This variation can be, in part,

attributed to common genetic polymorphisms (expression quantitative trait loci, eQTLs). If large expression deviation in duplication and deletion carriers is a risk factor for a disorder, we hypothesize that more modest expression variation in the same genes among non-carriers will be a modest risk factor for the same disorder or milder related traits. This idea is analogous to the well-supported observation that common polymorphisms of small effect associated with a common trait can overlap with Mendelian genes for a similar trait [55–57].

Here, we perform three *in silico* studies of the impact of predicted expression of individual 16p11.2 and 22q11.2 genes, in comparison with the diagnosed CNVs, on human traits (**Figure 2.1**). First, we identify genes associated with brain-related disorders via expression variation. Recent tools have leveraged the heritability of gene expression, allowing us to "impute" gene expression for genotyped individuals using eQTLs [58,59]. We perform association testing between imputed expression and five brain-related traits common to the 16p11.2 and 22q11.2 CNVs for which large amounts of genetic data have been amassed: schizophrenia, IQ, BMI, bipolar disorder, and ASD [60–64]. We find at least one 16p11.2 gene associated with schizophrenia, IQ, and BMI. Second, we use BioVU, a biobank containing electronic health records (EHRs) for over 3 million individuals, to determine the medical traits in CNV carriers detected in our EHR system, confirming canonical CNV features and discovering novel over-represented traits [65]. We also probe the consequences of expression variation of individual 16p11.2 and 22q11.2 genes on the medical phenome, by imputing gene expression in the >48,000 genotyped individuals in the BioVU health system and performing a phenome-wide association test across all available traits. We find that mental disorders are over-represented among top gene-trait association pairs, and we highlight genes associated with the traits over-

represented in CNV carriers. Taken together, our work provides a comprehensive catalog of associations of individual CNV genes to traits across the phenome.

METHODS

GWAS Data for schizophrenia, IQ, BMI, bipolar disorder, and ASD

We obtained the imputed individual-level genotypes for ASD, bipolar disorder, and schizophrenia from the Psychiatric Genomics Consortium in PLINK format (Additional file 1: Table S1). These datasets include mainly European populations and are comprised of several independent cohorts: 30 in bipolar disorder (N = 19,202 cases 30,472 controls, downloaded July 2019), 46 in schizophrenia (N = 31,507 cases 40,230 controls, downloaded July 2018), 14 in ASD (N = 7,386 cases, 8,566 controls, downloaded May 2019) [61,62,66]. For two additional traits, we used publicly available summary statistics: BMI from the Genetic Investigation of ANthropometric Traits (GIANT) consortium (2015, both sexes, n=339,224, downloaded June 2019) and IQ from Savage *et al* 2018 hosted by the Complex Trait Genomics lab at VU Amsterdam (n=269,867, downloaded May 2019) [63,64].

For replication studies and comparison of PheWAS results, we used the publicly available GWAS summary statistics for schizophrenia, IQ, BMI, bipolar disorder, and ASD from the UK Biobank [67]. We could not use the UK Biobank IQ data for replication of our discovery IQ data, as the datasets overlap. The list of UK Biobank phenotypes used is in Table S1 in Additional file 1. In addition, we used individual-level data from the UK Biobank (n = 408,375)

to perform conditional analysis for BMI fine-mapping, but chose not to use it for discovery analysis because of previously-observed high inflation of summary statistics [68,69].

Expression prediction models

In order to impute gene expression, we obtained PrediXcan models for 48 tissues based on GTEx v7 Europeans [58,59,70]. These models were generated by training an elastic net model that finds the best set of cis-SNP predictors for the expression of a gene in a tissue in the GTEx genotyped individuals [58]. Only models with predicted-observed correlation $R^2 > 0.01$ and cross-validated prediction performance $P < 0.05$ are kept.

Genes studied

We studied all coding and noncoding genes at the 16p11.2 and 22q11.2 copy number variant loci for which expression prediction models were available. We included flanking genes in a 200kb window upstream and downstream of the CNV breakpoints. Overall, 37 coding and 8 noncoding genes at or near 16p11.2, as well as 52 coding and 30 noncoding genes at or near 22q11.2, were tested. Not all genes in the CNV regions were available to be analyzed through our methods; noncoding genes were especially unlikely to have a high-quality predictive model in any tissue. Thirty-four genes (of which 27 were non-coding) at or near 16p11.2 lacked high-quality prediction models in every tissue. One hundred two genes (of which 90 were non-coding) at or near 22q11.2 lacked high-quality prediction models in every tissue. (Additional file 2:Table S2, Additional file 3: Fig. S1).

Comparison of observed expression correlations with predicted expression correlations

Observed expression correlations were calculated at a tissue-specific level on data from GTEx v7 [71]. Tissue-specific predicted expression was calculated by applying the appropriate GTEx predictive model on the GTEx v6p genotypes (dbgap id: phs000424.v6.p1) for 450 individuals. To minimize spurious correlations, the predicted expression levels were rigorously filtered and normalized. Specifically, the expression levels were filtered for outliers (values above 1.5*interquartile range, in either direction), adjusted for the principal components of both the predicted expression levels and the first 20 PCs of the GTEx genotypes, inverse-quantile normalized, re-adjusted for principal components, and re-filtered for outliers. We observed that normalization of the predicted expression reintroduced correlation between expression and the genotypic PCs, leading us to perform the correction twice.

Association analysis in individual-level data

Each of the three PGC collections went through quality control, filtering, and PCA calculation, as described previously [60–62]. In each individual cohort, the convert_plink_to_dosage.py script in PrediXcan was used to convert chromosome 16 and 22 genotypes from PLINK to dosage format, keeping only the SNPs used in at least one predictive model. Using these dosages, the --predict function in PrediXcan was used to generate predicted expressions of CNV genes for each individual. Genes with predicted expression of 0 for all individuals in a single tissue were filtered out. The average number of genes filtered out across tissue-cohort pairs was

0.89; the maximum was 11 in thyroid tissue in the Japanese schizophrenia cohort. Cross-tissue association studies between predicted expression and case-control status were performed using MultiXcan. In brief, MultiXcan takes the matrix of predicted expressions across tissues for a single gene, calculates the principal components of this matrix to adjust for collinearity, fits a model between these PCs and case-control status, and reports the results of the overall fit [59]. As in the PGC association studies, our analysis was adjusted by the principal components that were significantly associated with each trait– 7 for bipolar disorder, 10 for schizophrenia, and 8 for autism case-control studies (the autism trios were not adjusted for covariates). UK Biobank MultiXcan analysis was limited to individuals who reported their ethnicity as "white", and included age, age-squared, and 40 principal components as covariates.

Meta-analysis with METAL on the p-values from MultiXcan, weighted by the sample size of each cohort, was used to calculate a cross-cohort association statistic for each trait individually [72]. The joint fit in MultiXcan generates an F-statistic that is always greater than zero, while some of our traits of interest have a specific expected direction (only seen in deletion carriers or only seen in duplication carriers). Thus, a direction was assigned to each MultiXcan result. This was done by running a tissue-specific PrediXcan association analysis between predicted expressions and case-control status (using --logistic), which calculates a signed association Z-score for every gene. The sign of the mean Z-score for that gene across all tissues was the direction of association used for meta-analysis.

Association analysis in summary-level data

Both the single-tissue PrediXcan and the multi-tissue MultiXcan methods have been extended to estimate the association results between genetically regulated expression and a trait if only the summary statistics for the trait are available. For each trait's summary statistics, the summary version of PrediXcan (S-PrediXcan) and the associated MetaMany.py script was used to calculate the per-tissue association results for each gene in 48 GTEx tissues. Association results were aggregated across tissues using the summary version of MultiXcan (S-MultiXcan). The mean single-tissue Z-score (as reported in the zmean column in the S-MultiXcan output) was used as the direction of association. The UK Biobank replication studies were performed in the same way.

Conditional analysis to fine-map associations

Existing methods for fine-mapping PrediXcan associations (such as FOCUS [73] and MR-JTI [74]) are tissue-specific and focus on summary statistics. Given that we have individual-level data and use a cross-tissue approach, we chose to use a conditional analysis approach. In order to adapt the multi-tissue association analysis to perform conditional testing, "conditioned predicted expressions" were generated for a set of genes associated with the same trait. As an example, take the set of three genes [*INO80E, YPEL3, TMEM219*] associated with schizophrenia. In order to condition on *INO80E*, for example, the predicted expression of *INO80E* was regressed out of the predicted expressions of *YPEL3* and *TMEM219*. Conditioning was only done in tissues where the predicted expressions of the genes were correlated (Spearman correlation $P < 0.05$). Another

set of conditioned predicted expressions was generated by adjusting the predicted expression of *INO80E* by the predicted expressions of [*TMEM219, YPEL3*]. Separately, these per-tissue conditioned predicted expressions were used as inputs for a MultiXcan analysis and METAL meta-analysis on schizophrenia as described earlier. All three individually associated genes were tested in this manner. The same analysis was later used to test for independence of association between BMI in the UK Biobank as well as *psychosis* and *morbid obesity* traits in the PheWAS. The P_{cond} reported in the text is the p-value of a gene-trait pair when adjusting for all other genes considered for conditioning for this trait, unless otherwise stated. To validate that our approach explained all GWAS signal at the loci, we also conditioned the MultiXcan analysis on lead GWAS SNP(s) that were also eQTLs. The GWAS conditioning was performed in PLINK using the --condition function, with principal components (and age for BMI) as covariates.

Phenome-wide association studies

Vanderbilt University Medical Center (VUMC) houses de-identified phenotypic data in the form of the electronic health records (EHR) within the synthetic derivative (SD) system [75]. The SD contains EHR data including ICD9/10 billing codes, physician notes, lab results, and similar documentation for 3.1 million individuals. BioVU is a biobank at VUMC that is composed of a subset of individuals from the SD that have de-identified DNA samples linked to their EHR phenotype information. The clinical information is updated every 1-3 months for the de-identified EHRs. Detailed description of program operations, ethical considerations, and continuing oversight and patient engagement have been published [75]. At time of analysis, the biobank contained 48,725 individuals who had been genotyped. DNA samples were genotyped

with genome-wide arrays including the Multi-Ethnic Global (MEGA) array, and the genotype data were imputed into the HRC reference panel [76] using the Michigan imputation server [77]. Imputed data and the 1000 Genome Project data were combined to carry out principal component analysis (PCA) and European ancestry samples were extracted for analysis based on the PCA plot. GTEx v7 models from PredictDB were applied to the samples to calculate genetically regulated expression (GReX).

Phenome-wide association study (PheWAS) was carried out using 'phecodes', phenotypes derived from the International Code for Diseases version 9 (ICD-9) billing codes of EHRs. The PheWAS package for R, version 0.11.2-3 (2017) was used to define case, control and exclusion criteria [78,79]. We required two codes on different visit days to define a case for all conditions, and only phecodes with at least 20 cases were used for analysis (1,531 traits). The single-tissue predicted expressions were combined across tissues using MultiXcan, as was done to analyze individual-level GWAS data from the Psychiatric Genomics Consortium [59]. Covariates for this analysis were age, sex, genotyping array type/batch and three principal components of ancestry.

The top 1% (top 15 traits) of every gene's association results were kept for analysis. A binomial test was used to compare whether the number of traits in any clinical category (circulatory system, genitourinary, endocrine/metabolic, digestive, neoplasms, musculoskeletal, injuries & poisonings, mental disorders, sense organs, neurological, respiratory, infectious diseases, hematopoietic, symptoms, dermatologic, congenital anomalies, pregnancy complications) were over-represented in the top 1% of results compared to the proportion of each category among all 1,531 traits tested. The expected number of each clinical category as determined by [15 traits *

n_{genes}]*p_i where p_i is the probability of a randomly drawn (without replacement) code belongs to category i. P_i can be estimated by the number of codes belonging to category i divided by all codes tested (n=1,531). The significance threshold was 0.05/[17 categories] = 0.0029.

To analyze the overlap between PheWAS results and known Mendelian phenotypes associated with these genes, we used OMIM [80]. "16p11.2" and "22q11.2" were used as search terms and all CNV gene-trait pairs in the region with OMIM entries were used as the list of expected monogenic traits. For each gene-trait pair in OMIM, relevant similar traits (where available) were identified using the phecode catalog [81] and the top p-values for these gene-trait pairs in our PheWAS were selected and shown in Additional file 4: Table S3.

Determining traits over-represented in carriers

3.1 million electronic medical records from the SD at VUMC were queried for keywords corresponding to copy number variants at 16p11.2 and 22q11.2 (Additional file 5: Table S4). Individual charts identified as containing the keywords were manually reviewed and patients were labeled as cases if their medical records provided evidence of CNV carrier status. Patients identified in the queries with insufficient evidence of CNV carrier status were excluded from the analysis. Cases with positive 16p11.2 and 22q11.2 CNV carrier status were identified as: "16p11.2 duplication" (n=48, median age 11), "16p11.2 deletion" (n=48, median age 12), "22q11.2 duplication" (n=43, median age 11). Additional individuals in the 22q11.2 deletion case group were identified by querying the medical records for alternate terms including: "velocardiofacial", "DiGeorge", "conotruncal anomaly face", "Cayler", "Opitz G/BBB",

"Shprintzen", and "CATCH22" (n=388, average age 17). Individuals were excluded from case groups if they were included in the genotyped sample used for the gene-by-gene analysis, or if their records included a mention of additional CNVs. Individuals within the 16p11.2 case groups were also excluded if the size of the reported CNV was 200-250 kb. Individuals within the 22q11.2 case group were excluded if the size of the CNV was smaller than 500 kb or if there was a mention of "distal" when referring to the deletion or duplication. PheWAS was carried out, with each of the four carrier categories as cases and over 700,000 medical home individuals as controls, using age, sex, and self-reported race as covariates. The medical home individuals are patients seen at a Vanderbilt affiliated clinic on five different occasions over the course of three years. Because the sample size for this analysis was larger (700,000 individuals vs. 48,000), and we used traits that were present in 20 or more individuals, there were more traits available for analysis here, n=1,795. After calculating PheWAS, we excluded over-represented traits that were present in <5% of carriers from further analyses.

Comparing gene-specific PheWAS to carrier vs. non-carrier PheWAS

For the first comparison, for each of 16p11.2 duplications, 16p11.2 deletions, 22q11.2 duplications, 22q11.2 deletions, the entire carrier vs. non-carrier PheWAS results were ranked. All the traits in the top 1% of per-gene 16p11.2 and 22q11.2 PheWAS results were converted to a value corresponding to the rank of the trait in the carrier vs. non-carrier PheWAS. To determine whether the per-gene PheWAS top traits were distributed nonrandomly with respect to carrier association, the distribution of the ranks of the each CNV's per-gene PheWAS top traits

was compared to the ranks of all carrier vs. noncarrier PheWAS traits for the same CNV (a uniform distribution) using a one-tailed Wilcoxon rank sum test.

For the second comparison, individuals carrying "extreme" predicted expression across a CNV region were identified using a sequence of rankings. Each expression measurement (i.e. the expression of a single gene in a single tissue in a single individual) was classified as "extreme" if it ranked above the top 2nd percentile or below the bottom 2nd percentile of the BioVU cohort, "normal" if the measurement was between the 25th and 75th percentile, or "neither." For a gene expressed in only one tissue, the gene's "extreme" expression label is simply the same as the tissue's "extreme" label. For a gene with multiple tissue expressions, we counted the number of tissues with "extreme" expression and assigned a gene-level "extreme" label to individuals with the most tissues consistently expressing "extremes" for the gene. A gene-level "normal" label was assigned to half of the cohort who had no extreme-expression in any tissues and had the most tissues with "normal" expressions. The remaining individuals received a "neither" label for the gene. After obtaining the gene-level labels ("extreme", "normal", "neither"), we then ranked the individuals by the number of "extreme" expression genes, and labeled a subset of individuals (top 2% of the 48,600 individuals) as extreme-expression carriers. Note that we consider extreme high and extreme low predictions together due to prior data showing that eQTL direction can be specific to cell-types or tissues, which our cross-tissue approach cannot distinguish [78]. These were compared to a "control" group defined for each CNV region that included individuals with the fewest extreme-expressed genes and most "normal" expression genes who comprised about half of the cohort. PheWAS was performed to identify over-represented traits between the extreme-expression and control groups, analogously to the carrier vs. non-carrier PheWAS. The top 10% most associated traits in each category (16p11.2 extreme, 22q11.2 extreme) were assigned a value corresponding to the rank of the traits in the carrier vs. non-carrier association results, treating deletion and duplication CNV carrier traits separately. We used a one-tailed

Wilcoxon rank sum test to test whether the top 10% traits of each extreme category tend to have a shifted distribution for association with the (corresponding) carrier status (16p11.2 duplications and deletions for 16p11.2 extremes, 22q11.2 duplications and 22q11.2 deletions for 22q11.2 extremes).

Significance threshold for association studies

The significance threshold used for each discovery MultiXcan or S-MultiXcan association study and conditional analysis was 0.05/(number of traits*number of CNV genes tested). In practice, this usually meant 5 traits and 127 CNV genes, for a threshold of $P < 7.9 \times 10^{-5}$. For replication studies, the significance threshold was set at 0.05 in order to test a single gene. The exception was in the BMI UK Biobank dataset. We first tried a phenotype-swapping approach to generate an expected distribution for the 16p11.2 genes. The distributions were null and did not yield meaningful comparisons. Instead, 100 random subsets of adjacent genes of approximately the same length and gene count as the CNV were tested for association with BMI. The 95th percentile of the MultiXcan p-values for these genes was used as a permutation-based significance threshold.

In the gene-based PheWAS study, there were 1,531 phecodes (each with at least 20 cases) tested overall, corresponding to a Bonferroni-corrected phenome-wide significance threshold of 3.3×10^{-5}. For genes having no phenome-wide significant results, their top 15 associations, corresponding to the top 1% of the 1,531 phecodes, were used. In the carrier vs. non-carrier PheWAS, there were 1,795 phecodes tested overall, corresponding to a Bonferroni-corrected phenome-wide significance threshold of 2.79×10^{-5}. Additional traits meeting a false discovery rate threshold of

0.05 were considered in identifying traits both over-represented in carriers and represented in individual gene PheWAS.

Graphical summary of selected PheWAS results

The *chordDiagram* method in the *circlize* package was used to generate the circle summary plots [82]. The gene-trait pairs we selected for Tables 2.1 and 2.2 were used as inputs, with the -log10 p-value of association used as the weighting to determine the edge width. For the 22q11.2 circle plot, only associations with $P < 5x10^{-3}$ were used in order to create a legible plot. Descriptions were cut off at 55 characters; to read the entire descriptions see Tables 2.1 and 2.2.

RESULTS

Individual genes at 16p11.2 are associated with schizophrenia, IQ, and BMI

In order to find genes at copy number variant loci driving brain-related disorders, we performed an association analysis between imputed gene expression levels and five traits: schizophrenia, IQ, BMI, bipolar disorder, and ASD. It has been observed that copy number variants (including 16p11.2 and 22q11.2) affect expression of nearby genes [45,46,83]. As flanking genes affected by copy number variation may be relevant to phenotype, we additionally considered genes 200kb in each direction from each CNV [84]. Overall, we tested 52 coding and 30 noncoding genes at or near 22q11.2 and 37 coding and 8 noncoding genes at or near 16p11.2 for which a predictive model was available (Additional file 2: Table S2, Additional file 3: Fig. S1). As cis-eQTLs are

often shared among tissues, we pooled together information from all tissues in GTEx to boost our power to detect brain-related traits [59].

Two genes at 16p11.2 show predicted expression positively associated ($P < 7.9x10^{-5}$) with schizophrenia (**Figure 2.2**; Additional file 6: Table S5): *TMEM219* ($P = 1.5x10^{-5}$) and *INO80E* ($P = 5.3x10^{-10}$). This positive direction of effect is consistent with the association between 16p11.2 duplications and schizophrenia [2]. An additional gene, *YPEL3*, was significantly associated with schizophrenia in the negative direction ($P = 4.9x10^{-6}$). For IQ, there was one strong positive association at the 16p11.2 locus (**Figure 2.2**; Additional file 6: Table S5): *SPN* ($P = 2.9x10^{-22}$). Intellectual disability is observed in both deletions and duplications of 16p11.2, so there was no expected direction of effect [3,14]. Four genes showed negative association with BMI (**Figure 2.2**; Additional file 6: Table S5): *SPN* ($P = 6.2x10^{-18}$), *TMEM219* ($P = 2.2x10^{-5}$), *TAOK2* ($P = 8.5x10^{-11}$), and *INO80E* ($P = 1.0x10^{-7}$). We focused on genes with negative associations with BMI because, in humans, obesity is associated with deletions at 16p11.2 [1,17]. Two additional genes, *KCTD13* ($P = 9.5x10^{-6}$) and *MVP* ($P = 2.1x10^{-5}$), were significantly associated with BMI in the positive direction. No gene at 16p11.2 was significantly associated with bipolar disorder or ASD (Additional file 6: Table S5, Additional file 3: Fig. S3). No individual genes at or near 22q11.2 had predicted expression significantly associated with any of the five traits (Additional file 6: Table S5, Additional File 3: Fig. S4).

Follow-up conditional analyses narrow down genes driving schizophrenia and BMI

To replicate our analysis, we used a large cohort from the UK Biobank for which GWAS summary statistics were available for multiple brain-related traits (Additional file 1: Table S1) [67]. The predicted expression of *INO80E* and *TMEM219* from the discovery analyses were associated ($P < 0.05$) with having an ICD10 diagnosis of schizophrenia (ICD10: F20, 198 cases: *INO80E* $P = 0.04$, *TMEM219* $P = 0.03$, Additional file 7: Table S6). Although this is only nominally significant, it is notable that these genes are in the 3rd percentile of schizophrenia associations genome-wide within UK Biobank.

The UK Biobank GWAS of BMI is highly inflated, including in the 16p11.2 region. Nearly every 16p11.2 gene showed association at the previously used threshold ($P < 7.9 \times 10^{-5}$). Using a permutation-based approach within individual-level data, we adjusted the significance threshold to 8.8×10^{-11}. All genes from the discovery analysis replicated (Additional file 7: Table S6): *SPN* ($P = 6.1 \times 10^{-23}$), *KCTD13* ($P = 1.2 \times 10^{-30}$), *TMEM219* ($P = 7.1 \times 10^{-37}$), *MVP* ($P = 5.1 \times 10^{-11}$) *INO80E (*$P = 1.9 \times 10^{-27}$). We were not able to replicate the IQ result in the UK Biobank, because the UK Biobank sample overlapped with our discovery GWAS.

We performed an additional fine-mapping study on the three genes associated with schizophrenia. Linkage disequilibrium between the eQTL SNPs in predictive models may lead to correlation among predicted expressions for nearby genes, so it is possible that not all three detected association signals are independent. The predicted expressions of *INO80E, YPEL3*, and *TMEM219* were moderately correlated (the correlation of *INO80E* with the other genes is in the

range of -0.4 to 0.37 across GTEx tissues, for example), consistent with the relationships between the observed expressions of these genes (measured expression of *INO80E* is correlated with measured expression of the other genes in the range -0.36 to 0.31). In order to pick out the gene(s) driving the association signal, we used a conditional analysis approach (Additional file 8: Table S7). We observed that after adjusting the predicted expression of the other CNV genes for the predicted expression of *INO80E*, no gene was significantly associated with schizophrenia. However, when we adjusted the predicted expression of *INO80E* by the predicted expressions of the other two highly associated genes, *INO80E* remained significantly associated with schizophrenia ($P = 2.3\text{x}10^{-6}$). The same pattern was not observed for *TMEM219* or *YPEL3*, suggesting *INO80E* explains the entire 16p11.2 signal for schizophrenia.

While we did not have individual level data for the GIANT consortium, we obtained individual-level BMI data from the UK Biobank [69]. We performed an analogous conditional analysis on the six genes associated with BMI, *SPN, INO80E, TMEM219, TAOK2* in the negative direction, as well as *KCTD13* and *MVP* in the positive direction. Due to the inflation in the UK Biobank data, all these genes had very low p-values even after conditioning; however, we see that some genes' association results stayed in the same range, while others increased in p-value by five orders of magnitude or more after adjusting by the other five genes. Based on these observations, it is likely that *SPN* ($P_{UKBB} = 6.1\text{x}10^{-23}$, $P_{cond} = 7.5\text{x}10^{-21}$), *INO80E* ($P_{UKBB} = 1.9\text{x}10^{-27}$, $P_{cond} = 2.8\text{x}10^{-32}$), and *KCTD13* ($P_{UKBB} = 1.2\text{x}10^{-30}$, $P_{cond} = 4\text{x}10^{-27}$) were independently associated with BMI, while *TMEM219* ($P_{UKBB} = 7\text{x}10^{-37}$, $P_{cond} = 2.3\text{x}10^{-18}$), *TAOK2* ($P_{UKBB} = 4.2\text{x}10^{-29}$, $P_{cond} = 2.3\text{x}10^{-19}$), and *MVP* ($P_{UKBB} = 5.1\text{x}10^{-11}$, $P_{cond} = 5\text{x}10^{-6}$) were significant in the discovery analysis primarily due to correlation with one of the independent genes.

To validate that our approach explained all GWAS signal at the locus, we took two phenotypes in which we had both GWAS signal and individual level data available – PGC Schizophrenia and UK Biobank BMI – and conditioned the MultiXcan analysis on lead GWAS SNP(s) in those datasets that were also eQTLs. In schizophrenia (where *INO80E* is our proposed sole driver gene), conditioning on one GWAS SNP (rs4788200, GWAS $P = 2.8 \times 10^{-10}$) was sufficient to explain the GWAS peak in the region (Additional file 3: Fig. S2). Conditioning the MultiXcan analysis on this SNP successfully removed all association signal, including for *INO80E* (Additional file 3: Fig. S2). In BMI, (where we propose three independent genes, *INO80E, KCTD13, SPN*) conditioning on four GWAS/eQTL SNPs was sufficient to explain both the GWAS and MultiXcan signal (Additional file 3: Fig. S2). These were rs4787491 (GWAS $P = 7.6 \times 10^{-17}$), rs9936474 (GWAS $P = 5.1 \times 10^{-31}$), rs2008514 (GWAS $P = 3.3 \times 10^{-29}$), and rs8046707 (GWAS $P = 3.2 \times 10^{-19}$). The first two SNPs explain the GWAS signal within the region, and the latter two come from more distal GWAS peaks that are nevertheless involved in the expression prediction of 16p11.2 genes; as a result, four SNPs are needed to fully nullify the MultiXcan signal. The schizophrenia variant rs4788200 is not a strong eQTL for any gene-tissue pair, but it appears in the models for *INO80E* in 22/37 tissues where *INO80E* has models. Similarly, one of the BMI SNPs, rs4787491 is an expression-decreasing eQTL for *INO80E* in 35/37 tissues and is generally strong: the distribution of weights of this SNP was significantly different from the distribution of all *INO80E*-predicting SNPs, ($P = 4.8 \times 10^{-13}$, Kolmogorov-Smirnov test). We conclude that our approach is sufficient for explaining GWAS signal and that the multi-SNP predictive models involving both nearby and more distal SNPs are advantageous.

Phenome-wide association studies identify previously known and novel traits associated with 16p11.2 and 22q11.2 carrier status

While GWAS datasets provide insight into the impact of genes on ascertained brain-related traits, the 16p11.2 and 22q11.2 CNVs may contribute to a wide spectrum of traits, including milder manifestations of brain-related traits. Thus, biobanks containing both genetic and clinical data can tell us about broader clinical impacts on medical traits. We queried the de-identified electronic health records for 3.1 million patients at VUMC to explore the impacts of the 16p11.2 and 22q11.2 CNVs, as well as their individual genes, on the medical phenome in a representative population [75]. CNV diagnoses are documented in the medical records, which led us to ask: what are the specific clinical phenotypes that are common in individuals identified as 16p11.2 or 22q11.2 CNV carriers? Carriers were identified by diagnosis of 16p11.2 or 22q11.2 deletion/duplication (or syndromic names for 22q11.2, see methods) in their medical record, and over 700,000 individuals were used as controls. We performed a phenome-wide association study (PheWAS) between 16p11.2 and 22q11.2 deletion/duplication carriers and controls against 1,795 medical phenotype codes (Figs 3 and 4) [78,81]. Traits that were significantly over-represented in carriers ($P < 2.8\text{x}10^{-5}$) fell into three major categories: (1) known primary CNV clinical features, including possible reasons for the referral of the patient for genetic testing (i.e. neurodevelopmental concerns, epilepsy, congenital heart defects), (2) secondary CNV features known to be present in carriers but unlikely to be a primary reason for referral for genetic testing,

(3) novel diagnoses not previously reported (Fig. 3, Fig. 4, Additional file 9: Table S8). We chose to focus on traits present in at least 5% of carriers to avoid over-interpreting rare traits.

16p11.2 deletion carrier status was associated with developmental diagnoses (**Figure 2.3**): *lack of normal physiological development* ($P = 2.8\text{x}10^{-18}$), *developmental delays and disorders* ($P= 6.3\text{x}10^{-10}$), *delayed milestones* ($P = 1.4\text{x}10^{-11}$) [3]. In addition, 16p11.2 deletion carrier status was associated with *autism* ($P = 1.3\text{x}10^{-10}$) and *mental retardation* ($P = 7.9\text{x}10^{-13}$) [5]. The digestive diagnosis of *GERD* ($P = 1.1\text{x}10^{-5}$) has been previously observed in carriers but was unlikely to be a primary reason for genetic testing [85]. *GERD* was accompanied by other digestive diagnoses such as *dysphagia* ($P = 1.3\text{x}10^{-7}$) and *diseases of esophagus* ($P = 4.3\text{x}10^{-7}$). *Muscle weakness* ($P = 2.8\text{x}10^{-6}$) and *abnormal movements* ($P = 3.9\text{x}10^{-6}$) are consistent with neurological traits reported in 16p11.2 deletion carriers such as hypotonia and motor impairments [86]. *Sleep apnea* ($P = 8.9\text{x}10^{-6}$) was a novel phenotype, potentially related to increased BMI in deletion carriers.

16p11.2 duplication carrier status was similarly associated with developmental diagnoses (**Figure 2.3**): *lack of normal physiological development* ($P = 5.6\text{x}10^{-15}$), *developmental delays and disorders* ($P = 2.5\text{x}10^{-13}$), *delayed milestones* ($P = 9.0\text{x}10^{-13}$), *autism* ($P = 1.3\text{x}10^{-12}$), and *mental retardation* ($P = 1.6\text{x}10^{-7}$) [3,5]. 16p11.2 duplication carriers status was also associated with multiple heart defects, including *valvular heart disease/heart chambers* ($P = 4.6\text{x}10^{-10}$) and *cardiac shunt/heart septal defect* ($P = 3.2\text{x}10^{-8}$), both of which have been reported previously [87]. 16p11.2 duplications are known to be a risk factor for epilepsy, and were associated with an epilepsy-related diagnosis of *convulsions* ($P = 2.9\text{x}10^{-8}$) in the biobank [3,88]. *Infantile*

cerebral palsy ($P = 4.9x10^{-6}$), while a potential reason for genetic testing, has not previously been associated with 16p11.2 duplications. While the 16p11.2 CNV contains genes such as *SPN* and *MVP* that are active in the immune system, there is no prior evidence of the susceptibility of duplication carriers to infection, making the diagnosis *Bacterial infection NOS* ($P = 5.5x10^{-7}$) a novel finding.

For 22q11.2 deletion carriers, the canonical associated features were cardiac defects such as *cardiomegaly* ($P = 3.5x10^{-258}$) and *cardiac shunt/heart septal defects* ($P = 4.7x10^{-285}$) (**Figure 2.4)** [6,7]. Other highly associated diagnoses were developmental: *lack of normal physiological development* ($P = 1.7x10^{-47}$)*, developmental delays and disorders* ($P = 6.3x10^{-29}$)*, delayed milestones* ($P = 6.0x10^{-11}$) [6,7]. Congenital anomalies such as *cleft palate* ($P = 9.4x10^{-80}$) were also over-represented. The secondary known traits for 22q11.2 deletion carriers included *immunity deficiency* ($P < 10^{-285}$), and *disorders involving the immune mechanism* ($P < 10^{-285}$). Previously, it has been reported that 50% of 22q11.2 deletion carriers have T-cell dysfunction and 17% have humoral dysfunction [7]. Very few traits over-represented in 22q11.2 deletion carriers were novel; one of these was *hyperpotassemia* ($P = 1.4x10^{-10}$).

22q11.2 duplication carrier status was also associated with developmental diagnoses (**Figure 2.4**): *delayed milestones* ($P = 1.1x10^{-13}$), *lack of normal physiological development* ($P = 9.7x10^{-13}$)*, pervasive developmental disorders* ($P = 1.2x10^{-6}$) [8]. 22q11.2 duplication status was associated with cardiac phenotypes such as *cardiac shunt/ heart septal defect* ($P = 2.3x10^{-5}$). Cardiac features have not as often been reported in 22q11.2 duplication carriers compared to 22q11.2 deletion carriers [8]. Remaining traits such as *abnormality of gait* ($P = 3.1x10^{-12}$) and

hearing loss ($P = 2.1x10^{-7}$) have also been seen in 22q11.2, including as indications for genetic testing [89].

Phenome-wide association studies identify phenotypic consequences of expression variation in 16p11.2 and 22q11.2 genes

As our study of the impact of the entire CNV on phenotype confirmed our ability to detect important CNV-associated traits within the BioVU biobank, our next goal was to catalogue how each individual CNV gene might affect the medical phenome. We generated predicted expression for CNV and flanking genes, as in the initial GWAS analyses, for the 48,630 non-CNV-carrier individuals genotyped in BioVU. We tested 1,531 medical phenotypic codes meeting frequency criteria (n = 20 cases) in this subset. There were six phenome-wide significant ($P < 3.3x10^{-5}$) gene-trait associations at 16p11.2 including: *INO80E* with *skull and face fracture and other intercranial injury* ($P = 1.9x10^{-15}$), *NPIPB11* with *psychosis* ($P = 1.0x10^{-5}$), and *SLX1B* with *psychosis* ($P = 3.0x10^{-5}$). There were eleven phenome-wide significant gene-trait associations at 22q11.2 including: *AIFM3* with *renal failure* ($P = 2.3x10^{-5}$), *LZTR1* with *malignant neoplasm, other* ($P = 1.4x10^{-5}$), *SCARF2* with *mood disorders* ($P = 1.3x10^{-5}$), *PI4KA* with *disorders of iris and ciliary body* ($P = 1.1x10^{-7}$) and *disorders resulting from impaired renal function* ($P = 2.2x10^{-5}$). These include two renal traits, consistent with the 22q11.2 deletion carrier status association with *renal failure.* The associations of *LZTR1* and *PI4KA* with neoplasms and eye disorders correspond to similar traits associated with these genes in prior literature [90–92].

Previously established gene-trait associations came up as suggestive (top 1 percentile), although not phenome-wide significant, associations in the BioVU cohort. *TBX1*, a gene at 22q11.2 tied to heart development, had *other chronic ischemic heart disease, unspecified* ($P = 0.001$), *endocarditis* ($P = 0.0046$), *cardiomyopathy* ($P = 0.0055$), and *coronary atherosclerosis* ($P = 0.0076$) among its top 1% phenome associations [29–32]. *TBX6* at 16p11.2, which has a role in bone development and scoliosis, has *pathologic fracture of vertebrae* in its top 1% phenome associations ($P = 0.0028$) [93–95]. *TANGO2* mutations at 22q11.2 have been associated with metabolic abnormalities such as hypoglycemia, as well as epilepsy, and our PheWAS for *TANGO2* showed *abnormal glucose* ($P = 0.0013$) and *epilepsy, recurrent seizures, and convulsions* ($P = 0.0049$) as top phenotypes [96,97]. We identified additional genes at 16p11.2 and 22q11.2 that are associated with Mendelian traits, using OMIM [80], and browsed our PheWAS for potentially similar clinical traits, including those not meeting the top 1 percentile threshold. We find that of 13 such genes, 7 have a relevant clinical trait at $P < 0.05$, and 12 at $P < 0.1$. In 6 of the 13 genes, the relevant clinical traits are within the top 1% of PheWAS associations for the gene (Additional file 4: Table S3).

As few gene-trait pairs reached phenome-wide significance and established associations were present at more nominal levels, we also considered traits that did not meet the significance threshold in our analysis but were in the top 1% of phenotypic associations for a given gene (Additional file 10: Table S9). We found that traits categorized as “mental disorders” were over-represented in the top 1% of the phenome of CNV genes ($P = 5.2 \times 10^{-5}$). Of all 17 clinical categories tested, “mental disorders” was the only category with enrichment p-value meeting multiple testing thresholds (Additional file 11: Table S10). This suggested that the effect of CNV

genes is more widespread on brain-related traits than simply those detected as statistically significant.

Some of the top 1% PheWAS traits for CNV genes overlapped with the original five traits we studied: schizophrenia, IQ, BMI, bipolar disorder, and ASD. At 16p11.2, there were genes whose top PheWAS results included schizophrenia-related traits (*psychosis, schizophrenia and other psychotic disorders),* IQ-related traits (*developmental delays and disorders, mental retardation, delayed milestones*), BMI-related traits (*bariatric surgery, morbid obesity*), and ASD-related traits (*pervasive developmental disorders*) (**Table 2.1, Figure 2.5**). At 22q11.2, there were genes whose top PheWAS results included schizophrenia-related traits (*hallucinations*), BMI-related traits (*overweight, obesity and other hyperalimentation, morbid obesity*), ASD-related traits (*autism, speech and language disorder*), and bipolar-related traits (*mood disorders*) (**Table 2.2, Figure 2.5**). We could not perform strict independent replication for these associations because many of these traits are difficult to define in the same way across datasets (for example *Speech and language disorder* vs. *Autism*). Instead, we compared the top association statistics within our GWAS discovery and replication datasets for the genes identified to be associated with brain-related traits in PheWAS as an extension of this study (Additional file 8: Table S7). The following genes were associated at $P < 0.05$ and also in the top 5th percentile within at least one of the GWAS discovery or replication datasets (Additional file 8: Table S7): *SEPT1* (*psychosis* – in UK Biobank schizophrenia 20002_1289 $P = 0.03$), *AIFM3* (*mood disorders* – in UK Biobank bipolar F31 $P = 0.04$), *SCARF2* (*mood disorders* – in UK Biobank bipolar F31 $P = 0.003$), *HIC2* (*mood disorders* – in UK Biobank bipolar 20002_1991, $P = 0.004$), *ZNF48* (*bariatric surgery* –

in UK Biobank BMI $3.7x10^{-6}$). Of these, the association between *SCARF2* and *mood disorders* reached phenome-wide significance in the PheWAS.

Predicted expression may be correlated between nearby genes, thus multiple genes can share a PheWAS trait association due to correlation alone. We are underpowered for independence testing for the majority of our GWAS traits, but we selected several notable traits that appeared in multiple genes to test for independence, in the same way as in our GWAS analysis (Additional File 7: Table S6). We performed a conditional analysis on 16p11.2 genes whose top phenome associations included *psychosis: NPIPB11, BOLA2, MAPK3, SEPT1, SLX1B, TBC1D10B.* By comparing whether the p-value of association stayed constant vs. increased after conditioning, we found that *NPIPB11, SEPT1, SLX1B, TBC1D10B* were likely independent associations, whereas *BOLA2* and *MAPK3* may be associated with *psychosis* at least partly by correlation with the other four. We also performed the same analysis for 22q11.2 genes whose top phenome associations included *morbid obesity*: *SNAP29, P2RX6, P2RX6P.* Of these genes, the only one with a p-value increase was *P2RX6P*, suggesting that its association with *morbid obesity* may be explained at least in part by another gene. From conditional analysis, we see evidence of a multigenic contribution to both traits from CNV genes.

Genes in 16p11.2 and 22q11.2 are associated with traits that are also over-represented in carriers

We originally hypothesized that small variations in CNV gene expression would be associated with phenotypes resembling those that were present in CNV carriers, perhaps with smaller effects. Our use of electronic health records first on the entire CNV itself, then on individual

genes allows us to detect these potential effects across traits. Unlike the five brain-related traits that we originally chose, many of the traits in the EHRs do not have similar large GWAS datasets available. Considering that our non-ascertained biobank is not well-powered for less common traits, we chose to focus on the top one percentile of the phenome associations rather than the few associations that passed the phenome-wide significance threshold.

Traits that were found both in 16p11.2 carriers and in individual genes' PheWAS results included primary CNV traits such as *mental retardation* and *delayed milestones*, as well as secondary traits such as *dysphagia* and *convulsions* (**Table 2.1, Figure 2.5**). There were six genes (*ASPHD1, FAM57B, ALDOA, TBX6, MAPK3, SULT1A3)* whose top PheWAS associations included the 16p11.2 deletion-associated trait of *upper gastrointestinal congenital anomalies*, though we are underpowered to know whether all these signals are independent. Of the genes that we found as drivers in the first analysis of GWAS datasets, we note that *INO80E*'s top PheWAS results overlap the 16p11.2 deletion-associated trait *other specified cardiac dysrhythmias* and *SPN*'s top PheWAS results overlap the 16p11.2 duplication-associated trait of *failure to thrive (childhood*).

Over 30 genes at 22q11.2 had a top PheWAS trait overlapping a trait over-represented in 22q11.2 duplication or deletion carriers (**Table 2.2, Figure 5.5**). Top PheWAS results for 22q11.2 genes included primary cardiac traits such as *tachycardia* (*P2RX6P, GNB1L*) and primary brain-related traits such as *autism* (*TANGO2, ZDHHC8).* We also found genes with top PheWAS results overlapping secondary traits from the carrier screen, such as *diseases of the larynx and vocal cords* (*DGCR6, PRODH, ARVCF*).

It is difficult to meaningfully compare the carrier screen to the gene-based PheWAS results because the effects of modest expression variation in an individual gene are not necessarily expected to be the same as those of the deletion or duplication of an entire locus. We tested whether the top associations from individual gene PheWAS results were enriched for EHR phenotypes over-represented in carriers. We did this by analyzing where top PheWAS traits associated with CNV genes were ranked within PheWAS results of carrier status. We found no evidence for enrichment in 16p11.2 duplications, 16p11.2 deletions, 22q11.2 duplications, or 22q11.2 deletions (Additional file 3: Fig. S3). As an alternate way to compare the two PheWAS approaches by 'mimicking' the CNV effects, we identified individuals in the genotyped cohort in BioVU that had the most extreme (2^{nd} percentile) predicted expression across CNV genes in a region and were thus the most similar we could identify to true CNV carriers (see Methods). The top 10% of traits over-represented in this "extreme expression non-carrier" group were examined for their distribution within ranked (by p-value) lists of traits in CNV carriers. We found that in all four cases (16p11.2 deletions, 16p11.2 duplications, 22q11.2 deletions, 22q11.2 duplications), the top traits in the "extreme expression non-carrier" group were more likely to rank near the top of the CNV carrier traits than would be expected by chance; the distribution was significantly shifted for 22q11.2 genes ($P = 8.9 \times 10^{-15}$, mean rank 487/1795, 22q11.2 deletions; $P = 6.1 \times 10^{-8}$, mean rank 563/1784, 22q11.2 duplications; $P = 0.18$, mean rank 770/1816, 16p11.2 deletions; $P = 0.45$, mean rank 805/1816, 16p11.2 duplications; Additional file 3: Fig. S6) . These results demonstrate that within the same EHR system, expression prediction based on common SNPs independently shows enrichment for CNV carrier associated traits.

DISCUSSION

In this study, we sought to identify individual genes in the 16p11.2 and 22q11.2 regions driving brain-related disorders, as well as the impact of both the entire CNV and specific CNV genes on the medical phenome. In a novel *in-silico* approach to CNV fine-mapping, we tested whether genetically-driven predicted expression variation of the individual genes in each CNV was associated with ascertained brain-related disorders ascertained in GWAS data. We identified individual genes at 16p11.2 whose expression was associated with schizophrenia (*INO80E*), IQ (*SPN*), and BMI (*SPN, INO80E*) in the expected direction based on known 16p11.2 biology. We then used EHR data to detect (known and novel) traits overrepresented in 16p11.2 and 22q11.2 carriers for comparison with individual gene results. Third, we used the same EHR system biobank containing over 1,500 medical traits to explore the consequences of expression variation of 16p11.2 and 22q11.2 CNV genes in non-carriers, and we identified enrichment of brain-related traits as well as individual genes potentially driving carrier-associated traits. The results from the GWAS-derived and PheWAS analyses can be considered as independent ways to probe the function of CNV genes using expression imputation.

INO80E, the gene we identified as a driver of schizophrenia and BMI, is a chromatin remodeling gene and has rarely been considered in the context of brain-related traits [98]. Mice heterozygous for this gene have shown abnormal locomotor activation [99]. Locomotor activity in mice is a frequently used proxy for brain-related disorders including schizophrenia [100]. Our results are consistent with a previous observation that eQTLs from dorsolateral prefrontal cortex for *INO80E* co-localize with schizophrenia GWAS SNPs [101]. In addition, an analogous imputed expression based transcriptome-wide association study observed association between *INO80E*

and schizophrenia using summary statistics [102]. A third transcriptomic association study using prenatal and adult brain tissues also pointed to *INO80E* as a risk gene for schizophrenia [103]. By focusing on a specific schizophrenia-associated region, using individual level data, and performing a conditional analysis, we have obtained additional precision, and were able to fine-map the signal at 16p11.2 down to a single gene. Our study differs from Gusev *et al* and Walker *et al* in the expression prediction models used: we used 48 tissue models from the Genotype-Tissue Expression consortium, Gusev *et al* used brain, blood, and adipose tissues from other consortia, and Walker *et al* used prenatal and adult brain tissues only. The overlap in association results shows that our approach is robust to variation in predictive models. Furthermore, we find that the utilization of non-brain tissues in our analysis did not hinder our ability to detect this association. Mice with a heterozygous mutation in *Ino80e* showed increased body weight, consistent with our BMI association result for the same gene [99].

SPN, a gene highly associated with both IQ and BMI, is active in immune cells and is not known to play a role in brain-related disorders [104,105]. Recently, a large genome-wide analysis of rare CNVs fine-mapped *SPN* duplications as a driver of several phenotypic categories including *behavioral abnormality*[106]. We note that the association p-values for *SPN* are much lower than for any other genes showing association signal. This may be because our approach detected relatively few eQTLs for *SPN* (12 SNPs in two tissues), many of which overlapped with highly associated GWAS SNPs for both IQ and BMI, rather than contributing to noise.

Our results give evidence that pleiotropy is involved in the pathogenicity of 16p11.2, as opposed to a strictly “one gene, one trait” model. Specifically, *INO80E* was associated with both

schizophrenia and BMI, and *SPN* was associated with both BMI and IQ. Genetic correlations of at least -0.05 and as much as -0.5 have been estimated for the BMI/IQ and SCZ/BMI pairs, suggesting that pleiotropy may play a general role in these disorders [107–110]. Consistent with the genetic correlations, most (8/12) eQTL SNPs in our prediction models for *SPN* drove the associations with both IQ and BMI.

While most associations we detected were in the expected direction given previous knowledge, *MVP* and *KCTD13* were associated with BMI in the opposite (positive) direction, and *YPEL3* with schizophrenia in the negative direction. We resolved the schizophrenia result by conditional analysis, where we found that *YPEL3* was associated with schizophrenia simply due to correlation with *INO80E*. For BMI, we were able to use UK Biobank data to determine that *MVP* was not an independent association with BMI, while *KCTD13* remained. For an example like *KCTD13,* we offer three explanations: these results may be false-positives due to correlation-based "hitchhiking", they may demonstrate a limitation of our approach, or they may have a true BMI-increasing effect. First, we cannot rule out that it "hitchhikes" to statistical significance with other negatively-associated genes due to correlation but does not contribute to BMI itself. Second, this result might represent a limitation of our eQTL-based method. *KCTD13* is a highly brain-expressed gene, but had no high-quality brain prediction models [51]. The direction of the eQTLs regulating *KCTD13* expression in the brain may be brain-specific, and brain may be the only relevant tissue for the effect of *KCTD13* on BMI. That is, *KCTD13* may have a strong negative correlation with BMI, but falsely appears positive due to the specific eQTLs used for expression prediction. Such tissue-specific eQTL directions of effect have been observed for at least 2,000 genes [111]. Improved brain-specific prediction models will resolve this limitation.

Third, *KCTD13* could have a true BMI-increasing effect. If so, the 16p11.2 region contains both BMI-increasing and BMI-decreasing genes, and the effect of the BMI-decreasing genes is stronger. Such a model is a potential explanation for the observation that duplications at 16p11.2 in mice, unlike humans, are associated with obesity [44]. One set of genes may be the more influential determinant of the obesity trait in each organism.

Our PheWAS of traits overrepresented in 16p11.2 and 22q11.2 carriers served as a validation of our biobank EHR approach via detection of previously identified CNV-associated traits. Brain-related traits, such as *delayed milestones, mental retardation* and *pervasive developmental disorders*, were among the top over-represented traits in both 16p11.2 and 22q11.2 CNV carriers. 22q11.2 deletion carriers were strongly associated with *cardiac congenital anomalies* and *cleft palate*, two of the hallmark features of the CNV. Even though the total number of CNV carriers within the biobank was relatively small, the strong known clinical associations were observed. At the same time, we identified novel traits that may be confirmed in larger samples of CNV carriers such as *sleep apnea* in 16p11.2 deletions and *hyperpotassemia* in 22q11.2 deletions.

Our PheWAS between the predicted expressions of 16p11.2 and 22q11.2 genes and 1,500 medical phenotypic codes resulted in 17 phenome-wide significant gene-trait pairs. Some of these genes have been shown to drive similar traits in prior literature. The gene *AIFM3* at 22q11.2 was associated with *renal failure. AIFM3* is a gene in a proposed critical region for 22q11.2-associated kidney defects, and led to kidney defects in zebrafish [112]. *SNAP29,* another gene associated with kidney defects in the same study, had *renal failure, NOS* in its top 1%

phenome associations. *LZTR1* was significantly associated with *malignant neoplasm, other.* This gene is a cause of schwannomatosis, a disease involving neoplasms (albeit normally benign) [90]. Model organisms with defects in *PI4KA,* associated with *disorders of iris and ciliary body* in our study, showed eye-related phenotypes [91,92]. Because few genes had any associations which were phenome-wide significant, we elected to analyze the top 1% of associations of each gene. We noticed that our gene-by-gene PheWAS recapitulated known Mendelian effects of approximately half of Mendelian genes at the 16p11.2 and 22q11.2 CNVs, including the effect of *TBX1* on the circulatory system, of *TANGO2* on glucose and epilepsy, and of *TBX6* on the musculoskeletal system at this threshold [29–31,93–95]. There are three common SNPs at *TBX6* contributing to scoliosis (primarily in individuals who have additional disruptive mutations at the gene), and one was identified as an eQTL in our approach; perhaps an even stronger signal could have been observed if all three were included[113]. Notably, we found that clinical traits in the *mental disorders* category were over-represented in the top 1% of associations among all genes tested, and *mental disorders* was the only category significantly enriched. Some mental disorders, such as *psychosis*, were top PheWAS hits for multiple genes, but we were underpowered for rigorous independence testing. Moreover, three novel brain-related gene-trait pairs reached phenome-wide significance: *NPIPB11* and *SLX1B* near the CNV breakpoint at 16p11.2 with *psychosis*, as well as *SCARF2* at 22q11.2 with *mood disorders.* The expression of *SLX1B* is modified in 16p11.2 carriers; *NPIPB11* expression differences have not been detected in transcriptomic studies of 16p11.2 [44,46]. *SCARF2* has recently been proposed as a driver of schizophrenia within a fine-mapping study within CNV carriers[106]. Integrating genetic information with the diagnosis of *mood disorders* in the clinical data allowed us to find a new

candidate, *SCARF2*, at 22q11.2 that we were unable or underpowered to detect in the ascertained bipolar data alone.

We find that our results support the underlying hypothesis in which small changes in CNV gene expression affect risk for CNV-associated traits. In the three best-powered traits we had available – schizophrenia, BMI, and IQ – we were clearly able to prioritize individual gene(s) at 16p11.2. Similarly, we were able to detect PheWAS traits driven by small expression differences in CNV genes that were overlapping with traits in CNV carriers in the same biobank. Strikingly, we found that our gene-based PheWAS overlapped well with the carrier screen PheWAS for 22q11.2 when we found the most "CNV-like" extreme expression non-carriers. This observation validates our underlying model in which non-carriers with genetically-predicted expression differences are more likely to show carrier-like traits.

Limitations

The 16p11.2 and 22q11.2 CNVs are significant risk factors for ASD and schizophrenia, respectively, and yet no individual genes in either CNV were associated with case-control status for the associated trait in the best-powered datasets available to us. Assuming the true causal gene(s) for these disorders do exist within the CNV, limitations in our approach may preclude us from discovering them. As our predicted expressions are based on GWAS data, we end up underpowered to detect gene-based association signal where we are underpowered to detect SNP-based association signal. This is particularly true for ASD, in which the sample size is over 4 times less than that of schizophrenia. At the same time, predictive models for gene expression

are imperfect; while they capture some of the *cis*-heritability of gene expression, they may not capture the entire variability of the expression of a gene (the largest single-tissue prediction R^2 for our genes is 0.45, and the average R^2 is 0.07). For example, the expression predictions of these genes are calculated solely using *cis*-eQTLs within 1MB of the gene [58]. It may be necessary to consider the effect of *trans*-eQTLs to explore the genetic effect of expression variation accurately. Similarly, we have not considered *trans*-effects due to chromosome contacts, such as those that exist between the 16p11.2 region described here and another smaller CNV region elsewhere at 16p11.2 [114,115]. Moreover, there are genes in both regions for which no high-quality models exist. If the causal gene is among the genes that cannot be well-predicted, we cannot detect this gene by our approach. One category of genes that are not represented in our study are microRNAs. 22q11.2 carriers have a unique microRNA signature, and the contribution of microRNA to 22q11.2-CNV associated schizophrenia has been previously hypothesized [116,117]. If the microRNAs are important regulatory elements for 22q11.2-associated traits, our approach is insufficient to detect them.

Rather than focusing on any specific tissue(s), we chose to perform a cross-tissue analysis, an approach that improves power to detect gene-trait associations and detected 16p11.2 genes associated with schizophrenia, IQ, and BMI [59]. While we might expect that brain-specific models would be best at detecting relevant genes for brain-related traits, we are limited by the amount of data available – brain tissue transcriptomes are available for fewer than half of the GTEx individuals [52]. An underlying assumption behind the use of all tissues (rather than just brain tissues) for these mental disorders is that eQTLs for our genes of interest are shared across tissues, and that the same eQTLs affect the expression of a gene in the brain as in other tissues.

In general, eQTLs tend to be either highly shared between tissues or highly tissue-specific, largely as a function of the gene being expressed exclusively or nearly exclusively in a single tissue [118]. The GTEx correlation of eQTL effect sizes between brain and non-brain tissues is 0.499 (Spearman) [52]. We may miss genes of interest that have brain-specific expression but not enough power to detect eQTLs. Furthermore, as these eQTLs come from adult tissues, we would miss genes where effects on brain-related traits are specific to early developmental timepoints.

A further limitation is that the variation in expression that can be modeled using eQTLs may be considerably smaller for some genes than the effect of deletions and duplications. For example, there may be a gene at 22q11.2 for which decreases in expression contribute to schizophrenia, but only when expression levels are reduced beyond a threshold, e.g., to nearly 50% of the expression levels of non-carriers. We saw an improvement in the overlap between the gene-by-gene and carrier/non-carrier PheWAS traits when we restricted our analyses to the individuals with the most extreme CNV gene expression across the region, supporting this threshold hypothesis which could be pursued in further study.

Alternatively, the overlap with carrier phenotypes observed when considering predictions across the CNV region could support a multi-gene hypothesis. So far, we have considered the effect of each CNV gene independently, when the genes may not be acting independently. A *Drosophila* model for 16p11.2 genes has shown evidence of epistasis between genes within a CNV as a modifier of phenotype [40]. If there are 16p11.2 traits in humans also driven by epistasis, our single-gene screen would not have detected the appropriate genes for those traits. Similarly, traits

driven by multiple genes would be detectable in our carrier screen but not in our gene-by-gene PheWAS. Given the strong possibility that there are multiple genetic drivers for each trait, efficient ways to consider multiple genes are necessary [119,120].

Because the CNV carrier individuals in our biobank are young (median age < 18), we don't yet know what traits might commonly occur once individuals reach older age. There were traits in our analysis that were over-represented in older CNV carriers, but difficult to interpret as they didn't meet our frequency threshold, including: *dementia with cerebral degenerations* in 22q11.2 deletion carriers, *anterior horn cell disease* in 16p11.2 deletion carriers, and *cerebral degenerations, unspecified* in 16p11.2 duplication carriers. These findings show a need for longitudinal studies of carrier cohorts and studies of carriers in older age. Such additional data may point to additional clinical features of 16p11.2 and 22q11.2 CNV carriers.

CONCLUSION

In developing our approach, we hypothesized that naturally occurring variation in gene expression of CNV genes in non-carriers would convey risk for traits seen in CNV carriers. We found that this was true for at least three 16p11.2 associated traits: BMI, schizophrenia, and IQ. Promisingly, the direction of association was generally consistent with whether the trait was found in duplication or deletion carriers. Our approach is computationally efficient, extendable to other CNV-trait pairs, and overcomes one limitation of animal models by testing the effect of CNV genes specifically in humans.

In this study, we synthesized information from both large GWAS studies and EHR-linked biobanks, benefiting from the strengths of both approaches. Psychiatric brain-related disorders such as autism, schizophrenia, and bipolar disorder have a population frequency below 5%, so large datasets specifically ascertained for brain-related disorders are better at providing sufficient statistical power for association analysis, especially when the effect of each gene is small. On the other hand, the presence of many diagnostic codes in a biobank help identify brain-related traits that may be relevant to CNVs but not the primary reported symptoms, such as speech and language disorder. We were also able to carry out two distinct and complementary analyses using the same dataset. The presence of CNV carrier status in the EHR-linked biobank allowed us to probe the phenotypic consequences of the entire deletion or duplication. Then, we were able to test each CNV gene for association with the same diagnostic descriptions.

Our novel approach provided insights into how individual genes in the 16p11.2 and 22q11.2 CNVs may drive health and behavior in a human population. Expression imputation methods allowed us to study the predicted effects of individual CNV genes in large human populations. The incorporation of medical records into biobanks provided a way to determine clinical symptoms and diagnoses to which expression differences in the genes may contribute. We expect our ability to detect genes with this type of approach to increase in the coming years, as more individuals in biobanks are genotyped, the number of individuals contributing to large cohorts grow, and the methods to more finely and accurately predict gene expression improve. Additional experiments on our newly prioritized genes are necessary to determine their specific functional impact on brain-related disorders and to evaluate their value as putative therapeutic targets.

REFERENCES

1. Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z, et al. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature. 2011;478:97–102.

2. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, et al. Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet. 2009;41:1223–7.

3. Shinawi M, Liu P, Kang S-HL, Shen J, Belmont JW, Scott DA, et al. Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. J Med Genet. 2010;47:332–41.

4. Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA, et al. Recurrent 16p11.2 microdeletions in autism. Hum Mol Genet. 2007;17:628–38.

5. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, et al. Association between Microdeletion and Microduplication at 16p11.2 and Autism. N Engl J Med. 2008;358:667–75.

6. Bassett AS, Chow EWC, Husted J, Weksberg R, Caluseriu O, Webb GD, et al. Clinical features of 78 adults with 22q11 deletion syndrome. Am J Med Genet Part A. 2005;138A:307–13.

7. Campbell IM, Sheppard SE, Crowley TB, McGinn DE, Bailey A, McGinn MJ, et al. What is new with 22q? An update from the 22q and You Center at the Children's Hospital of Philadelphia. Am J Med Genet Part A. 2018;176:2058–69.

8. Wentzel C, Fernström M, Öhrner Y, Annerén G, Thuresson A-C. Clinical variability of the 22q11.2 duplication syndrome. Eur J Med Genet. 2008;51:501–10.

9. Schneider M, Debbané M, Bassett AS, Chow EWC, Fung WLA, van den Bree MBM, et al. Psychiatric Disorders From Childhood to Adulthood in 22q11.2 Deletion Syndrome: Results

From the International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome. Am J Psychiatry. 2014;171:627–39.

10. Voll SL, Boot E, Butcher NJ, Cooper S, Heung T, Chow EWC, et al. Obesity in adults with 22q11.2 deletion syndrome. Genet Med. 2017;19:204–8.

11. Carlson C, Papolos D, Pandita RK, Faedda GL, Veit S, Goldberg R, et al. Molecular analysis of velo-cardio-facial syndrome patients with psychiatric disorders. Am J Hum Genet. 1997;60:851–9.

12. Sahoo T, Theisen A, Rosenfeld JA, Lamb AN, Ravnan JB, Schultz RA, et al. Copy number variants of schizophrenia susceptibility loci are associated with a spectrum of speech and developmental delays and behavior problems. Genet Med. 2011;13:868–80.

13. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet. 2008;84:148–61.

14. Bijlsma EK, Gijsbers ACJ, Schuurs-Hoeijmakers JHM, van Haeringen A, Fransen van de Putte DE, Anderlid B-M, et al. Extending the phenotype of recurrent rearrangements of 16p11.2: Deletions in mentally retarded patients without autism and in normal individuals. Eur J Med Genet. 2009;52:77–87.

15. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, et al. Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet. 2009;41:1223–7.

16. Rees E, Kirov G, Sanders A, Walters JTR, Chambert KD, Shi J, et al. Evidence that duplications of 22q11.2 protect against schizophrenia. Mol Psychiatry. 2014;19:37–40.

17. Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature.

2010;463:671–5.

18. Smith ACM, McGavran L, Robinson J, Waldstein G, Macfarlane J, Zonona J, et al. Interstitial deletion of (17)(p11.2p11.2) in nine patients. Am J Med Genet. 1986;24:393–414.

19. Potocki L, Chen KS, Park SS, Osterholm DE, Withers MA, Kimonis V, et al. Molecular mechanism for duplication 17p11.2 - The homologous recombination reciprocal of the Smith-Magenis microdeletion. Nat Genet. 2000;24:84–7.

20. Slager RE, Newton TL, Vlangos CN, Finucane B, Elsea SH. Mutations in RAI1 associated with Smith–Magenis syndrome. Nat Genet. 2003;33:466–8.

21. Walz K, Paylor R, Yan J, Bi W, Lupski JR. Rai1 duplication causes physical and behavioral phenotypes in a mouse model of dup(17)(p11.2p11.2). J Clin Invest. 2006;116:3035–41.

22. Williams JCP, Barratt-Boyes BG, Lowe JB. Supravalvular Aortic Stenosis. Circulation. 1961;24:1311–8.

23. Beuren AJ, Apitz J, Harmjanz D. Supravalvular Aortic Stenosis in Association with Mental Retardation and a Certain Facial Appearance. Circulation. 1962;26:1235–40.

24. Curran ME, Atkinson DL, Ewart AK, Morris CA, Leppert MF, Keating MT. The elastin gene is disrupted by a translocation associated with supravalvular aortic stenosis. Cell. 1993;73:159–68.

25. Ewart AK, Morris CA, Atkinson D, Jin W, Sternes K, Spallone P, et al. Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. Nat Genet. 1993;5:11–6.

26. Koolen DA, Vissers LELM, Pfundt R, De Leeuw N, Knight SJL, Regan R, et al. A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. Nat Genet. 2006;38:999–1001.

27. Koolen DA, Sharp AJ, Hurst JA, Firth H V, Knight SJL, Goldenberg A, et al. Clinical and

molecular delineation of the 17q21.31 microdeletion syndrome. J Med Genet. 2008;45:710–20.

28. Koolen DA, Kramer JM, Neveling K, Nillesen WM, Moore-Barton HL, Elmslie F V, et al. Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. Nat Genet. 2012;44:639–41.

29. Jerome LA, Papaioannou VE. DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1. Nat Genet. 2001;27:286–91.

30. Lindsay EA, Vitelli F, Su H, Morishima M, Huynh T, Pramparo T, et al. Tbx1 haploinsufficiency in the DiGeorge syndrome region causes aortic arch defects in mice. Nature. 2001;410:97–101.

31. Merscher S, Funke B, Epstein JA, Heyer J, Puech A, Lu MM, et al. TBX1 Is Responsible for Cardiovascular Defects in Velo-Cardio-Facial/DiGeorge Syndrome. Cell. 2001;104:619–29.

32. Paylor R, Glaser B, Mupo A, Ataliotis P, Spencer C, Sobotka A, et al. Tbx1 haploinsufficiency is linked to behavioral disorders in mice and humans: implications for 22q11 deletion syndrome. Proc Natl Acad Sci U S A. 2006;103:7729–34.

33. Ma G, Shi Y, Tang W, He Z, Huang K, Li Z, et al. An association study between the genetic polymorphisms within TBX1 and schizophrenia in the Chinese population. Neurosci. Lett. 2007.

34. Consortium SWG of the PG, Ripke S, Walters JT, O'Donovan MC. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. medRxiv. 2020;2020.09.12.20192922.

35. Clements CC, Wenger TL, Zoltowski AR, Bertollo JR, Miller JS, de Marchena AB, et al. Critical region within 22q11.2 linked to higher rate of autism spectrum disorder. Mol Autism. 2017;8:58.

36. Crepel A, Steyaert J, De la Marche W, De Wolf V, Fryns J-P, Noens I, et al. Narrowing the

critical deletion region for autism spectrum disorders on 16p11.2. Am J Med Genet Part B Neuropsychiatr Genet. 2011;156:243–5.

37. Pucilowska J, Vithayathil J, Tavares EJ, Kelly C, Colleen Karlo J, Landreth GE. The 16p11.2 deletion mouse model of autism exhibits altered cortical progenitor proliferation and brain cytoarchitecture linked to the ERK MAPK pathway. J Neurosci. 2015;35:3190–200.

38. Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S, et al. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature. 2012;485:363–7.

39. Blaker-Lee A, Gupta S, McCammon JM, De Rienzo G, Sive H. Zebrafish homologs of genes within 16p11.2, a genomic region associated with brain disorders, are active during brain development, and include two deletion dosage sensor genes. Dis Model Mech. 2012;5.

40. Iyer J, Singh MD, Jensen M, Patel P, Pizzo L, Huber E, et al. Pervasive genetic interactions modulate neurodevelopmental defects of the autism-Associated 16p11.2 deletion in Drosophila melanogaster. Nat Commun. 2018;9:1–19.

41. Paylor R, McIlwain KL, McAninch R, Nellis A, Yuva-Paylor LA, Baldini A, et al. Mice deleted for the DiGeorge/velocardiofacial syndrome region show abnormal sensorimotor gating and learning and memory impairments. Hum Mol Genet. 2001;10:2645–50.

42. Guna A, Butcher NJ, Bassett AS. Comparative mapping of the 22q11.2 deletion region and the potential of simple model organisms. J Neurodev Disord. 2015;7:18.

43. McCammon JM, Blaker-Lee A, Chen X, Sive H. The 16p11.2 homologs fam57ba and doc2a generate certain brain and body phenotypes. Hum Mol Genet. 2017;26:3699–712.

44. Arbogast T, Ouagazzal A-M, Chevalier C, Kopanitsa M, Afinowi N, Migliavacca E, et al. Reciprocal Effects on Neurocognitive and Metabolic Phenotypes in Mouse Models of 16p11.2

Deletion and Duplication Syndromes. Barsh GS, editor. PLOS Genet. 2016;12:e1005709.

45. Ward TR, Zhang X, Leung LC, Zhou B, Muench K, Roth JG, et al. Genome-wide molecular effects of the neuropsychiatric 16p11 CNVs in an iPSC-to-iN neuronal model. bioRxiv. 2020;2020.02.09.940965.

46. Blumenthal I, Ragavendran A, Erdin S, Klei L, Sugathan A, Guide JR, et al. Transcriptional Consequences of 16p11.2 Deletion and Duplication in Mouse Cortex and Multiplex Autism Families. Am J Hum Genet. 2014;94:870–83.

47. Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, et al. Genome-wide Transcriptome Profiling Reveals the Functional Impact of Rare De Novo and Recurrent CNVs in Autism Spectrum Disorders. Am J Hum Genet. 2012;91:38–55.

48. Zhang X, Zhang Y, Zhu X, Purmann C, Haney MS, Ward T, et al. Local and global chromatin interactions are altered by large genomic deletions associated with human brain development. Nat Commun. 2018;9:5356.

49. Jalbrzikowski M, Lazaro MT, Gao F, Huang A, Chow C, Geschwind DH, et al. Transcriptome Profiling of Peripheral Blood in 22q11.2 Deletion Syndrome Reveals Functional Pathways Related to Psychosis and Autism Spectrum Disorder. van Amelsvoort T, editor. PLoS One. 2015;10:e0132542.

50. Migliavacca E, Golzio C, Männik K, Blumenthal I, Oh EC, Harewood L, et al. A Potential Contributory Role for Ciliary Dysfunction in the 16p11.2 600 kb BP4-BP5 Pathology. Am J Hum Genet. 2015;96:784–96.

51. Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv. 2019;787903.

52. Aguet F, Ardlie KG, Cummings BB, Gelfand ET, Getz G, Hadley K, et al. Genetic effects on

gene expression across human tissues. Nature. 2017;550:204–13.

53. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. 2013. p. 580–5.

54. Ardlie KG, DeLuca DS, Segrè A V., Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-). 2015;348:648–60.

55. Freund MK, Burch K, Shi H, Mancuso N, Kichaev G, Garske KM, et al. Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits.

56. Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA. Clan genomics and the complex architecture of human disease. Cell. 2011. p. 32–43.

57. Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, et al. A nondegenerate code of deleterious variants in mendelian loci contributes to complex disease risk. Cell. 2013;155:70–80.

58. Gamazon ER, Wheeler HE, Shah KP, Mozaffari S V., Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47:1091–8.

59. Barbeira AN, Pividori MD, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. Plagnol V, editor. PLOS Genet. 2019;15:e1007889.

60. Schizophrenia Working Group of the Psychiatric Genomics Consortium S. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421–7.

61. Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. Nat Genet. 2019;51:793–

803.

62. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51:431–44.

63. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518:197–206.

64. Savage JE, Jansen PR, Stringer S, Watanabe K, Bryois J, De Leeuw CA, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat Genet. 2018;50:912–9.

65. Roden D, Pulley J, Basford M, Bernard G, Clayton E, Balser J, et al. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. Clin Pharmacol Ther. 2008;84:362–9.

66. Schizophrenia Working Group of the Psychiatric Genomics Consortium {fname}. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421–7.

67. UK Biobank — Neale lab [Internet]. [cited 2020 Mar 28]. Available from: http://www.nealelab.is/uk-biobank

68. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. Hum Mol Genet. 2018;27:3641–9.

69. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nat 2018 5627726. 2018;562:203–9.

70. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9:1825.

71. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.

72. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26:2190–1.

73. Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, et al. Probabilistic fine-mapping of transcriptome-wide association studies. Nat Genet. 2019;51:675–82.

74. Zhou D, Jiang Y, Zhong X, Cox NJ, Liu C, Gamazon ER. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. Nat Genet. 2020;52:1239–46.

75. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008;84:362–9.

76. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48:1279–83.

77. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–7.

78. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010/03/24. 2010;26:1205–10.

79. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics. 2014/04/14. 2014;30:2375–6.

80. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore M.

Online Mendelian Inheritance in Man, OMIM®.

81. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013;31:1102–11.

82. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R . Bioinformatics. 2014;30:2811–2.

83. Dantas AG, Santoro ML, Nunes N, de Mello CB, Pimenta LSE, Meloni VA, et al. Downregulation of genes outside the deleted region in individuals with 22q11.2 deletion syndrome. Hum Genet. 2019;138:93–103.

84. Merla G, Howald C, Henrichsen CN, Lyle R, Wyss C, Zabot MT, et al. Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. Am J Hum Genet. 2006;79:332–41.

85. Roth JG, Muench KL, Asokan A, Mallett VM, Gai H, Verma Y, et al. Copy Number Variation at 16p11.2 Imparts Transcriptional Alterations in Neural Development in an hiPSC-derived Model of Corticogenesis. bioRxiv. 2020;2020.04.22.055731.

86. Steinman KJ, Spence SJ, Ramocki MB, Proud MB, Kessler SK, Marco EJ, et al. 16p11.2 deletion and duplication: Characterizing neurologic phenotypes in a large clinically ascertained cohort. Am J Med Genet Part A. 2016;170:2943–55.

87. Karunanithi Z, Vestergaard EM, Lauridsen MH. Transposition of the great arteries - a phenotype associated with 16p11.2 duplications? World J Cardiol. 2017;9:848–52.

88. Fernandez BA, Roberts W, Chung B, Weksberg R, Meyn S, Szatmari P, et al. Phenotypic spectrum associated with de novo and inherited deletions and duplications at 16p11.2 in individuals ascertained for diagnosis of autism spectrum disorder. J Med Genet. 2010;47:195–

203.

89. Wenger TL, Miller JS, DePolo LM, de Marchena AB, Clements CC, Emanuel BS, et al. 22q11.2 duplication syndrome: elevated rate of autism spectrum disorder and need for medical screening. Mol Autism. 2016;7:27.

90. Piotrowski A, Xie J, Liu YF, Poplawski AB, Gomes AR, Madanecki P, et al. Germline loss-of-function mutations in LZTR1 predispose to an inherited disorder of multiple schwannomas. Nat Genet. 2014;46:182–7.

91. Ma H, Blake T, Chitnis A, Liu P, Balla T. Crucial role of phosphatidylinositol 4-kinase IIIα in development of zebrafish pectoral fin is linked to phosphoinositide 3-kinase and FGF signaling. J Cell Sci. 2009;122:4303–10.

92. Bojjireddy N, Botyanszki J, Hammond G, Creech D, Peterson R, Kemp DC, et al. Pharmacological and genetic targeting of the PI4KA enzyme reveals its important role in maintaining plasma membrane phosphatidylinositol 4-phosphate and phosphatidylinositol 4,5-bisphosphate levels. J Biol Chem. 2014;289:6120–32.

93. Chen W, Liu J, Yuan D, Zuo Y, Liu Z, Liu S, et al. Progress and perspective of TBX6 gene in congenital vertebral malformations. Oncotarget. 2016;7:57430–41.

94. Liu J, Wu N, Yang N, Takeda K, Chen W, Li W, et al. TBX6-associated congenital scoliosis (TACS) as a clinically distinguishable subtype of congenital scoliosis: further evidence supporting the compound inheritance and TBX6 gene dosage model. Genet Med. 2019;21:1548–58.

95. Watabe-Rudolph M, Schlautmann N, Papaioannou VE, Gossler A. The mouse rib-vertebrae mutation is a hypomorphic Tbx6 allele. Mech Dev. 2002;119:251–6.

96. Dines JN, Golden-Grant K, LaCroix A, Muir AM, Cintrón DL, McWalter K, et al.

TANGO2: expanding the clinical phenotype and spectrum of pathogenic variants. Genet Med. 2019;21:601–7.

97. Lalani SR, Liu P, Rosenfeld JA, Watkin LB, Chiang T, Leduc MS, et al. Recurrent Muscle Weakness with Rhabdomyolysis, Metabolic Crises, and Cardiac Arrhythmia Due to Bi-allelic TANGO2 Mutations. Am J Hum Genet. 2016;98:347–57.

98. Ayala R, Willhoft O, Aramayo RJ, Wilkinson M, McCormack EA, Ocloo L, et al. Structure and regulation of the human INO80-nucleosome complex. Nature. 2018;556:391–5.

99. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE. Mouse Genome Database (MGD) 2019. Nucleic Acids Res. 2018;47:D801–6.

100. Powell CM, Miyakawa T. Schizophrenia-relevant behavioral testing in rodent models: a uniquely human disorder? Biol Psychiatry. 2006;59:1198–207.

101. Dobbyn A, Huckins LM, Boocock J, Sloofman LG, Glicksberg BS, Giambartolomei C, et al. Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS. Am J Hum Genet. 2018;102:1169–84.

102. Gusev A, Mancuso N, Won H, Kousi M, Finucane HK, Reshef Y, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat Genet. 2018;50:538–48.

103. Walker RL, Ramaswami G, Hartl C, Mancuso N, Gandal MJ, de la Torre-Ubieta L, et al. Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. Cell. 2019;179:750-771.e22.

104. Pallant A, Eskenazi A, Mattei MG, Fournier REK, Carlsson SR, Fukuda M, et al. Characterization of cDNAs encoding human leukosialin and localization of the leukosialin gene to chromosome 16. Proc Natl Acad Sci U S A. 1989;86:1328–32.

105. Park JK, Rosenstein YJ, Remold-O'Donnell E, Bierer BE, Rosen FS, Burakoff SJ. Enhancement of T-cell activation by the CD43 molecule whose expression is defective in Wiskott-Aldrich syndrome. Nature. 1991;350:706–9.

106. Collins RL, Glessner JT, Porcu E, Niestroj L-M, Ulirsch J, Kellaris G, et al. A cross-disorder dosage sensitivity map of the human genome. medRxiv. 2021;2021.01.26.21250098.

107. Marioni RE, Yang J, Dykiert D, Mõttus R, Campbell A, Davies G, et al. Assessing the genetic overlap between BMI and cognitive function. Mol Psychiatry. 2016;21:1477–82.

108. Sabia S, Kivimaki M, Shipley MJ, Marmot MG, Singh-Manoux A. Body mass index over the adult life course and cognition in late midlife: the Whitehall II Cohort Study. Am J Clin Nutr. 2009;89:601–7.

109. Ikeda M, Tanaka S, Saito T, Ozaki N, Kamatani Y, Iwata N. Re-evaluating classical body type theories: Genetic correlation between psychiatric disorders and body mass index. Psychol Med. 2018;48:1745–8.

110. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47:1236–41.

111. Mizuno A, Okada Y. Biological characterization of expression quantitative trait loci (eQTLs) showing tissue-specific opposite directional effects. Eur J Hum Genet. 2019;

112. Lopez-Rivera E, Liu YP, Verbitsky M, Anderson BR, Capone VP, Otto EA, et al. Genetic Drivers of Kidney Defects in the DiGeorge Syndrome. N Engl J Med. 2017;NEJMoa1609009.

113. Wu N, Ming X, Xiao J, Wu Z, Chen X, Shinawi M, et al. TBX6 Null Variants and a Common Hypomorphic Allele in Congenital Scoliosis . N Engl J Med. 2015;372:341–50.

114. Loviglio MN, Leleu M, Männik K, Passeggeri M, Giannuzzi G, van der Werf I, et al. Chromosomal contacts connect loci associated with autism, BMI and head circumference

phenotypes. Mol Psychiatry. 2016;

115. Bachmann-Gagescu R, Mefford HC, Cowan C, Glew GM, Hing A V, Wallace S, et al. Recurrent 200-kb deletions of 16p11.2 that include the SH2B1 gene are associated with developmental delay and obesity. Genet Med. 2010;12:641–7.

116. Forstner AJ, Degenhardt F, Schratt G, Nöthen MM. MicroRNAs as the cause of schizophrenia in 22q11.2 deletion carriers, and possible implications for idiopathic disease: a mini-review. Front Mol Neurosci. 2013;6:47.

117. De la Morena MT, Eitson JL, Dozmorov IM, Belkaya S, Hoover AR, Anguiano E, et al. Signature MicroRNA expression patterns identified in humans with 22q11.2 deletion/DiGeorge syndrome. Clin Immunol. 2013;147:11–22.

118. Consortium TGte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.

119. Gokhale A, Hartwig C, Freeman AAH, Bassell JL, Zlatic SA, Savas CS, et al. Systems analysis of the 22q11.2 microdeletion syndrome converges on a mitochondrial interactome necessary for synapse function and behavior. J Neurosci. 2019;39:3561–81.

120. Jensen M, Girirajan S. An interaction-based model for neuropsychiatric features of copy-number variants. bioRxiv. 2018;459958.

121. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26:2336–7.

FIGURES

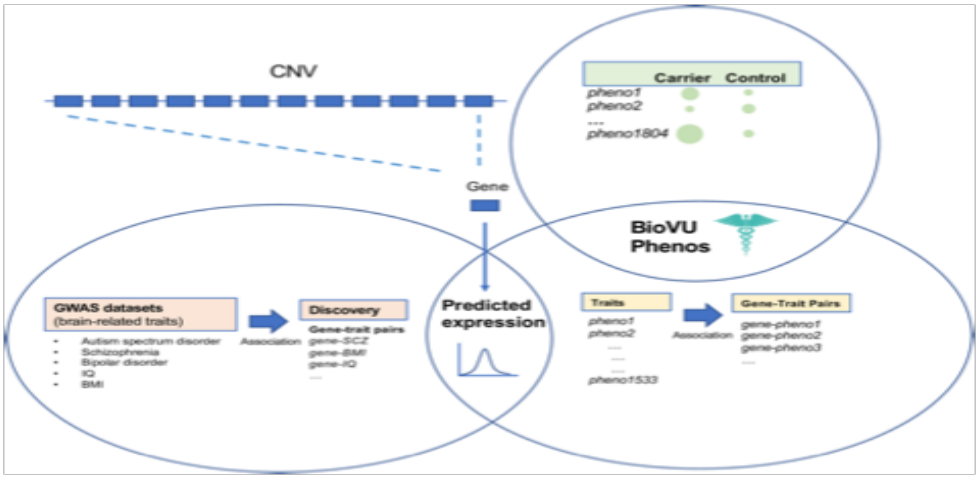
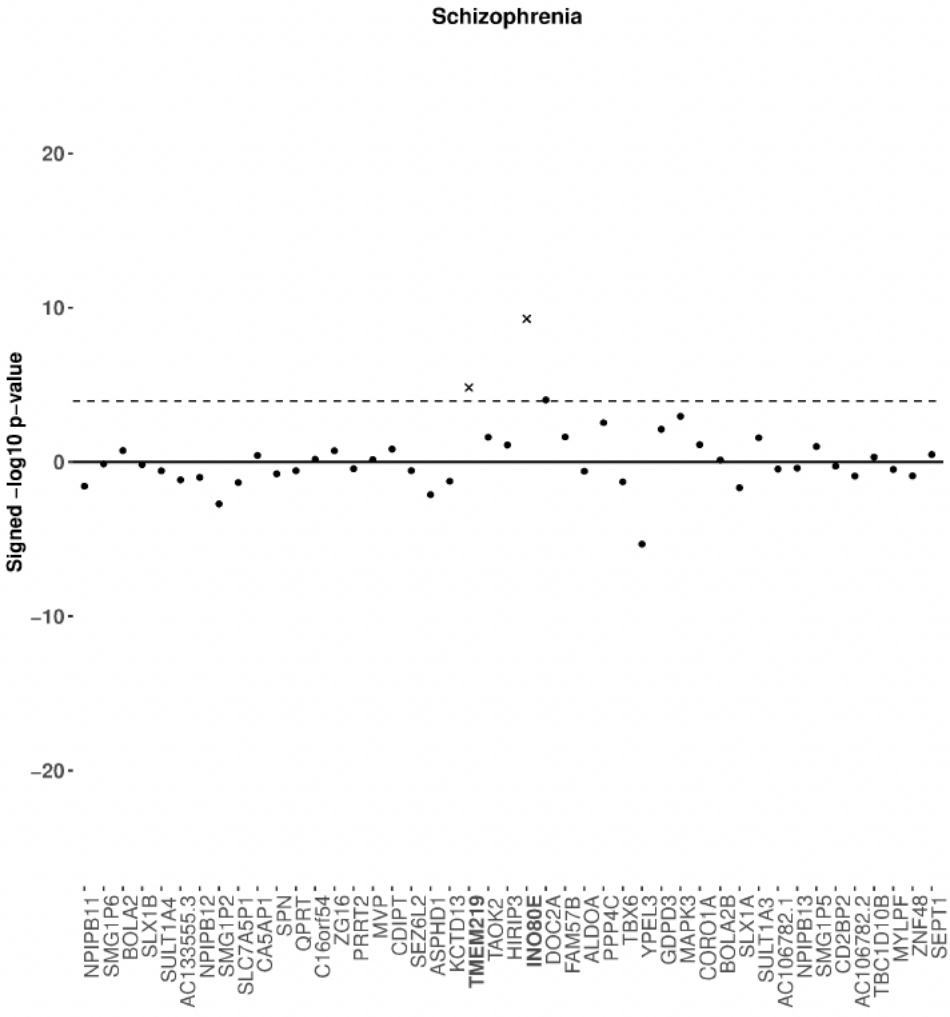
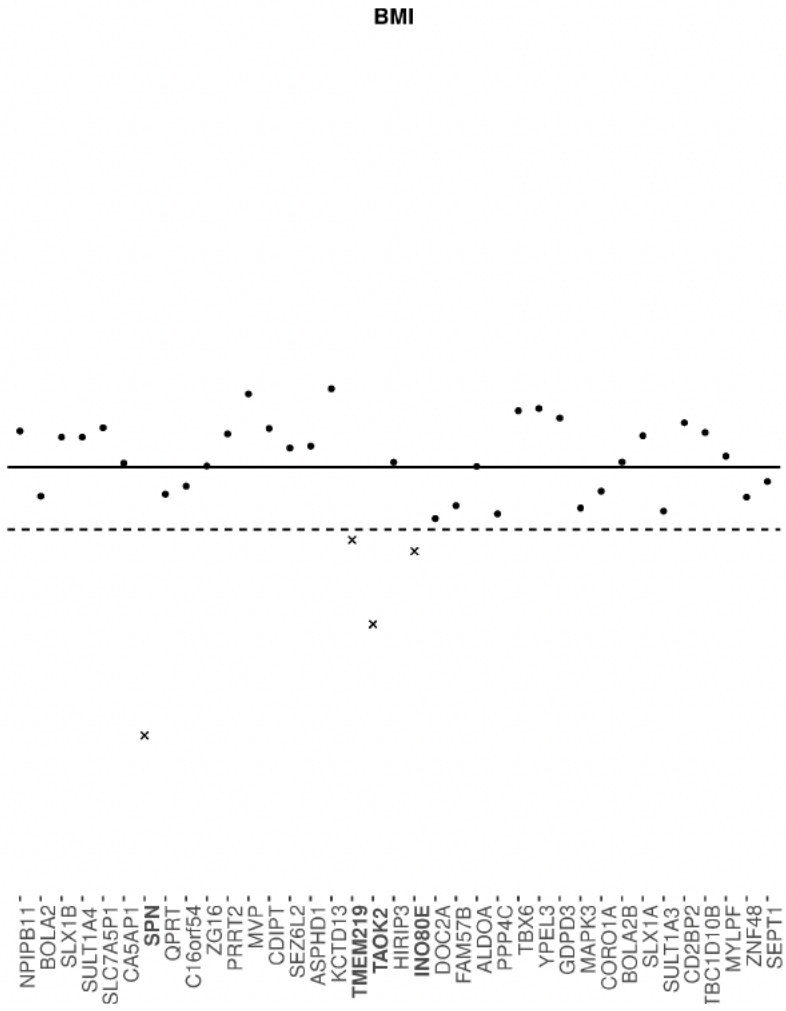


Figure 2.1: An overview of the three components of this study.
We probed the effects of individual genes in the 16p11.2 and 22q11.2 CNVs on phenotype in two ways. First (bottom left), we used large GWAS datasets for brain-related traits associated with both CNVs to determine whether variation in predicted expression in any of the individual genes in each CNV was associated with case-control status for each trait. In the second component of this study (top right), we used a biobank containing clinical and genotypic data to identify the individuals with 16p11.2 and 22q11.2 duplications or deletions and determined the clinical traits that were over-represented in CNV-carriers. Third (bottom right), we used the biobank to perform a phenome-wide association study to determine clinical traits that are driven by the predicted expression of individual CNV genes, as well as whether these traits overlapped with traits over-represented in CNV carriers. Analyses one and three are integrated in their use of imputed expression; analyses two and three are integrated in their use of electronic health data.


Schizophrenia
Signed −log10 p−value
20
10
0
−10
−20
NPIPB11
SMG1P6
BOLA2
SLX1B
SULT1A4
AC133555.3
NPIPB12
SMG1P2
SLC7A5P1
CA5AP1
SPN
QPRT
C16orf54
ZG16
PRRT2
MVP
CDIPT
SEZ6L2
ASPHD1
KCTD13
TMEM219
TAOK2
HIRIP3
INO80E
DOC2A
FAM57B
ALDOA
PPP4C
TBX6
YPEL3
GDPD3
MAPK3
CORO1A
BOLA2B
SLX1A
SULT1A3
AC106782.1
NPIPB13
SMG1P5
CD2BP2
AC106782.2
TBC1D10B
MYLPF
ZNF48
SEPT1


BMI
NPIPB11
BOLA2
SLX1B
SULT1A4
SLC7A5P1
CA5AP1
SPN
QPRT
C16orf54
ZG16
PRRT2
MVP
CDIPT
SEZ6L2
ASPHD1
KCTD13
TMEM219
TAOK2
HIRIP3
INO80E
DOC2A
FAM57B
ALDOA
PPP4C
TBX6
YPEL3
GDPD3
MAPK3
CORO1A
BOLA2B
SLX1A
SULT1A3
CD2BP2
TBC1D10B
MYLPF
ZNF48
SEPT1

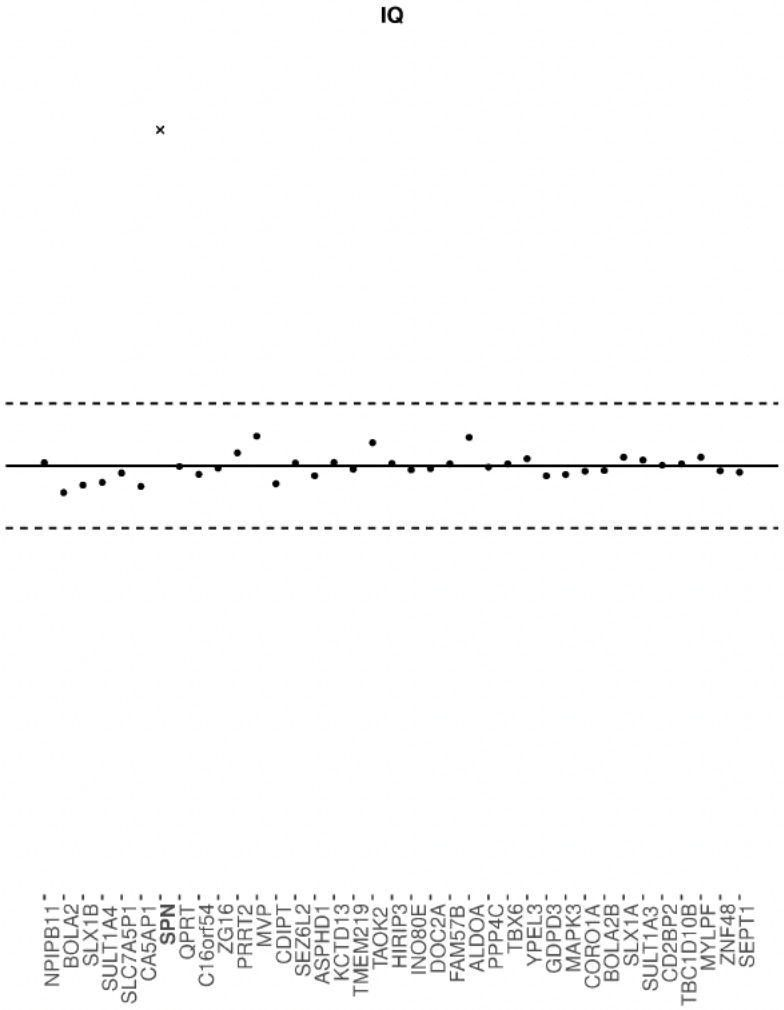
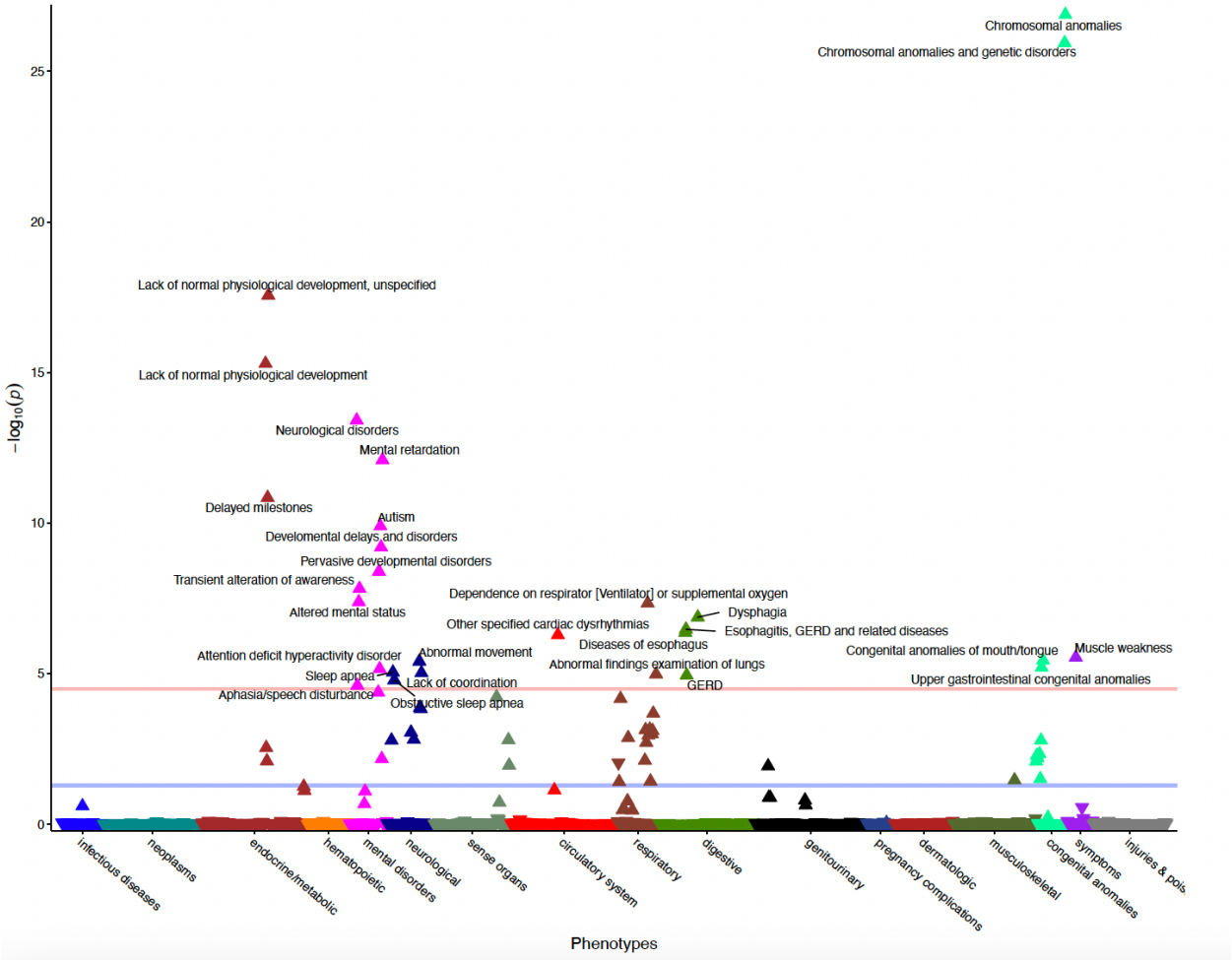


Figure 2.2. Association between 16p11.2 genes and three brain-related traits.

Association between predicted expression of 16p11.2 genes and schizophrenia (top), BMI (middle), IQ (bottom) using MultiXcan (schizophrenia) and S-MultiXcan (BMI, IQ). Genes are listed on the horizontal access in order of chromosomal position. The -log10 p-values on the vertical axis are given a positive or negative direction based on the average direction of the single-tissue results. The significance threshold, $P < 7.9 \times 10^{-5}$, is a Bonferroni correction on the total number of 16p11.2 and 22q11.2 genes (127) tested across 5 traits (0.05/(5*127)). Genes exceeding the significance threshold in the expected direction (positive for schizophrenia, negative for BMI, either for IQ) are denoted as x's.


Chromosomal anomalies
Chromosomal anomalies and genetic disorders
Lack of normal physiological development, unspecified
Lack of normal physiological development
Neurological disorders
Mental retardation
Delayed milestones
Autism
Developmental delays and disorders
Pervasive developmental disorders
Transient alteration of awareness
Altered mental status
Dependence on respirator [Ventilator] or supplemental oxygen
Dysphagia
Other specified cardiac dysrhythmias
Esophagitis, GERD and related diseases
Diseases of esophagus
Congenital anomalies of mouth/tongue
Muscle weakness
Attention deficit hyperactivity disorder
Abnormal movement
Abnormal findings examination of lungs
Sleep apnea
Lack of coordination
GERD
Upper gastrointestinal congenital anomalies
Aphasia/speech disturbance
Obstructive sleep apnea
25
20
15
10
5
0
$-\log_{10}(p)$
infectious diseases
neoplasms
endocrine/metabolic
hematopoietic
mental disorders
neurological
sense organs
circulatory system
respiratory
digestive
genitourinary
pregnancy complications
dermatologic
musculoskeletal
congenital anomalies
symptoms
injuries & pois
Phenotypes

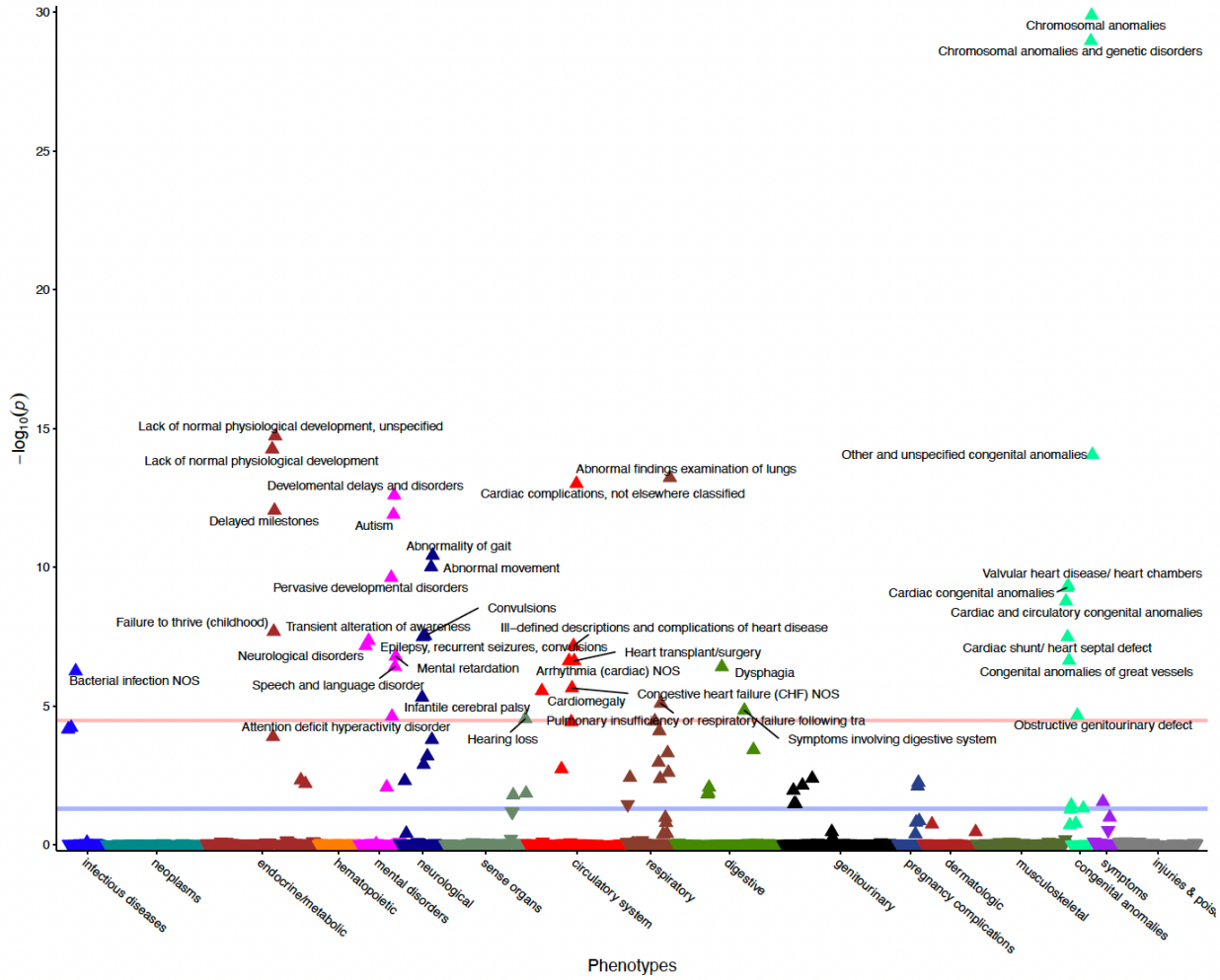
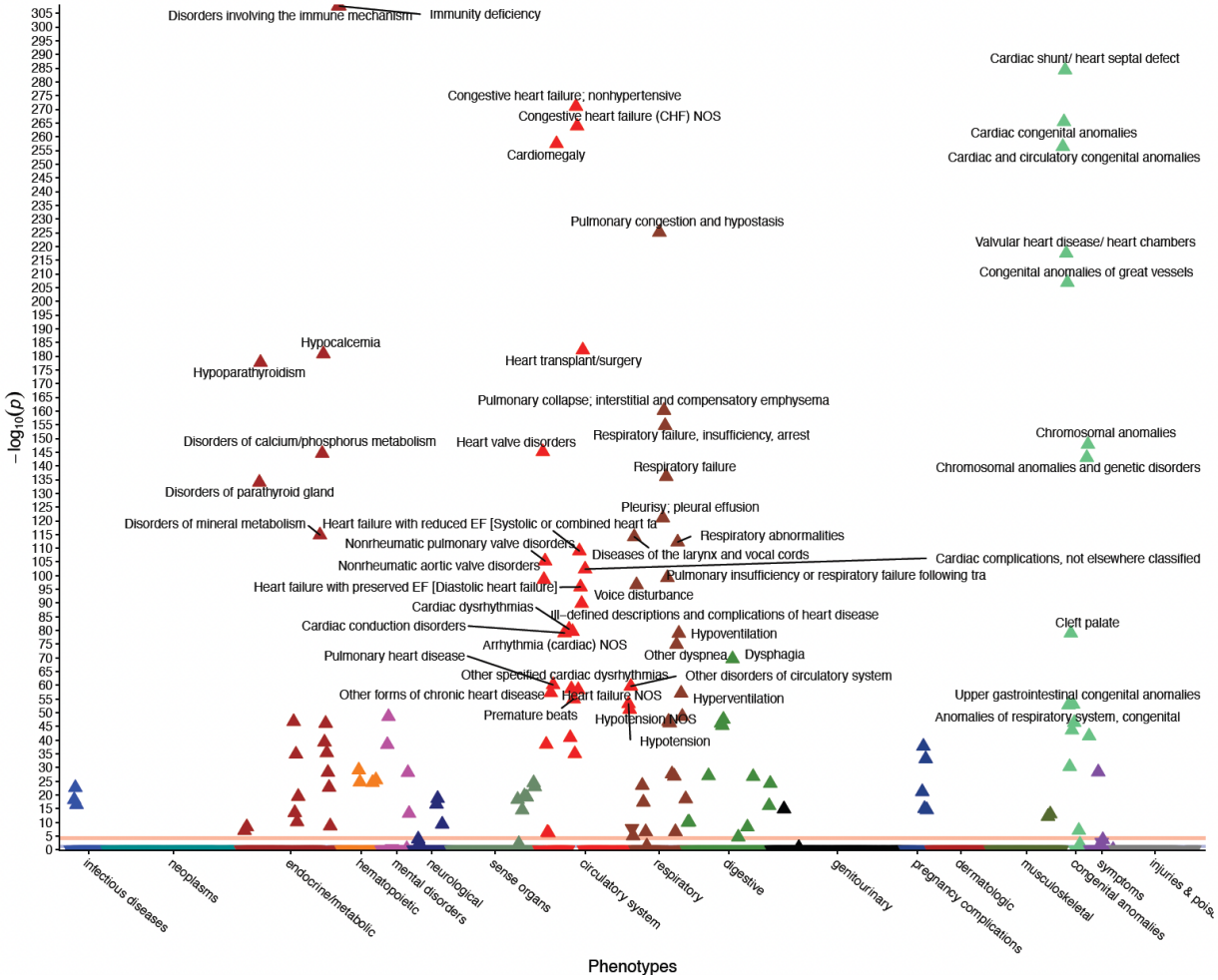


Figure 2.3. Clinical traits over-represented in 16p11.2 deletion and duplication carriers.

CNV carriers were identified in the EHR by keyword search and chart review (top, 16p11.2 deletions [n=48], bottom, 16p11.2 duplications [n=48], see Methods). Controls included all individuals without the CNV within the medical home population at Vanderbilt (n~707,000). The x-axis represents the PheWAS codes that are mapped from ICD-9/ICD-10 codes, grouped and color-coded by organ system. The y-axis represents the level of significance ($-\log_{10}p$). The horizontal red line indicates a Bonferroni correction for the number of phenotypes tested in this PheWAS ($p = 0.05/1{,}795 = 2.8\text{x}10^{-5}$); the horizontal blue line indicates $p = 0.05$. Each triangle represents a phenotype. Triangles represent direction of effect; upward pointing arrows indicate phenotypes more common in cases. Covariates included age, sex, and self-reported race extracted from the EHR. Phenotypes reaching Bonferroni-corrected significance level are labeled in plot.


Disorders involving the immune mechanism
Immunity deficiency
Cardiac shunt/ heart septal defect
Congestive heart failure; nonhypertensive
Congestive heart failure (CHF) NOS
Cardiomegaly
Cardiac congenital anomalies
Cardiac and circulatory congenital anomalies
Pulmonary congestion and hypostasis
Valvular heart disease/ heart chambers
Congenital anomalies of great vessels
Hypocalcemia
Hypoparathyroidism
Heart transplant/surgery
Pulmonary collapse; interstitial and compensatory emphysema
Disorders of calcium/phosphorus metabolism
Heart valve disorders
Respiratory failure, insufficiency, arrest
Chromosomal anomalies
Respiratory failure
Chromosomal anomalies and genetic disorders
Disorders of parathyroid gland
Pleurisy; pleural effusion
Disorders of mineral metabolism
Heart failure with reduced EF [Systolic or combined heart fa
Respiratory abnormalities
Nonrheumatic pulmonary valve disorders
Diseases of the larynx and vocal cords
Cardiac complications, not elsewhere classified
Nonrheumatic aortic valve disorders
Pulmonary insufficiency or respiratory failure following tra
Heart failure with preserved EF [Diastolic heart failure]
Voice disturbance
Cardiac dysrhythmias
Ill-defined descriptions and complications of heart disease
Cleft palate
Cardiac conduction disorders
Hypoventilation
Arrhythmia (cardiac) NOS
Pulmonary heart disease
Other dyspnea
Dysphagia
Other specified cardiac dysrhythmias
Other disorders of circulatory system
Other forms of chronic heart disease
Heart failure NOS
Hyperventilation
Upper gastrointestinal congenital anomalies
Premature beats
Hypotension NOS
Anomalies of respiratory system, congenital
Hypotension
$-\log_{10}(p)$
infectious diseases
neoplasms
endocrine/metabolic
hematopoietic
mental disorders
neurological
sense organs
circulatory system
respiratory
digestive
genitourinary
pregnancy complications
dermatologic
musculoskeletal
congenital anomalies
symptoms
injuries & pois.
Phenotypes

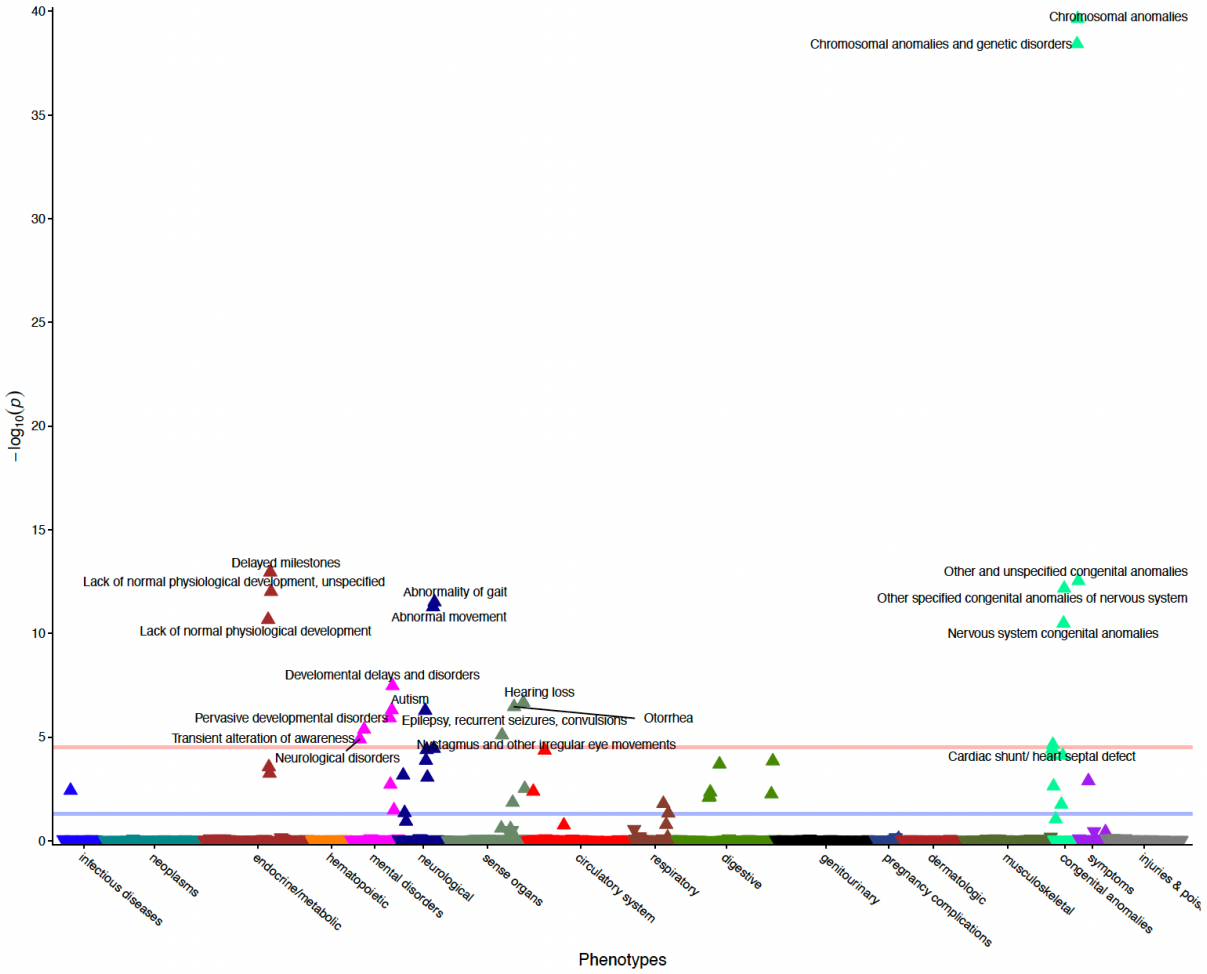
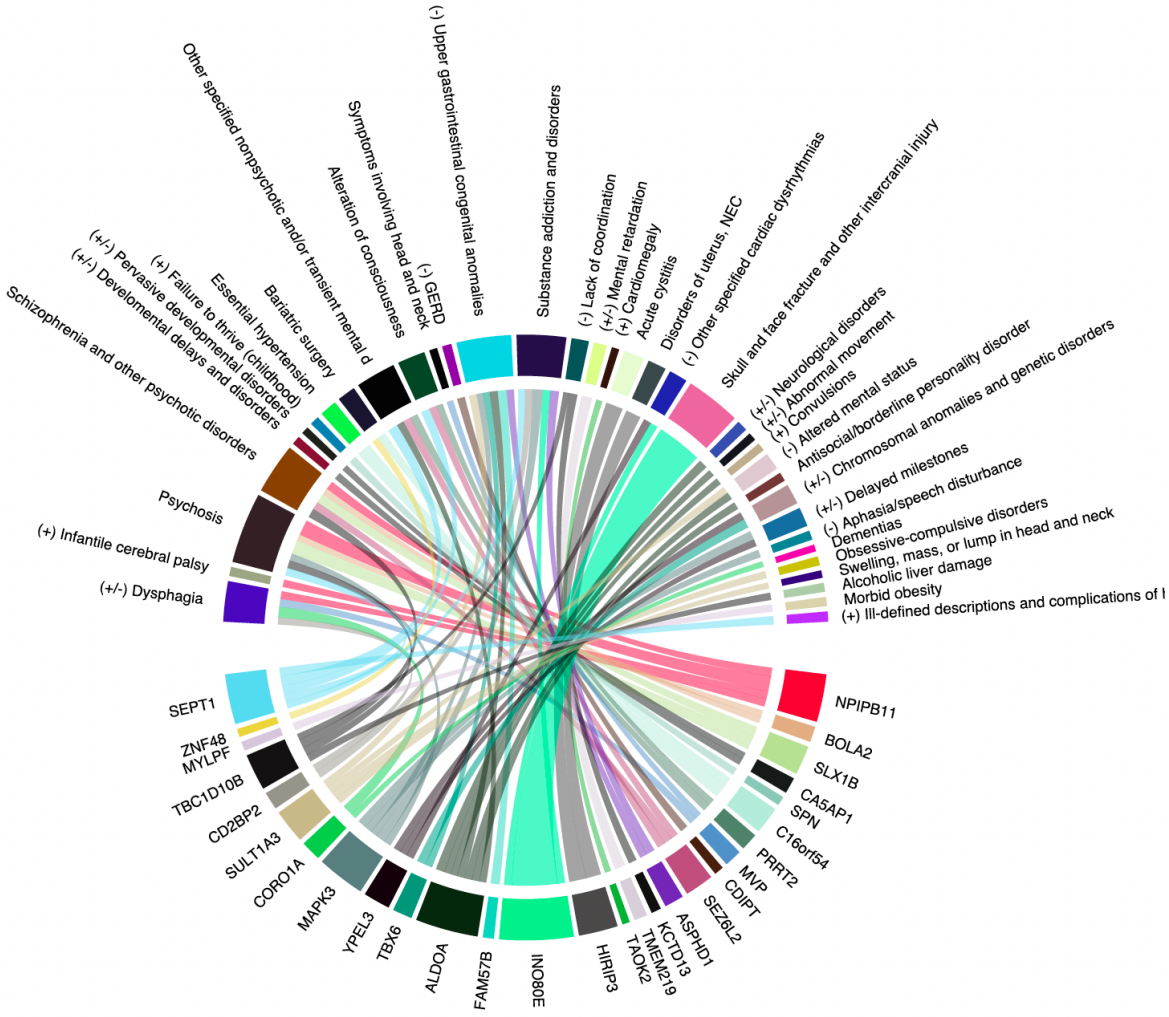


Figure 2.4. Clinical traits over-represented in 22q11.2 deletion and duplication carriers.

CNV carriers were identified in the EHR by keyword search and chart review (left, 22q11.2 deletions [n=388], right, 22q11.2 duplications [n=43], see Methods). Controls included all individuals without the CNV within the medical home population at Vanderbilt (n~707,000). The x-axis represents the PheWAS codes that are mapped from ICD-9/ICD-10 codes, grouped and color-coded by organ system. The y-axis represents the level of significance ($-\log_{10}p$). The horizontal red line indicates a Bonferroni correction for the number of phenotypes tested in this PheWAS ($p = 0.05/1{,}795 = 2.8\text{x}10^{-5}$); the horizontal blue line indicates $p = 0.05$. Each triangle represents a phenotype. Triangles represent direction of effect; upward pointing arrows indicate phenotypes more common in cases. Covariates included age, sex, and self-reported race extracted from the EHR. Top phenotypes ($P < 1.0\text{x}10^{-50}$) are labeled in the 22q11.2 deletion plot (left). Phenotypes reaching Bonferroni-corrected significance level are labeled in the 22q11.2 duplication plot (right).


(-) Upper gastrointestinal congenital anomalies
Substance addiction and disorders
(-) Lack of coordination
(+/-) Mental retardation
(+) Cardiomegaly
Acute cystitis
Disorders of uterus, NEC
(-) Other specified cardiac dysrhythmias
Skull and face fracture and other intercranial injury
(+/-) Neurological disorders
(+/-) Abnormal movement
(+) Convulsions
(-) Altered mental status
Antisocial/borderline personality disorder
(+/-) Chromosomal anomalies and genetic disorders
(+/-) Delayed milestones
(-) Aphasia/speech disturbance
Dementias
Obsessive-compulsive disorders
Swelling, mass, or lump in head and neck
Alcoholic liver damage
Morbid obesity
(+) Ill-defined descriptions and complications of
NPIPB11
BOLA2
SLX1B
CA5AP1
SPN
C16orf54
PRRT2
MVP
CDIPT
SEZ6L2
ASPHD1
KCTD13
TMEM219
TAOK2
HIRIP3
INO80E
FAM57B
ALDOA
TBX6
YPEL3
MAPK3
CORO1A
SULT1A3
CD2BP2
TBC1D10B
MYLPF
ZNF48
SEPT1
(+/-) Dysphagia
(+) Infantile cerebral palsy
Psychosis
Schizophrenia and other psychotic disorders
(+/-) Developmental delays and disorders
(+/-) Pervasive developmental disorders
(+) Failure to thrive (childhood)
Essential hypertension
Bariatric surgery
Alteration of consciousness
Symptoms involving head and neck
Other specified nonpsychotic and/or transient mental d
(-) GERD

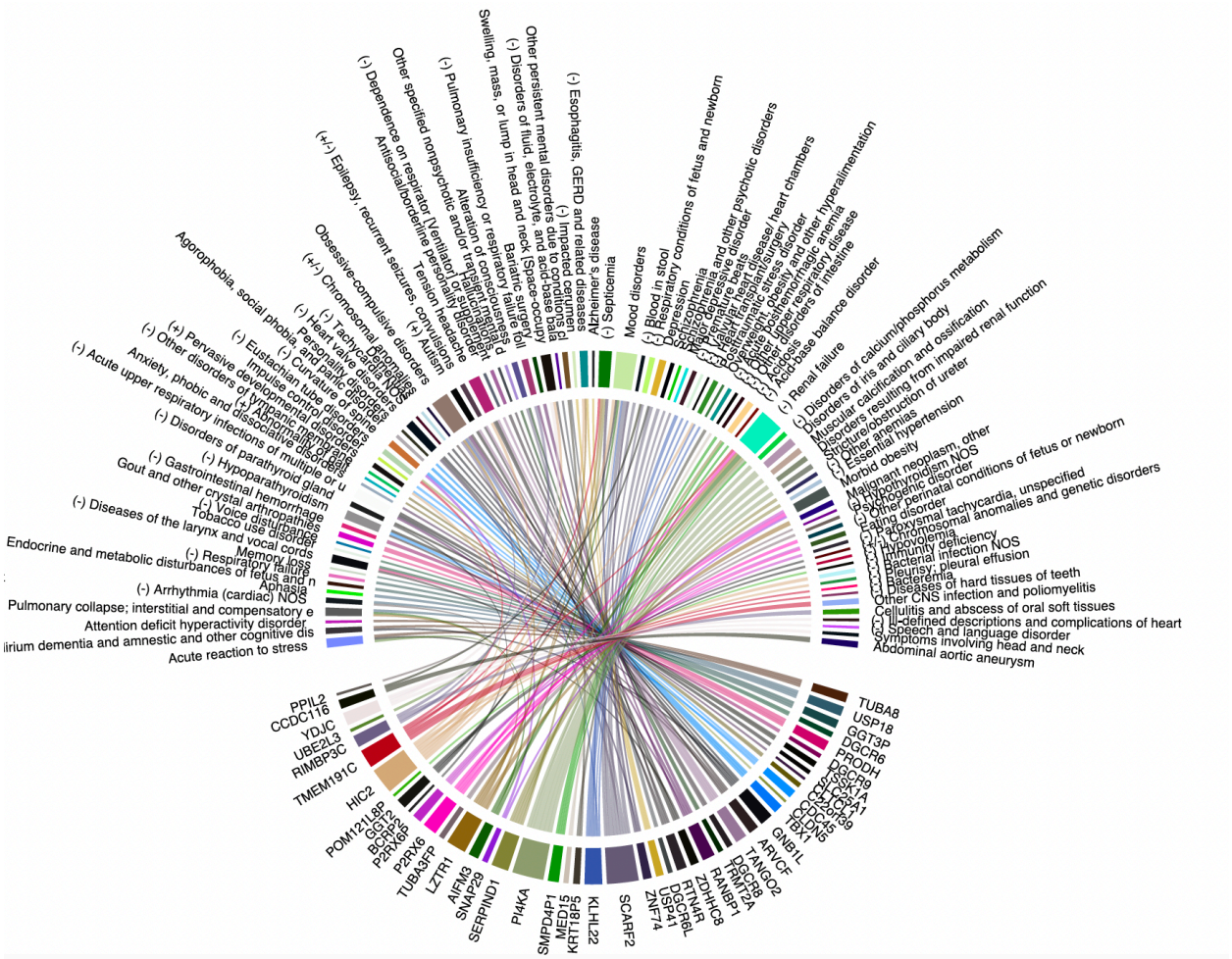


Figure 2.5. Graphical summary of selected PheWAS results by gene.

Each circle contains the CNV genes, in chromosomal order, on the bottom, and their associated PheWAS traits at the top. Genes are connected to their PheWAS-associated traits, with the width of the line proportional to the -log10 p-value of the association. If a trait is also over-represented in duplication and/or deletion carriers it is marked with a + (duplications), - (deletions), or +/- (both). The complete list of gene-trait pairs can be found in Tables 1 and 2, and Supplemental Table S7 in Additional File 14.

TABLES

Table 2.1: Selected 16p11.2 gene associations with PheWAS traits.

Possible reasons for inclusion are (1) del, dup, or del/dup: trait is over-represented in 16p11.2 deletion carriers, duplication carriers, or both ($P < 2.8x10^{-5}$); (2) brain-related trait; (3) PheWS, phenome-wide significant
[a]Phenome-wide significant gene-trait pair ($P < 3.3x10^{-5}$)
[b]not significant after conditional analysis
[c]In an independent dataset, this brain-related gene-trait pair reached $P < 0.05$ and was in the top 5% of genes associated with this trait overall

Gene	*PheWAS Trait*	*P-value*	*Reason for inclusion*
NPIPB11	Psychosis[a]	$1.04x10^{-5}$	brain-related, PheWS
	Schizophrenia and other psychotic disorders	0.0016	brain-related
	Dysphagia	0.0031	del/dup
	Infantile cerebral palsy	0.0039	dup, brain-related
BOLA2	Schizophrenia and other psychotic disorders	0.0082	brain-related
	Psychosis[b]	0.0083	brain-related
SLX1B	Psychosis[a]	$3.03x10^{-5}$	brain-related, PheWS
	Schizophrenia and other psychotic disorders	0.000606	brain-related
CA5AP1	Developmental delays and disorders	0.005	del/dup, brain-related
	Pervasive developmental disorders	0.01	del/dup, brain-related
SPN	Failure to thrive (childhood)	0.0039	dup
C16orf54	Essential hypertension[a]	$2.8x10^{-5}$	PheWS
	Bariatric surgery	0.0019	brain-related
PRRT2	Other specified nonpsychotic and/or transient mental disorders	0.0031	brain-related

Gene	*PheWAS Trait*	*P-value*	*Reason for inclusion*
	Alteration of consciousness	0.0079	brain-related
MVP	Dysphagia	0.003	del/dup
	Symptoms involving head and neck	0.0073	brain-related
CDIPT	GERD	0.0032	del
SEZ6L2	Other specified nonpsychotic and/or transient mental disorders	0.0025	brain-related
	Schizophrenia and other psychotic disorders	0.0029	brain-related
	Alteration of consciousness	0.0029	brain-related
ASPHD1	Substance addiction and disorders	0.0015	brain-related
	Upper gastrointestinal congenital anomalies	0.0044	del
KCTD13	Lack of coordination	0.0023	del, brain-related
TMEM219	Mental retardation	0.00034	del/dup, brain-related
TAOK2	Cardiomegaly	0.01	dup
HIRIP3	Acute cystitis[a]	2.9×10^{-6}	PheWS
	Disorders of uterus, NEC[a]	1.3×10^{-5}	PheWS
INO80E	Skull and face fracture and other intercranial injury	1.9×10^{-15}	brain-related, PheWS
	Substance addiction and disorders	0.0032	brain-related
	Other specified cardiac dysrhythmias	0.0034	del
FAM57B	Upper gastrointestinal congenital anomalies	0.0011	del
ALDOA	Neurological disorders	0.0014	del/dup, brain-related
	Upper gastrointestinal congenital anomalies	0.0029	del
	Antisocial/borderline personality disorder	0.0043	brain-related
	Altered mental status	0.0043	del, brain-related

Gene	*PheWAS Trait*	*P-value*	*Reason for inclusion*
	Other specified nonpsychotic and/or transient mental disorders	0.0052	brain-related
	Abnormal movement	0.007	del/dup, brain-related
	Convulsions	0.0072	dup, brain-related
TBX6	Chromosomal anomalies and genetic disorders	0.0011	del/dup
	Upper gastrointestinal congenital anomalies	0.0059	del
YPEL3	Chromosomal anomalies and genetic disorders	0.0035	del/dup
	Other specified cardiac dysrhythmias	0.0038	del
	Delayed milestones	0.0053	del/dup, brain-related
MAPK3	Substance addiction and disorders	0.00063	brain-related
	Delayed milestones	0.0014	del/dup, brain-related
	Aphasia/speech disturbance	0.0036	del, brain-related
	Psychosis[b]	0.0054	brain-related
	Upper gastrointestinal congenital anomalies	0.0092	del
CORO1A	Dysphagia	0.00034	del/dup
	Dementias	0.013	brain-related
SULT1A3	Upper gastrointestinal congenital anomalies	0.0033	del
	Obsessive-compulsive disorders	0.0042	brain-related
	Altered mental status	0.006	del, brain-related
	Swelling, mass, or lump in head and neck [Space-occupying lesion, intracranial NOS]	0.01	brain-related
CD2BP2	Substance addiction and disorders	0.0034	brain-related

Gene	*PheWAS Trait*	*P-value*	*Reason for inclusion*
	Dysphagia	0.0055	del/dup
TBC1D10B	Schizophrenia and other psychotic disorders	0.0013	brain-related
	Psychosis	0.0028	brain-related
	Alcoholic liver damage	0.0045	brain-related
	Lack of coordination	0.011	del, brain-related
MYLPF	Morbid obesity	0.0037	brain-related
ZNF48	Bariatric surgery[c]	0.0071	brain-related
SEPT1	Other specified nonpsychotic and/or transient mental disorders	0.00055	brain-related
	Alteration of consciousness	0.0018	brain-related
	Ill-defined descriptions and complications of heart disease	0.0019	dup
	Psychosis[c]	0.0035	brain-related
	Substance addiction and disorders	0.0068	brain-related

Table 2.2. Selected 22q11.2 gene associations with PheWAS traits.

Possible reasons for inclusion are (1) del, dup, or del/dup: trait is over-represented in 16p11.2 deletion carriers, duplication carriers, or both ($P < 2.8x10^{-5}$); (2) brain-related trait; (3) PheWS, phenome-wide significant

[a]Phenome-wide significant gene-trait pair ($P < 3.3x10^{-5}$)

[b]not significant after conditional analysis

[c]In an independent dataset, this brain-related gene-trait pair reached $P < 0.05$ and was in the top 5% of genes associated with this trait overall.

Gene	*PheWAS Trait*	*P-value*	*Reason for inclusion*
TUBA8	Acute reaction to stress	0.0006	brain-related
	Delirium dementia and amnestic and other cognitive disorders	0.0015	brain-related

Gene	*PheWAS Trait*	*P-value*	*Reason for inclusion*
	Attention deficit hyperactivity disorder	0.0031	brain-related
USP18	Aphasia	0.00066	brain-related
	Pulmonary collapse; interstitial and compensatory emphysema	0.00091	del
	Arrhythmia (cardiac) NOS	0.0026	del
GGT3P	Endocrine and metabolic disturbances of fetus and newborn	0.00068	del
	Respiratory failure	0.0015	del
	Memory loss	0.016	brain-related
DGCR6	Diseases of the larynx and vocal cords	0.0014	del
	Tobacco use disorder	0.0086	brain-related
PRODH	Gout and other crystal arthropathies[a]	$1.3x10^{-5}$	PheWS
	Diseases of the larynx and vocal cords	0.005893	del
	Voice disturbance	0.00801	del
DGCR9	Gastrointestinal hemorrhage	0.00016	del
TSSK1A	Hypoparathyroidism	0.0011	del
	Disorders of parathyroid gland	0.0029	del
SLC25A1	Acute upper respiratory infections of multiple or unspecified sites	0.00015	del
CLTCL1	Anxiety, phobic and dissociative disorders	0.0054	brain-related
C22orf39	Other disorders of tympanic membrane	0.0051	del
	Abnormality of gait	0.0092	dup, brain-related
CDC45	Hypoparathyroidism	0.00061	del
	Impulse control disorder	0.0035	brain-related
	Pervasive developmental disorders	0.011	dup, brain-related
CLDN5	Eustachian tube disorders	0.0078	del
TBX1	Curvature of spine	0.00083	del
	Agorophobia, social phobia, and panic disorder	0.0013	brain-related
	Personality disorders	0.0043	brain-related
GNB1L	Delirium dementia and amnestic and other cognitive disorders	0.0023	brain-related

Gene	*PheWAS Trait*	*P-value*	*Reason for inclusion*
	Heart valve disorders	0.0029	del
	Dementias	0.0047	brain-related
	Acute upper respiratory infections of multiple or unspecified sites	0.0071	del
	Tachycardia NOS	0.0074	del
ARVCF	Obsessive-compulsive disorders	0.0024	brain-related
	Diseases of the larynx and vocal cords	0.0041	del
	Chromosomal anomalies	0.0075	del/dup
	Hypoparathyroidism	0.0094	del
TANGO2	Autism	0.0011	dup, brain-related
	Tension headache	0.002	brain-related
	Antisocial/borderline personality disorder	0.0028	brain-related
	Epilepsy, recurrent seizures, convulsions	0.0049	del/dup, brain-related
DGCR8	Dependence on respirator [Ventilator] or supplemental oxygen	0.00059	del
	Hallucinations	0.0061	brain-related
TRMT2A	Other specified nonpsychotic and/or transient mental disorders	0.0033	brain-related
	Alteration of consciousness	0.0061	brain-related
RANBP1	Bariatric surgery	0.00034	brain-related
	Obsessive-compulsive disorders	0.0011	brain-related
	Pulmonary insufficiency or respiratory failure following trauma and surgery	0.0026	del
	Acute upper respiratory infections of multiple or unspecified sites	0.0035	del
ZDHHC8	Autism	0.0013	dup, brain-related
	Tension headache	0.0035	brain-related

Gene	*PheWAS Trait*	*P-value*	*Reason for inclusion*
	Acute reaction to stress	0.0049	brain-related
RTN4R	Heart valve disorders	0.0035	del
	Swelling, mass, or lump in head and neck [Space-occupying lesion, intracranial NOS]	0.0044	brain-related
	Tension headache	0.0065	brain-related
	Epilepsy, recurrent seizures, convulsions	0.0084	del/dup, brain-related
DGCR6L	Disorders of fluid, electrolyte, and acid-base balance	0.0065	del
	Other persistent mental disorders due to conditions classified elsewhere	0.0077	brain-related
USP41	Impacted cerumen	0.0026	del
	Esophagitis, GERD and related diseases	0.006	del
	Alzheimer's disease	0.0072	brain-related
ZNF74	Septicemia	0.00061	del
	Mood disorders	0.0053	brain-related
	Heart valve disorders	0.0057	del
SCARF2	Mood disorders[a,c]	1.3×10^{-5}	brain-related, PheWS
	Depression	0.00014	brain-related
	Schizophrenia	0.00027	brain-related
	Blood in stool	0.00071	del
	Obsessive-compulsive disorders	0.001	brain-related
	Alteration of consciousness	0.0011	brain-related
	Schizophrenia and other psychotic disorders	0.003	brain-related
	Major depressive disorder	0.0033	brain-related
	Respiratory conditions of fetus and newborn	0.0035	del

Gene	*PheWAS Trait*	*P-value*	*Reason for inclusion*
KLHL22	Premature beats	0.00013	del
	Valvular heart disease/ heart chambers	0.0051	del
	Overweight, obesity and other hyperalimentation	0.0064	brain-related
	Mood disorders	0.01	brain-related
	Heart transplant/surgery	0.011	del
	Posttraumatic stress disorder	0.012	brain-related
	Obsessive-compulsive disorders	0.012	brain-related
KRT18P5	Acute posthemorrhagic anemia	0.00048	del
	Other persistent mental disorders due to conditions classified elsewhere	0.0016	brain-related
MED15	Other upper respiratory disease	0.0019	del
	Mood disorders	0.0120	brain-related
SMPD4P1	Other disorders of intestine	0.001	del
	Acidosis	0.0039	del
	Acid-base balance disorder	0.0054	del
	Renal failure	0.0059	del
POM121L4P	Acute reaction to stress	0.0022	brain-related
	Convulsions	0.0072	del, brain-related
PI4KA	Disorders of iris and ciliary body[a]	1.1×10^{-7}	PheWS
	Muscular calcification and ossification[a]	7.3×10^{-6}	PheWS
	Disorders resulting from impaired renal function[a]	2.2×10^{-5}	PheWS
	Stricture/obstruction of ureter[a]	3.1×10^{-5}	PheWS
	Disorders of calcium/phosphorus metabolism	5.7×10^{-5}	del
	Renal failure	0.0007	del
SERPIND1	Other anemias	0.00044	del
	Essential hypertension	0.00045	del
	Renal failure	0.0009	del
	Acidosis	0.001	del
	Septicemia	0.0011	del
SNAP29	Curvature of spine	0.0015	del
	Morbid obesity	0.0045	brain-related

Gene	*PheWAS Trait*	*P-value*	*Reason for inclusion*
AIFM3	Renal failure[a]	2.3×10^{-5}	del, PheWS
	Pulmonary collapse; interstitial and compensatory emphysema	0.0053	del
	Mood disorders[c]	0.006	brain-related
LZTR1	Malignant neoplasm, other[a]	1.4×10^{-5}	PheWS
	Renal failure	0.00077	del
	Septicemia	0.0014	del
	Obsessive-compulsive disorders	0.0018	brain-related
	Esophagitis, GERD and related diseases	0.0054	del
	Pulmonary collapse; interstitial and compensatory emphysema	0.0056	del
TUBA3FP	Psychogenic disorder	0.0017	brain-related
	Hypothyroidism NOS	0.0074	del
P2RX6	Morbid obesity	0.00012	brain-related
	Other perinatal conditions of fetus or newborn	0.00022	del
	Renal failure	0.00067	del
	Eating disorder	0.0065	brain-related
P2RX6P	Morbid obesity[b]	0.00043	brain-related
	Paroxysmal tachycardia, unspecified	0.0014	del
	Eating disorder	0.0072	brain-related
BCRP2	Disorders of parathyroid gland	0.0078	del
GGT2	Depression	0.0038	brain-related
	Hypovolemia	0.0043	del
	Chromosomal anomalies and genetic disorders	0.0059	del/dup
	Mood disorders	0.0064	brain-related
POM121L8P	Immunity deficiency	0.0063	del
HIC2	Bacterial infection NOS	0.00023	del
	Mood disorders[c]	0.000464	brain-related

Gene	*PheWAS Trait*	*P-value*	*Reason for inclusion*
	Tension headache	0.00069	brain-related
	Swelling, mass, or lump in head and neck [Space-occupying lesion, intracranial NOS]	0.00091	brain-related
	Esophagitis, GERD and related diseases	0.002	del
	Pleurisy; pleural effusion	0.0023	del
	Posttraumatic stress disorder	0.0028	brain-related
	Pervasive developmental disorders	0.0031	dup, brain-related
TMEM191C	Other CNS infection and poliomyelitis[a]	7.2×10^{-6}	PheWS
	Eustachian tube disorders	0.0022	del
	Renal failure	0.0029	del
	Septicemia	0.0038	del
	Bacteremia	0.0073	del
	Diseases of hard tissues of teeth	0.008431	del
RIMBP3C	Cellulitis and abscess of oral soft tissues[a]	1.8×10^{-5}	PheWS
	Pulmonary insufficiency or respiratory failure following trauma and surgery	0.00047	del
	Obsessive-compulsive disorders	0.0018	brain-related
UBE2L3	Acute reaction to stress	0.0019	brain-related
YDJC	Swelling, mass, or lump in head and neck [Space-occupying lesion, intracranial NOS]	0.00025	brain-related
	Symptoms involving head and neck	0.00072	brain-related
	Ill-defined descriptions and complications of heart disease	0.0027	del
	Speech and language disorder	0.0042	del, brain-related
CCDC116	Abdominal aortic aneurysm[a]	1.9×10^{-6}	PheWS
	Respiratory conditions of fetus and newborn	0.0032	del
PPIL2	Arrhythmia (cardiac) NOS	0.006	del

Supplementary material for this work that is referenced in this chapter can be found at https://doi.org/10.1186/s13073-021-00972-1

Chapter three: Neurobehavioral traits are driven by combinations of genes at 16p11.2 and 22q11.2

ABSTRACT

Background

The 16p11.2 and 22q11.2 copy number variants (CNVs) are associated with neurobehavioral traits including autism spectrum disorders (ASD), schizophrenia, bipolar disorder, obesity, and intellectual disability. Identifying specific genes contributing to each disorder and dissecting the architecture of CNV-trait association has been difficult, inspiring hypotheses more complex models, such as the effects of pairs of genes.

Methods

To model pairs of CNV genes upregulated or downregulated in the same direction (as would occur in CNV carriers), we trained elastic net prediction models using SNPs to impute summed gene expression across two genes in control individuals, and then applied these prediction models to large GWAS cohorts for five traits: ASD, bipolar disorder, schizophrenia, BMI (obesity), and IQ (intellectual disability). We compared the adjusted R2 values of the associations between each trait and these imputed pairs across the region with the adjusted R2 values of the trait association with single genes across the region and with traditional interaction models. To analyze region-wide effects, we ranked predicted expressions of single genes and summed ranks for an individual across the genes in the region to comprise a score. We compared case-control score distributions and calculated the correlation between regionwide score and quantitative traits.

Results

We found that in all CNV-trait pairs except for bipolar disorder at 22q11.2, pairwise effects explain more trait variance than single genes, although for schizophrenia and IQ in 22q11.2 this was not CNV-specific. We observed three patterns for individual gene frequency of being in significant pairs: similar set of genes contributing to single and pairwise associations, different genes contributing, and one gene contributing disproportionately. We also found that BMI and IQ have a significant association with the regionwide score.

Conclusions

Insights into the mechanisms of CNV pathogenicity might result from studying combinations of the genes in and near these CNVs. The genetic architecture differs by trait and region, but nine of the 10 CNV-trait combinations we investigated showed greater variance explained by pairwise models than single genes, and two traits showed regionwide signal. The importance of combinatorial contributions appears to be unique to CNV regions in 7/9 examples and did not extend to well-matched control regions for the same traits.

INTRODUCTION

Copy number variants (CNVs) at 16p11.2 and 22q11.2 contribute to neurobehavioral disorders including autism spectrum disorder (ASD), schizophrenia, bipolar disorder, intellectual disability, and obesity [1–11]. Specific gene-trait contributions at these regions have proven difficult to find. Single-gene fine-mapping approaches have been difficult due to a lack of highly-penetrant point

mutations in these genes and inconsistent findings in animal models [12–15]. A potential source for the lack of clear gene-phenotype relationships is that the architecture may be more complicated than single-gene contributions to each trait [16]. More complex models are good candidates for *in silico* analysis, as multiple hypotheses can be efficiently assessed in parallel.

Data in humans and mice suggest that the expression of 16p11.2 and 22q11.2 CNV genes is consistently upregulated/downregulated in duplication/deletion carriers [17–20]. From this observation, we can propose that gene expression dysregulation (and potential downstream protein expression) is likely to be a pathophysiological mechanism of CNV-associated traits. This implies that examination of the consequences of gene expression variation may uncover the genetic architecture of CNV-phenotype association. However, gene expression data for cases affected with neurobehavioral traits remains limited in availability and ambiguous with respect to causality. Instead, we can use expression-imputation methodology to use genetic data, available for a far greater number of (control) individuals, to determine gene expression under the assumption that genetic regulation is similar in cases and controls. This method allows us to analyze the architecture at a gene level (rather than individual SNPs) and because it is based on germline genetics, is not affected by potential confounding influences on gene expression such as age, chronic illness, medication use, and circumstances of death and tissue preservation. eQTLs (in our case, SNPs used for expression prediction) are less likely to affect genes in a context-dependent manner, as eQTL-linked genes are less likely to be affected by enhancer activity compared to GWAS-linked genes [21]. Given that our regions of interest have trait associations via CNVs but very limited GWAS signal for the same traits, using eQTLs and expression prediction is likely to find additional information missed by GWAS analyses.

Previously, we used expression imputation to test whether individual genes at the 16p11.2 and 22q11.2 CNV regions were contributing to our five traits of interest (ASD, schizophrenia, bipolar disorder, intellectual disability, and obesity) [22]. We found contributions of *INO80E* to schizophrenia and body mass index (BMI) and of *SPN* to BMI and IQ, both at 16p11.2. However, no individual genes were associated with 22q11.2 traits, despite using equally-powered genetic datasets. No genes at 16p11.2 were significantly associated with ASD or bipolar disorder using our experiment-wide threshold. These lack of findings in light of the overall success of our approach were disappointing given the high prevalence of traits such as ASD in 16p11.2 CNV carriers and schizophrenia in 22q11.2 deletion carriers. One explanation for lack of gene-trait association is that individual genes may not be independent contributors to these traits, rather the genetic architecture is combinatorial. Promisingly, it was found that several pairs of 16p11.2 genes in *Drosophila* showed evidence of stronger effects on eye phenotypes than individual genes, and double mutants of 16p11.2 genes in zebrafish led to hyperactivity and body size phenotypes [15,23]. Thus, we aimed to investigate combinatorial associations in our traits of interest in humans.

In a CNV carrier, all genes within the breakpoints are duplicated or deleted, typically with a similar increase/decrease of expression across all genes (**Figure 3.1**). In our previous study, we considered the level of expression of any individual gene, and its effect on relevant phenotypes in non-carriers. Here, we consider two additional models in non-carriers. First, as a feasible way to model multigene effects at specific pairs of genes, for each gene pair we look for trait association with expression increases or decreases across two genes. Second, we analyze

association patterns when gene expression trends towards being upregulated or downregulated across the whole region as a way to capture effects of more than two genes (**Figure 3.1**).

METHODS

Genes studied

We selected genes at the 16p11.2 and 22q11.2 CNV regions that fell into one of these annotation categories: protein-coding, lincRNA, pseudogene, antisense, miRNA. These were consistent with what was used for PrediXcan modeling previously, with miRNA included given the strong representation of miRNAs at 22q11.2 [24,25]. We included noncoding genes, as they have not received significant attention in studies of these regions, despite some evidence of miRNA contribution to 22q11.2 phenotypes. In addition, we considered flanking genes within 200kb of the region, as we previously found evidence of flanking gene involvement in psychosis, while others noticed that a 16p11.2 deletion has transcriptional impacts throughout the 16p chromosome arm [22,26,27].

Phenotypes and datasets

Imputed genotypes from the Psychiatric Genomics Consortium were used to study schizophrenia (wave 3 freeze), bipolar disorder (wave 2 freeze), and ASD (2019 freeze, used for analysis of variance explained only) [28–30]. An additional joint PGC-iPsych ASD summary statistic set was used to boost power for ASD analyses (www.med.unc.edu/pgc/download-results/) [30]. Summary statistics from the GIANT consortium (2015 freeze, www.portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files)

were used to study BMI, and a VU-Amsterdam University cohort (wave 2 freeze, www.ctg.cncr.nl/software/summary_statistics) was used for IQ [31,32]. Individual-level IQ and BMI data from the UK Biobank were used for validating discoveries in individual-level data on phenotypes for which individual-level data were not available [33].

Predicting the expression levels of individual 16p11.2 and 22q11.2 CNV genes

Analyses of single genes were performed using elastic net models from www.predictdb.org trained on the GTEx version 8 data [34]. These prediction models were available for up to 42 16p11.2 genes and up to 65 22q11.2 genes in at least one tissue. The elastic net approach was chosen for consistency with our pairwise model training approach. Gene expression for each CNV gene in each individual was predicted using the --predict option in PrediXcan, for each cohort [35]. Analyses on summary statistics did not require expression prediction.

Finding control gene sets

To create control gene sets to use in a permutation-based analysis, the 16p11.2 and 22q11.2 regions were matched on three categories: (1) number of genes (exact), (2) length of the region (in bases, 80-120% of the region), (3) ratio of coding to non-coding genes (at least 80% that of the region to avoid picking up dense regions of noncoding genes). Gene sets that overlapped the distal 16p11.2 region or the Major Histocompatibility Complex (a known gene-dense major GWAS-identified locus for schizophrenia) were excluded [36]. Overall we found 41 comparable regions to 16p11.2 and 29 to 22q11.2.

Predicting the expression of pairs of 16p11.2 and 22q11.2 CNV genes

As a simple way to model pairwise gene expression, we took every pair of genes in each CNV and defined pairwise "joint genes" with expression equal to the inverse-normalized sum of the expressions of each gene in GTEx. Normalized GTEx gene expression sums were used as inputs to the PrediXcan elastic net model training pipeline (www.github.com/hakyimlab/PredictDB_Pipeline_GTEx_v7), with covariates used for the GTEx v8 analyses downloaded from www.gtexportal.org/home/datasets. Given that our goal was to evaluate the contribution of these pairwise genes to specific traits, rather than a general-use pairwise model training process, a high overlap between the SNPs in our models and the GWAS datasets was vital. For that reason, we chose to repeat the training process for each trait, leaving only the SNPs in each GWAS dataset as inputs for model training. We repeated this model training process again on the control pairs of genes.

Association studies between predicted expression and traits

Individual level

Each PGC cohort was converted from PLINK to dosage format for PrediXcan input. Tissue-specific prediction models were applied to each tissue in each cohort. MultiXcan, a cross-tissue implementation of PrediXcan, was used to combine predicted expressions across tissues and perform association with trait [37]. Using the MultiXcan p-value and the average direction of effect of each gene across tissues, we used METAL to determine a per-gene result [38]. Both single gene and pairwise analyses were performed in the same way.

Summary level

The ‘MetaMany’ option in the MetaXcan package was applied to summary-level data using single-tissue prediction models to generate gene-level association results for each tissue [39]. S-MultiXcan, a cross-tissue implementation of PrediXcan for summary level data, was used to combine across tissues for tissue-wide association results [37]. Cross-tissue covariances were downloaded from PredictDB for single-gene models and generated from single tissue covariances for pairwise models using the covariance builder script available at www.github.com/hakyimlab/MetaXcan/blob/master/software/CovarianceBuilder.py. Both single gene and pairwise analyses of summary statistics were performed in the same way.

UK Biobank additional expression prediction

While the best-powered GWAS meta-analyses of BMI and IQ were available as summary statistics, certain analyses such as interaction models and percent variance explained require individual-level data. We obtained IQ and BMI measurements from the UK Biobank and took an average across visits for people with multiple measurements. Analysis was limited to individuals who reported their ethnicity as “white”, and included age, age-squared, and 40 principal components as covariates. A large number of principal components was used due to the correlation between the BMI phenotype and components in the PC 30-40 range. Expression imputation for single genes and pairs was performed with PrediXcan as described above.

Significance thresholds for association studies

For all association studies, a permutation-based threshold was determined using the control gene sets. After association testing between control gene sets for a CNV and phenotype, the median of the 5th percentile of control sets was used as a 5% significance threshold for the true CNV region. As control genes are chosen independently of association with trait, using a median across all regions will counteract bias caused by any control gene set overlapping a strong GWAS peak for a trait.

Estimating variance explained by pairwise models

Variance in phenotype explained by imputed expression was measured as the R^2 of the linear model between case-control status and imputed expression for all genes in the CNV. Specifically, the adjusted R^2 was used, as using all pairs of genes involves a large number of variables. For every tissue-cohort pair, R^2 values were calculated using all single genes, all pairwise genes, and interaction terms. The number of times a model (single, pairwise, or interaction) had the greatest R^2 for a cohort-tissue pair was tallied. The same process was implemented for control gene sets. A chi-square test was performed to determine whether the proportion of best pairwise models being "best" in a CNV region was different from the proportion in control regions. This approach required individual-level data, and as we used summary level data for ASD, IQ, and BMI, we performed it in PGC ASD individual-level data (without iPSYCH), and UK Biobank for IQ and BMI (each of which was treated as one single cohort).

We acknowledge that previous attempts to solve the problem of variance explained by predicted expression were made by Liang et al [40]. We attempted this method and found extremely large estimates for variance explained. This inflation might be due to our relatively small (<5 MB) regions of interest with high SNP and predicted-expression correlation structure, as opposed to a predicted transcriptome-wide screen. The estimates provided by our approach, where the adjusted R^2 rarely exceeds 0.05, are a more reasonable estimate of the effect of one small set of genes on a trait.

Testing a region-wide model

We estimated a region-wide score for each individual using their single-gene predicted expressions. First, we found the normalized rank of an individual for the expression of a gene; the median rank was used for genes expressed in multiple tissues. The sum of an individual's gene-specific rankings became the individual's region-wide score; these scores were converted to normalized (between 0 and 1) ranks. For quantitative traits, we quantified the relationship between score and phenotype as a Pearson correlation. For binary traits, we tested for a difference in score distribution between cases and controls (Kolmogorov-Smirnov test), as well as for a difference in score means between cases and controls (t-test).

In addition, we attempted to approach region-wide association in the same way as pairwise association for schizophrenia. Region-wide sums of GTEx expressions of all CNV genes were used as inputs into elastic net models from GTEx SNPs, with the same covariates as before. After model quality filtering, models in only 5 tissues at 16p11.2 and 13 tissues at 22q11.2 remained, all with $R^2 < 0.1$. As a result, we did not further pursue this method.

RESULTS

Summary of individual gene results

We have updated our single-gene prioritization from our previous study using new models from GTEx version 8 and new data from schizophrenia PGC wave 3 [28,34]. With this enhancement, we find one 22q11.2 gene (*PPIL2*) significantly associated with schizophrenia at a permutation-based threshold. We note that the permutation-based threshold is less conservative than the experiment-wide thresholds used in previous analysis [22]. However, we can identify five top genes at 22q11.2 associated with BMI (*YDJC, CCDC116, PPIL2, THAP7, UBE2L3*), primarily located outside the canonical CNV region (LCR D-E), three with bipolar disorder (*TMEM191B, TUBA8, PPIL2*), six with ASD (*CLTCL1, AC004471.10, UFD1L, DGCR14, CCDC188, DGCR9*), and two with IQ (*SEPT5, LINC00896*). The top genes associated with ASD at 22q11.2 are located in the LCR A-B part of the variant, consistent with a previous study [41]. At 16p11.2, the majority of genes tested (30/38) show an association with BMI. We find that, after updating single-gene prediction models to GTEx v8, *SPN* is no longer a major driver of BMI and IQ, as the best predictive SNPs in the most up-to-date version of GTEx did not overlap with top *SPN* SNPs as before; however, new models for *SULT1A4* indicated this gene as a major contributor to both BMI and IQ. *INO80E* and *KCTD13* remained associated with BMI. We find that *INO80E* is a top association with bipolar disorder and ASD; this gene previously showed suggestive bipolar disorder association but did not meet experiment-wide significance criteria even with the updated models [22].

Predicting expression of pairs of 16p11.2 and 22q11.2 genes

We trained elastic net models for pairs of 16p11.2 and 22q11.2 genes (both coding and non-coding when possible) using SNPs from each dataset's SNP list to maximize overlap. **Table 3.1** shows the number of gene pairs that had high-quality predictions and were used in analysis. In general, the model quality (as measured by the performance R^2) of pairwise models was in-between that of the two genes that it comprised. In addition, we trained pairwise models for control gene regions (N=41 for 16p11.2 and N=29 for 22q11.2).

Pairwise prediction models explain more trait variance than single-gene or interaction models

To assess whether that analyzing pairs of genes provided more information than individual genes, we calculated how much variance in CNV-associated traits was explained by predicted gene expression as the adjusted R^2 of linear models of individual gene expression predictions, pairwise additive gene expression predictions, and pairwise interaction models. We calculated the proportion of tissue-cohort pairs for which pairwise gene expression was the best predictor. In all trait-region pairs, with the exception of bipolar disorder at 22q11.2, we found that the trait variance explained was greater for gene pairs proportionally more often than either single genes or interactions (**Table 3.2**). To confirm that this phenomenon was CNV region-specific and not a polygenic property of the trait, we additionally performed this analysis for control gene sets (**Table 3.2**). For all traits tested at 16p11.2, the proportion of pairwise models exceeding single or interaction was greater than that of control regions ($P < 0.05$). At 22q11.2, the CNV region performed better than controls in ASD ($P = 2.2\text{x}10^{-9}$) and BMI ($P < 2.2\text{x}10^{-16}$), but schizophrenia, bipolar disorder, and IQ did not have a greater proportion pairwise in the region as opposed to controls.

Pairwise association signal is oligogenic

Using our pairwise models to perform association analysis, we found that there were 269 16p11.2 and 278 22q11.2 pairs significantly associated with ASD, 204 16p11.2 and 132 22q11.2 pairs associated with bipolar disorder, 695 16p11.2 and 129 22q11.2 pairs associated with schizophrenia, 74 16p11.2 and 30 22q11.2 pairs associated with IQ and 1206 16p11.2 and 162 22q11.2 pairs associated with BMI. The proportion of gene pairs exceeding the significance threshold was consistent with that of single genes, and in the cases where the proportions differed (such schizophrenia and IQ at 16p11.2, BMI and IQ at 22q11.2), the pairwise analysis had the lower proportion of significantly associated genes/pairs. We thus find that pairwise association signal is oligogenic, spread across many pairs rather than enrichment specific to top outlier results. Due both to the eQTL sharing between pairwise prediction models as well as to the sharing of genes across pairs, we are unable to use our approach to confidently identify specific candidate gene pairs; several pairs of potential interest are noted in the Discussion section.

Patterns of genes most represented in associated pairs differ by phenotype

We wanted to know whether the pairwise associations were primarily comprised of genes with independent association signal or indicated genes with uniquely combinatorial effects. The results were strikingly different across traits (**Figure 3.2**). For example, schizophrenia and bipolar disorder for 16p11.2 had a large overlap between the top single genes and genes involved in the most top pairs; both traits were primarily driven by pairs involving *INO80E.* BMI showed a similar pattern, with the top single gene result (*SULT1A4*) appearing frequently in top pairs.

Genes that frequently appeared in top pairs of ASD and IQ did overlap with top single genes, but many other genes appeared in pairs at similar frequency. Though *INO80E* was not a top ASD single gene, it showed strong contribution to pairs. Both schizophrenia and bipolar disorder 22q11.2 top pairs were dominated by those which involved *PPIL2*; this flanking gene was the only single-gene association with schizophrenia and one of three for bipolar disorder. *DGCR2* and *DGCR6* contributed to ASD through pairs, but were not single gene associations. Although our pairwise analyses included a greater number of non-coding genes, we did not find that non-coding genes were individually common contributors to pairs at either 16p11.2 or 22q11.2.

Region-wide contributions of 16p11.2 and 22q11.2 CNVs to phenotype

After comparing the impacts of single genes and pairs of CNV genes on neurobehavioral traits, we wanted to test combinations greater than pairwise, but feasibility limited our combinatorial testing. Therefore, we considered a broad region-wide model: whether the average deviation of the multigenic region contributes to a phenotype. We assigned a region-wide score to each individual and tested whether scores were significantly different between cases and controls or correlated with quantitative traits (**Figure 3.3**). We found that the region-wide score was positively correlated with BMI for 16p11.2 genes ($P = 2.0\text{x}10^{-11}$) and negatively correlated for 22q11.2 genes ($P = 0.0001$). IQ was also negatively correlated with region-wide score for 16p11.2 genes ($P = 8.7\text{x}10^{-15}$). None of the categorical traits showed an effect of a region-wide contribution.

DISCUSSION

Our study aimed to provide insight into the genetic architecture of the 16p11.2 and 22q11.2 copy number variants. **Figure 3.4** summarizes our findings. We modeled the neurobehavioral trait consequences of pairs of genes expressed in the same direction, extending our previous single-gene analysis. We found that for nearly all traits tested, variance in phenotype was better explained by pairs of genes than by single genes or traditional interaction models. The only exception was bipolar disorder at 22q11.2, where single genes explain more variance. However, for schizophrenia and IQ at 22q11.2 the pairwise model was not specific to the CNV regions but appeared to be a trait-based property extending to control regions with similar properties. The advantage of summed pair models in control regions compared to interaction models was somewhat surprising due to our hypothesis that CNV regions have the unique property of dysregulation of nearby genes in the same direction. However, perhaps regulatory landscape across many regions of the genome also biases towards expression dysregulation of physically colocalized genes in the same direction.

As we observed neither enrichment in the proportion of significant pairwise tests nor outlier top signal in the QQ plots, the pairwise contribution to explaining trait variance seems to be oligogenic across the region. However, in some cases we did observe outliers when examining the frequency of specific genes involved in top pairs. There was striking variation across traits and regions in terms of whether the top single genes were also the top contributors to pairs or novel genes were equally likely to contribute. A single gene was repeatedly contributing to top pairs for bipolar disorder at 16p11.2 (*INO80E*, 26% of top pairs) and schizophrenia at 22q11.2 (*PPIL2*, 42% of top pairs*)*. The individual association with these genes was not detected, but the

recurrent role of these genes in pairs suggests an important trait contribution. In contrast, for schizophrenia at 16p11.2 and ASD at 22q11.2, multiple top single genes participate disproportionately in top pairs. Intriguingly, although pairwise models show similar advantages for ASD at 16p11.2 and IQ at 22q11.2, genes across the region are more evenly represented in top pairs. Bipolar disorder at 22q11.2 (with single genes models most often explaining variance) showed association with flanking genes on either side, *TUBA8*, *TMEM191B*, and *PPIL2*; *PPIL2* appeared in most of the pairs, as well. Because we did not find overall support for a pairwise model for bipolar disorder at 22q11.2, this may simply reflect the independent association of *PPIL2*. Our finding of *PPIL2* as a bipolar disorder driving gene is supported by this gene's over-representation of rare protein truncating variants in the Bipolar Exome sequencing consortium data [42]. We note that evidence for the association of 22q11.2 with bipolar disorder is weaker than that of schizophrenia and ASD [43]. If 22q11.2 does not drive bipolar disorder, our inability to find single or pairwise signal is consistent with biology, and the contribution of *PPIL2* to bipolar disorder are independent of copy number changes.

Given that the pairwise signal tended to be oligogenic and that expression imputation of adjacent genes has high correlation, it is difficult to confirm the association of specific pairs of genes. For ASD at 16p11.2, the top 15 pairs include four with *FAM57B*. This gene was previously shown to have multiple within-region interactions in zebrafish [44]. Here, we find that the top pairwise contributions are with coding and non-coding genes in repetitive or flanking regions (*RP11-347C12.3, TBC1D10B, BOLA2B, NPIPB12*). Studies of 16p11.2 CNV genes rarely include these flanking genes, but our data suggest that they may contribute to trait association. Notably, our expectation of expression dysregulation in the same direction would be less strong for flanking

genes, so expanded testing of flanking regions may be worthwhile. The *FAM57B* and *DOC2A* pair driving hyperactivity, head size, and length in the zebrafish study was in the top quarter of associated pairs for BMI and IQ. We note that McCammon *et al* specifically excluded additive effects, while our study is based on genes contributing additively to pairs (which we find explains more variance than traditional interactions). For BMI at 16p11.2, the top ranked pair is *CDIPT* with *ALDOA*. It is notable that these two genes were not top-ranked individual genes for BMI, demonstrating the utility of our pairwise approach to prioritize pairs that might not be detected as individual genes. The top pair for IQ, *MVP* and *KCTD13*, on the other hand, includes one top IQ-associated gene (*MVP*) and one gene (*KCTD13*) not associated with IQ. This finding is similar to an observation in zebrafish, where the expressivity of head-size phenotypes driven by *KCTD13* overexpression was increased by additional overexpression of *MVP* [13]. For IQ at 22q11.2, several top pairs contain *COMT* along with a non-coding gene. *COMT* is a gene with variants believed to affect multiple traits, including IQ [45], and whose expression is associated with IQ in the general population [46].Our data provides a refined hypothesis that the relationship between *COMT* and IQ is dependent on additional non-coding genes at 22q11.2.

We also wondered whether there was a general contribution across many genes in the region. In our analyses, we found that there was a region-wide contribution to both BMI and IQ in both CNVs. The large number of 16p11.2 genes associated with BMI in both single and pairwise models was consistent with a region-wide signal. From previously established associations in CNV carriers, we would expect a negative correlation between increased expression and BMI for both 16p11.2 and 22q11.2 CNVs. However, we saw this only at 22q11.2 in the region-wide model. Previously, we found individual genes independently associated with both increases and

decreases at BMI at 16p11.2 [22]. We hypothesized that there may be both BMI-increasing and BMI-decreasing genes in the 16p11.2 region due to our observation of association in both directions in single-gene models (and BMI decreases in syntenic deletion mice [12]), in which case we might have been picking up more BMI-decreasing genes in our region-wide score. However, one potential limitation of our cross-tissue expression prediction approach is that our results may not be driven by the biologically-relevant tissues and thus appear to be opposite in direction [47]. We also note that BMI and IQ are quantitative traits with high sample size, and so we likely had power limitations in other traits.

Previously, we proposed that *INO80E* at 16p11.2 is a driver of schizophrenia and BMI, a finding that has been corroborated by similar analyses by others. However, we found that pairwise models explained more trait variance in both schizophrenia and BMI at 16p11.2, so it is possible that the pathophysiological contribution of *INO80E* will be better explained in combination with other genes than independently, a hypothesis that might be of interest for experimental design. Our pairwise findings also suggest that *INO80E* has an important contribution to at least two other traits. In bipolar disorder, *INO80E* is the top individual associated gene and is the most disproportionate contributor to pairs. In ASD, *INO80E* is not a top individual gene but is the most frequent (albeit not strongly disproportionate) contributor to significant pairs. This finding suggests that four traits may be influenced by the *INO80E* gene, and at least in the case of ASD, this gene works in combination with other genes. However, we have not found evidence of the involvement of *INO80E* in IQ, showing that the neurobehavioral phenotypes of 16p11.2 may be broader than the impact of this single gene, under the assumption that IQ in the general population is a good representation of the 16p11.2-mediated impact on intellectual ability.

There are a number of limitations in our approach to probing the architecture of 16p11.2 and 22q11.2 CNVs using pairs of gene expression predictions and region-wide gene expression scores. There are numerous combinatorial models that have not been tested, and the true architecture of gene-trait pairs may lie anywhere in between what we can capture in simplified models. In fact, given the observation that the entire 16p chromosome arm is enriched for ASD risk signal and has high amount of chromosomal contact, the region itself, as we had defined it, could be insufficient [26,48].

Another potential model that we have not tested is that only the extremes of the distribution – either in pairwise sums or region-wide scores – will impact a phenotype, and more modest increases and decreases in gene expression are buffered. For example, the BMI-16p11.2 panel in Figure 3 suggests a difference in the top and bottom decile compared to the BMI-score relationship in the intermediate deciles. Our study using all individuals has an advantage in statistical power if more typical gene expression levels are relevant to the trait, but a disadvantage in the potential noise that is introduced if only extreme expression deviation is relevant to uncommon traits such as schizophrenia, bipolar disorder, and ASD. This is motivated by a previous study we performed in a clinical biobank. There, we identified traits over-represented in CNV carriers, as well as traits associated with the predicted expression of CNV genes. We found that when we restricted only to the individuals with extreme predicted expressions, their associated traits were better at matching traits over-represented in true CNV carriers, compared to people with more "average" predicted expression levels. Further motivation for focusing on extremes in follow-up analyses comes from an observation that

predicting expression in a quadratic rather than linear way improves gene identification in TWAS [49].

A technical limitation of our study design is that available datasets are not always ideal for our approach. For BMI, IQ, and ASD, the best-powered datasets are summary statistics. We use the summary statistics for single and pairwise association testing, determining permutation-based significance cutoffs, and finding top individual genes that are represented in pairs. However, in order to measure variance explained and region-wide scoring, we use individual-level data. We have to consider heterogeneity across the cohorts as a caveat when comparing results. Still, for both ASD and IQ the individual level data used is a subset of the full cohort comprising summary level statistics, minimizing the differences. Finally, our study is based on multiple tissues derived from adults, rather than more targeted analyses of the brain at early development. Similarly, when we decide which model explains more variance, we do not weight tissues differently (according to trait relevance, sample size, etc.). Despite the limitations, we may be detecting signal driven by a subset of the data; for example, ASD-donor cerebral organoids show cell-type specificity of *INO80E* to neuroepithelial cells during development, yet we detect a pairwise contribution in cross-tissue analysis [50].

The 16p11.2 and 22q11.2 regions are highly penetrant for neurobehavioral traits, but require a better understanding of genetic architecture to indicate key biological pathways. By extending transcription imputation to study a simple summed model of pairwise gene expression, we uncover a consistent pattern of higher variance explained by gene pairs than either single genes or traditional interaction models and several traits showing region-wide association signal. Most

of these patterns appear specific to CNV regions and did not appear to represent the genetic architecture in matched control regions. Our study suggests that pathobiological insights might result from studying combinations of the genes in and near these CNVs, albeit with potentially differing genetic architecture across traits and regions.

REFERENCES

1. Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z, et al. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature. 2011;478:97–102.

2. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, et al. Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet. 2009;41:1223–7.

3. Carlson C, Papolos D, Pandita RK, Faedda GL, Veit S, Goldberg R, et al. Molecular analysis of velo-cardio-facial syndrome patients with psychiatric disorders. Am J Hum Genet. 1997;60:851–9.

4. Shinawi M, Liu P, Kang S-HL, Shen J, Belmont JW, Scott DA, et al. Recurrent reciprocal 16p11.2 rearrangements associated with global developmental delay, behavioural problems, dysmorphism, epilepsy, and abnormal head size. J Med Genet. 2010;47:332–41.

5. Kumar RA, KaraMohamed S, Sudi J, Conrad DF, Brune C, Badner JA, et al. Recurrent 16p11.2 microdeletions in autism. Hum Mol Genet. 2007;17:628–38.

6. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, et al. Association between Microdeletion and Microduplication at 16p11.2 and Autism. N Engl J Med. 2008;358:667–75.

7. Bassett AS, Chow EW. 22q11 deletion syndrome: a genetic subtype of schizophrenia. Biol Psychiatry. 1999;46:882–91.

8. Campbell IM, Sheppard SE, Crowley TB, McGinn DE, Bailey A, McGinn MJ, et al. What is new with 22q? An update from the 22q and You Center at the Children's Hospital of Philadelphia. Am J Med Genet Part A. 2018;176:2058–69.

9. Wentzel C, Fernström M, Öhrner Y, Annerén G, Thuresson A-C. Clinical variability of the 22q11.2 duplication syndrome. Eur J Med Genet. 2008;51:501–10.

10. Schneider M, Debbané M, Bassett AS, Chow EWC, Fung WLA, van den Bree MBM, et al. Psychiatric Disorders From Childhood to Adulthood in 22q11.2 Deletion Syndrome: Results From the International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome. Am J Psychiatry. 2014;171:627–39.

11. Voll SL, Boot E, Butcher NJ, Cooper S, Heung T, Chow EWC, et al. Obesity in adults with 22q11.2 deletion syndrome. Genet Med. 2017;19:204–8.

12. Pucilowska J, Vithayathil J, Tavares EJ, Kelly C, Colleen Karlo J, Landreth GE. The 16p11.2 deletion mouse model of autism exhibits altered cortical progenitor proliferation and brain cytoarchitecture linked to the ERK MAPK pathway. J Neurosci. 2015;35:3190–200.

13. Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S, et al. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature. 2012;485:363–7.

14. Blaker-Lee A, Gupta S, McCammon JM, De Rienzo G, Sive H. Zebrafish homologs of genes within 16p11.2, a genomic region associated with brain disorders, are active during brain development, and include two deletion dosage sensor genes. Dis Model Mech. 2012;5.

15. Iyer J, Singh MD, Jensen M, Patel P, Pizzo L, Huber E, et al. Pervasive genetic interactions modulate neurodevelopmental defects of the autism-Associated 16p11.2 deletion in Drosophila melanogaster. Nat Commun. 2018;9:1–19.

16. Jensen M, Girirajan S. An interaction-based model for neuropsychiatric features of copy-number variants. PLoS Genet. 2019;15.

17. Ward TR, Zhang X, Leung LC, Zhou B, Muench K, Roth JG, et al. Genome-wide molecular effects of the neuropsychiatric 16p11 CNVs in an iPSC-to-iN neuronal model. bioRxiv. 2020;2020.02.09.940965.

18. Blumenthal I, Ragavendran A, Erdin S, Klei L, Sugathan A, Guide JR, et al. Transcriptional Consequences of 16p11.2 Deletion and Duplication in Mouse Cortex and Multiplex Autism Families. Am J Hum Genet. 2014;94:870–83.

19. Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, et al. Genome-wide Transcriptome Profiling Reveals the Functional Impact of Rare De Novo and Recurrent CNVs in Autism Spectrum Disorders. Am J Hum Genet. 2012;91:38–55.

20. Zhang X, Zhang Y, Zhu X, Purmann C, Haney MS, Ward T, et al. Local and global chromatin interactions are altered by large genomic deletions associated with human brain development. Nat Commun. 2018;9.

21. Mostafavi H, Spence JP, Naqvi S, Pritchard JK. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. bioRxiv. 2022;

22. Vysotskiy M, Zhong X, Miller-Fleming TW, Zhou D, Cox NJ, Weiss LA. Integration of genetic, transcriptomic, and clinical data provides insight into 16p11.2 and 22q11.2 CNV genes. Genome Med 2021 131. 2021;13:1–26.

23. McCammon JM, Blaker-Lee A, Chen X, Sive H. The 16p11.2 homologs fam57ba and doc2a generate certain brain and body phenotypes. Hum Mol Genet. 2017;26:3699–712.

24. Forstner AJ, Degenhardt F, Schratt G, Nöthen MM. MicroRNAs as the cause of schizophrenia in 22q11.2 deletion carriers, and possible implications for idiopathic disease: a mini-review. Front Mol Neurosci. 2013;6:47.

25. Barbeira AN, Melia OJ, Liang Y, Bonazzola R, Wang G, Wheeler HE, et al. Fine-mapping and QTL tissue-sharing information improves the reliability of causal gene identification. Genet Epidemiol. 2020;44:854–67.

26. Weiner DJ, Ling E, Erdin S, Tai DJC, Yadav R, Grove J, et al. Statistical and functional convergence of common and rare variant risk for autism spectrum disorders at chromosome 16p. medRxiv. 2022;13:2022.03.23.22272826.

27. Tai DJC, Razaz P, Erdin S, Gao D, Wang J, Nuttle X, et al. Tissue and cell-type specific molecular and functional signatures of 16p11.2 reciprocal genomic disorder across mouse brain and human neuronal models. bioRxiv. 2022;2022.05.12.491670.

28. Consortium TSWG of the PG, Ripke S, Walters JT, O'Donovan MC. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. medRxiv. 2020;2020.09.12.20192922.

29. Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. Nat Genet. 2019;51:793–803.

30. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51:431–44.

31. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518:197–206.

32. Savage JE, Jansen PR, Stringer S, Watanabe K, Bryois J, De Leeuw CA, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat Genet. 2018;50:912–9.

33. UK Biobank — Neale lab [Internet]. [cited 2020 Mar 28]. Available from: http://www.nealelab.is/uk-biobank

34. Barbeira AN, Bonazzola R, Gamazon ER, Liang Y, Park YS, Kim-Hellmuth S, et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Genome Biol. 2021;22:1–24.

35. Gamazon ER, Wheeler HE, Shah KP, Mozaffari S V., Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47:1091–8.

36. Mokhtari R, Lachman HM. The Major Histocompatibility Complex (MHC) in Schizophrenia: A Review. J Clin Cell Immunol. 2016;7.

37. Barbeira AN, Pividori MD, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. Plagnol V, editor. PLOS Genet. 2019;15:e1007889.

38. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26:2190–1.

39. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9:1825.

40. Liang Y, Pividori M, Manichaikul A, Palmer AA, Cox NJ, Wheeler HE, et al. Polygenic transcriptome risk scores (PTRS) can improve portability of polygenic risk scores across ancestries. Genome Biol. 2022;23:1–18.

41. Clements CC, Wenger TL, Zoltowski AR, Bertollo JR, Miller JS, de Marchena AB, et al. Critical region within 22q11.2 linked to higher rate of autism spectrum disorder. Mol Autism. 2017;8:58.

42. Palmer DS, Howrigan DP, Chapman SB, Adolfsson R, Bass N, Blackwood D, et al. Exome sequencing in bipolar disorder reveals shared risk gene AKAP11 with schizophrenia. medRxiv. 2021;2021.03.09.21252930.

43. Schneider M, Debbané M, Bassett AS, Chow EWC, Fung WLA, Van Den Bree MBM, et al. Psychiatric disorders from childhood to adulthood in 22q11.2 deletion syndrome: Results from the international consortium on brain and behavior in 22q11.2 deletion syndrome. Am J Psychiatry. 2014;171:627–39.

44. McCammon JM, Blaker-Lee A, Chen X, Sive H. The 16p11.2 homologs fam57ba and doc2a generate certain brain and body phenotypes. Hum Mol Genet. 2017;26:3699–712.

45. Carmel M, Zarchi O, Michaelovsky E, Frisch A, Patya M, Green T, et al. Association of COMT and PRODH gene variants with intelligence quotient (IQ) and executive functions in 22q11.2DS subjects. J Psychiatr Res. 2014;56:28–35.

46. Ni P, Liu M, Wang D, Tian Y, Zhao L, Wei J, et al. Association Analysis Between Catechol-O-Methyltransferase Expression and Cognitive Function in Patients with Schizophrenia, Bipolar Disorder, or Major Depression. Neuropsychiatr Dis Treat. 2021;17:567–74.

47. Mizuno A, Okada Y. Biological characterization of expression quantitative trait loci (eQTLs) showing tissue-specific opposite directional effects. Eur J Hum Genet. 2019;

48. Loviglio MN, Leleu M, Männik K, Passeggeri M, Giannuzzi G, van der Werf I, et al. Chromosomal contacts connect loci associated with autism, BMI and head circumference phenotypes. Mol Psychiatry. 2016;

49. Lin Z, Xue H, Malakhov MM, Knutson KA, Pan W. Accounting for nonlinear effects of gene expression identifies additional associated genes in transcriptome-wide association studies. Hum Mol Genet. 2022;31:2462–70.

50. Lim ET, Chan Y, Burns MJ, Guo X, Erdin S, Tai DJC, et al. Identifying cell type specific driver genes in autism-associated copy number loci from cerebral organoids. bioRxiv. 2020;2020.11.15.375386.

FIGURES

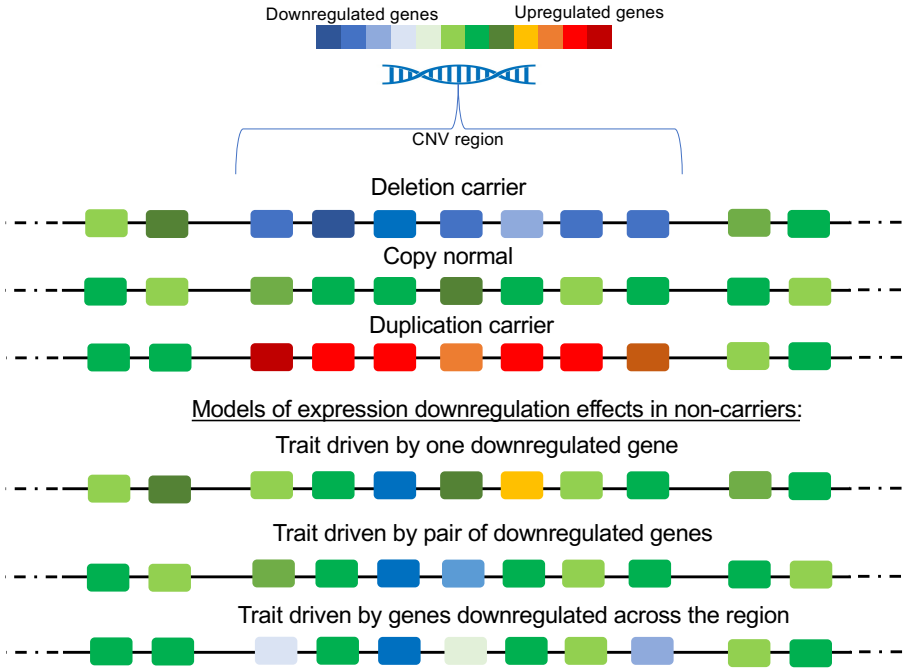


Figure 3.1: An overview of models of CNV pathogenicity due to gene expression.

Rectangles represent individual genes in a chromosomal location. Warmer colors represent increased mRNA expression. Cooler colors represent decreased mRNA expression. Greens represent population average mRNA expression.
Top: Within a CNV region, deletion carriers have reduced expression across the majority of genes, duplication carriers have increased expression across the majority of genes, and copy normal individuals have "average" levels of expression across the majority of genes. These increases and decreases are specific to the CNV region experiencing increased or decreased DNA copies (potential positional effects on flanking genes not shown).

Bottom: Three models of how gene expression downregulation in a CNV region may influence a CNV-associated trait in non-carriers. In the first model, decreased expression of a single gene is sufficient. In the second model, a trait is impacted when two specific genes both have reduced expression. In the third model, the trait becomes more likely due to reduction of expression in many genes across the region. These three models are utilized in our study.

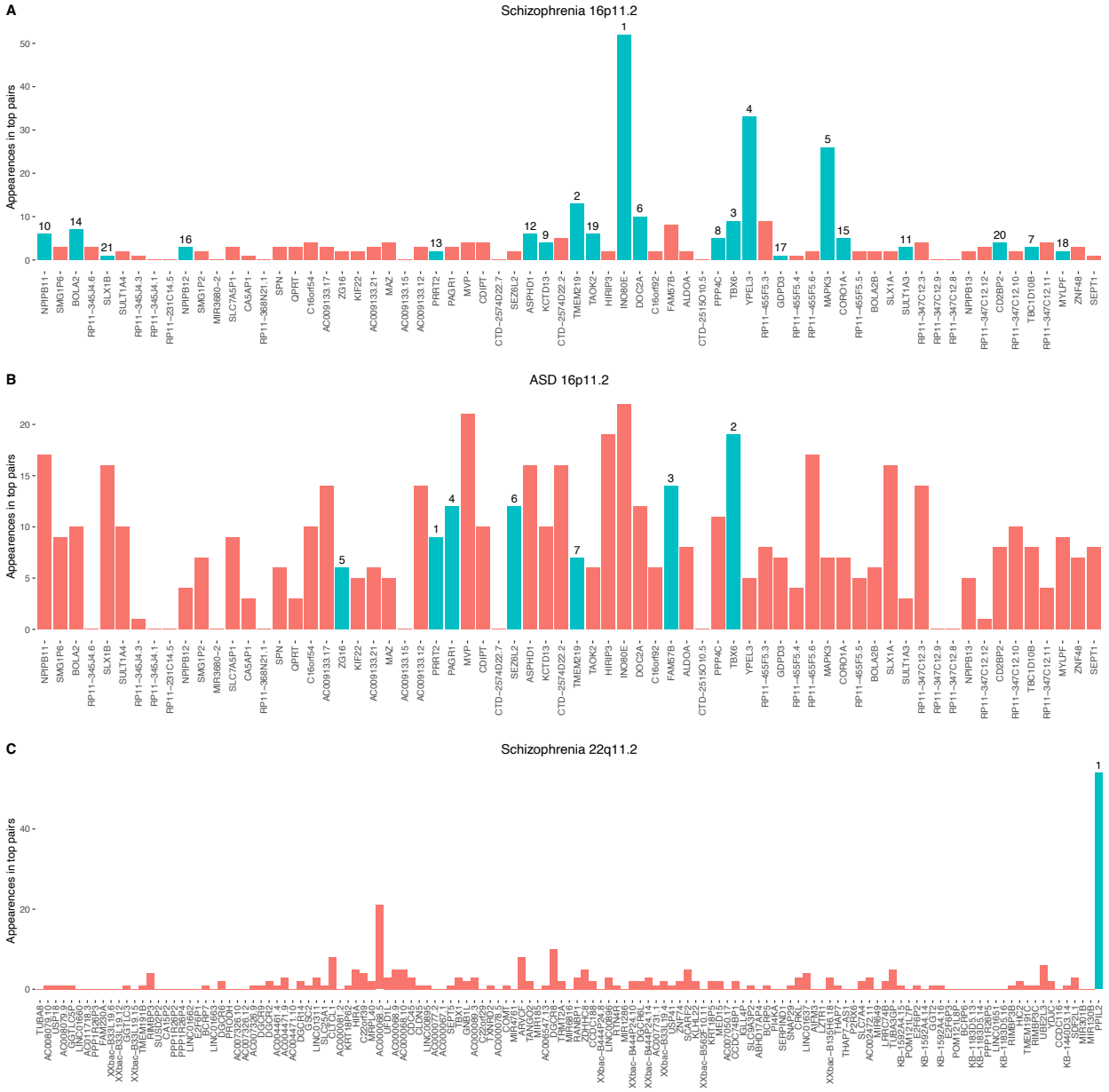


Figure 2.2 Genes most likely to appear in top pairs.

Y-axis: counts of the number of times each gene contributes to a significant pair (permutation *P*-value < median of 5th percentiles of control region p-values). Bars in teal represent genes

significant (permutation *P*-value < median of 5th percentiles of control region p-values) in a single gene model for the same trait, with rank indicated above the bar. Bars in salmon represent genes not significant in a single gene model. X-axis: genes in chromosomal order.
Top: For schizophrenia at 16p11.2, the three disproportionately represented genes are also significant in a single gene model. As the number of pairs associated with schizophrenia at 16p11.2 is large, only the top 10% of schizophrenia pairs are plotted here.
Middle: For ASD at 16p11.2, genes significant in a single gene model are not disproportionately represented in significant pairs, with no disproportionate outliers evident.
Bottom: For schizophrenia at 22q11.2, one gene that was also individually significant, *PPIL2,* appears in a large fraction of significant pairs. The second-most represented gene, AC000068.5, was not among significant single genes.

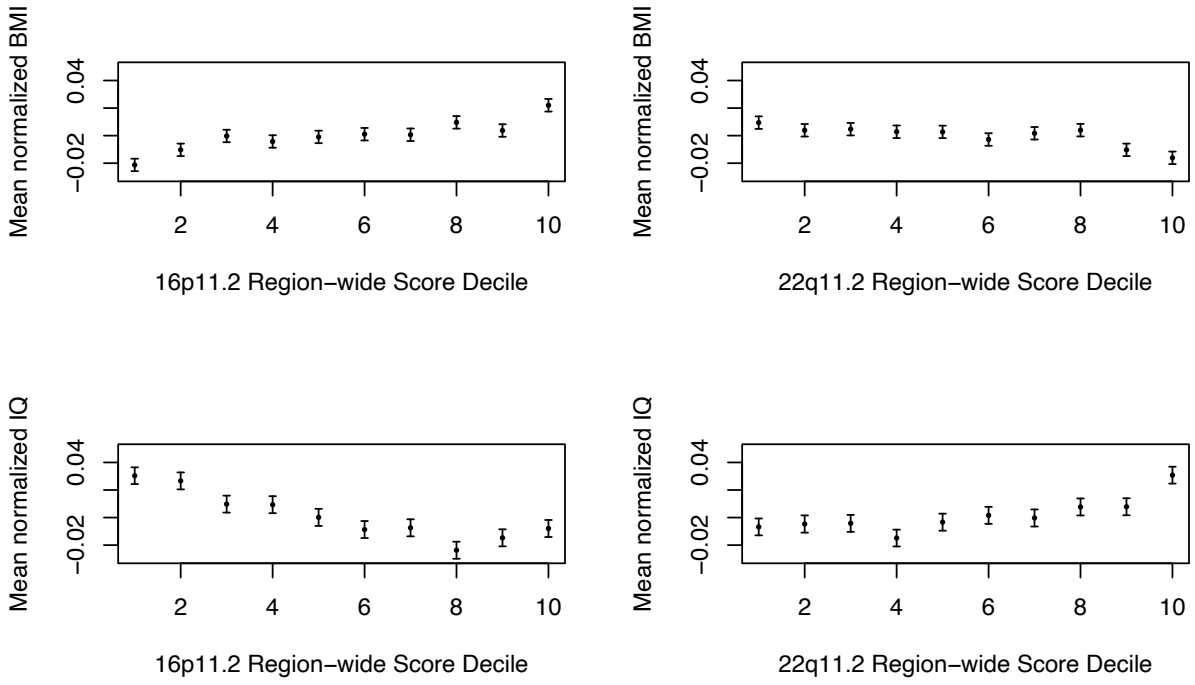


Figure 3.3 IQ and BMI values are associated with region-wide score.

Region-wide scores across individuals were binned into deciles and the mean (dot) and standard error (bars) of BMI and IQ values for each decile are plotted.

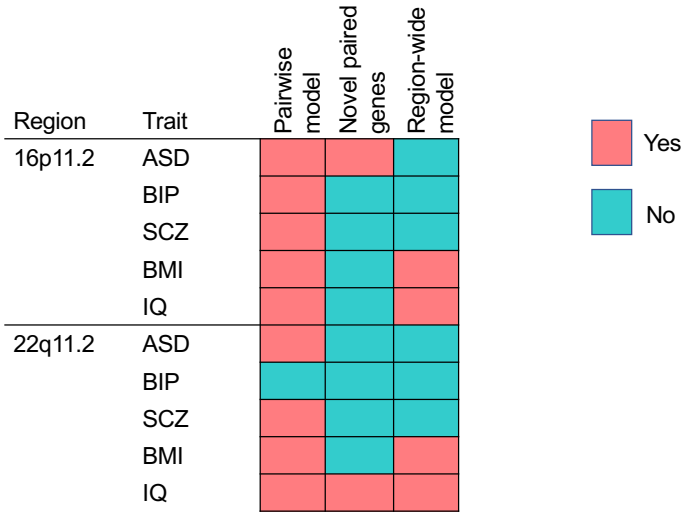


Figure 4.4: Insights gained into CNV-trait pairs

For each CNV-trait pair, we specify whether pairwise models performed better than single gene models (left column), whether genes represented disproportionately in significant pairs primarily represented genes significant in a single gene model (middle column), and whether region-wide association with a trait was significant. Yes: salmon, No: teal.

TABLES

Table 3.1: Proportion of significantly associated (permutation $P <$ median of 5th percentiles of control region p-values) single genes (singles) and pairwise gene sums (pairs) for each trait and CNV.

	16p11.2		22q11.2	
Trait	**N single (%)**	**N pairs (%)**	**N single (%)**	**N pairs (%)**
ASD	8/42 (19%)	273/1542 (18%)	6/65 (9%)	282/3654 (8%)
Bipolar	5/37 (14%)	142/1536 (9%)	3/59 (5%)	137/3669 (4%)
Schizophrenia	21/37 (57%)	702/1543 (45%)	1/59 (2%)	129/4267 (3%)
BMI	31/38 (82%)	1212/1554 (78%)	5/52 (10%)	176/3229 (5%)
IQ	5/38 (13%)	74/1545 (5%)	2/65 (3%)	33/4052 (1%)

Table 3.2: Counts of the model estimated to explain most trait variance for each tissue-cohort pair. *P*-value represents a chi-square test comparing pairwise to non-pairwise counts.

		CNV Region	**All Control Regions**	
Region	Trait	**single/interaction/pairwise (% pairwise)**	**single/interaction/pairwise (% pairwise)**	**P-value**
16p11.2	ASD	205/169/243 (39%)	5891/7588/6387 (32%)	0.00012
	Bipolar	359/390/721 (49%)	14806/19593/19631 (36%)	$< 2.2x10^{-16}$
	Schizophrenia	754/730/1554 (51%)	28288/39477/51938 (43%)	$< 2.2x10^{-16}$
	BMI	0/0/49 (100%)	48/159/1744 (89%)	0.016
	IQ	0/0/49 (100%)	98/232/1565 (83%)	0.0013
22q11.2	ASD	174/196/267 (42%)	4909/7016/5313 (31%)	$1.3x10^{-9}$
	Bipolar	536/435/499 (34%)	11642/15167/14381 (35%)	0.44
	Schizophrenia	871/816/1155 (41%)	19632/28147/35053 (42%)	0.07
	BMI	0/0/49 (100%)	17/68/1258 (94%)	0.069
	IQ	7/17/25 (51%)	15/64/605 (88%)	$2.5x10^{-16}$

Chapter four: Conclusion

In the preceding chapters, I used a novel predicted expression-based approach to fine-map the impacts of 16p11.2 and 22q11.2 on five neurobehavioral traits: ASD, schizophrenia, bipolar disorder, BMI, and IQ. In chapter two, we found that, as we hypothesized, the 16p11.2 region (at least) has genes that seem to drive the same traits in non-carriers. Although we might have hoped that we could identify individual genes for all ten CNV-trait pairs, our results point to a more complex reality. Multi-gene combinations appear to be the biggest contributors to traits in both CNVs. Much further work remains to untangle these genes further.

We saw intriguing pleiotropic effects among the phenotypes. There has been consistent phenotypic evidence of a relationship between IQ and BMI, backed up by genetic studies [1,2]. Here, at the 16p11.2 region, regardless of the reference panel used, the top associated genes between these studies were the same (*SPN* in chapter two and *SULT1A4* and *MVP* in chapter three – these genes are on the same end of the CNV and better eQTL mapping can improve the precision of the GWAS SNP-to-gene matching). At 22q11.2, we saw a large single and pairwise contribution from *PPIL2* for bipolar disorder and schizophrenia, two traits where we might expect to see genetic overlap [3]. While *PPIL2* has evidence of a rare-variant contribution to bipolar disorder, there is no such evidence in schizophrenia data [4]. However, in the specific case of 22q11.2 carriers, a combinatorial effect including this gene might contribute to both traits. Finally, we found *INO80E*, a gene that contributes to BMI, schizophrenia, and bipolar disorder (with weaker evidence for contribution to ASD), likely through pairwise effects. Although schizophrenia is the only trait where we found a large individual contribution (and even that seemed to explain less variance than the pairwise contribution), the idea of a "master

gene" driving pleiotropy within the region is intriguing. We have not identified similar genes in 22q11.2, including oft-proposed candidate genes such as *COMT, TBX1,* or *DGCR8*.

I am particularly curious about what further research will discover regarding *INO80E*, the chromatin remodeling gene at 16p11.2 that appears important for at least four of the five main phenotypes we studied (it did not appear in our IQ analyses). We are not the only ones who have implicated this gene using similar studies, yet it does not seem like there are currently any functional hypotheses [5,6]. At the same time, there are several curiosities that are worth pursuing further. Our lab's attempted knockdown of this gene in iPSCs (unpublished) did not recapitulate any of the previously observed 16p11.2 knockdown cellular phenotypes. In Drosophila, *INO80E* is one of the few genes that was not available for study [7]. In zebrafish, *INO80E* appears to be the only one of the CNV genes that is not located on the same chromosome as any of the others [8]. In large exome sequencing studies of ASD, bipolar disorder, and schizophrenia, no over-representation of rare variants in *INO80E* has been found.

Another hypothesis derived from both studies is the potential that the 16p11.2 region has both BMI-increasing and BMI-decreasing genes. The contribution of 16p11.2 to obesity is of note as it's both bi-directional (deletions associated with obesity and duplications associated with low weight) and has the opposite effect in mice (deletions with low weight and duplications with obesity) [9,10]. Our single-gene analysis pointed to genes with expression associated with BMI in the expected negative direction (such as *INO80E*) and the unexpected positive direction (such as *KCTD13*). Then, in the region-wide analysis, we found that higher region-wide expression scores are associated with increased BMI. While it is possible that our observations are a specific

product of our study design – focusing across all tissues rather than the most relevant – our hypothesis of both increasing and decreasing genes is worth considering when trying to learn why mice and humans have different effects from deletion of the same set of genes.

One follow-up question we may ask is: how generalizable are our results to other traits and CNVs? The five traits in this study were chosen (apart from practical factors such as data availability) because they frequently appear in CNV carriers and they have not been fine-mapped to any consistent small subset of genes. However, the BioVU carrier screen picked up a wide array of phenotypes, especially at 22q11.2, including some that have not been reported before. While heart and kidney-related traits at this CNV may have known drivers, traits such as immune dysfunction and hearing loss can benefit from fine-mapping approaches such as ours. Similarly, could CNVs other than 16p11.2 and 22q11.2 be analyzed in this way? We found evidence in chapter three that 16p11.2 and 22q11.2 are different from matched controls in their patterns of trait variance explained. What we do not yet know is whether this is a phenomenon of 16p11.2 and 22q11.2 or one of CNVs in general. Additional variants with unclear genetic architecture, such as 1q21.1, would benefit from similar analyses.

Our studies have barely scratched the surface of the overall impact of CNVs on neurobehavior. One of the main reasons is that the focus has been on genes at or near the CNV regions. This is a reasonable search space, yet it may be an incomplete picture of CNV gene activity. We have genes in these regions that are have important effects that may not be localized to the CNV region, such as T-box transcription factors, as well as a major miRNA processing gene. One of the projects that I started was looking at downstream impacts of CNV gene expression,

analogous to trans-eQTL mapping. My preliminary analyses showed that CNV genes seemed to have more downstream genes compared to controls. However, I did not end up performing association testing with these networks, so I would still be curious to know whether or not the observation of larger networks would be helpful in predicting phenotype. It is additionally becoming appreciated that CNV mutations may have additional modifiers elsewhere in the genome [11,12]. Ultimately, the search space of CNV-relevant loci may have to expand well beyond the small 0.5-3MB regions. Even within our data, there is a large combinatorial space between pairs and all genes in the region that can be explored. Theoretically, this can be done *in silico*, although it would be a very intensive process. For example, one can utilize a backwards stepwise approach, starting with all of the genes and removing/adding back genes one by one to find a final best combination of genes with the most variance explained.

While this study leaves off far from translatability, the idea of using our results to inform treatments for CNV carriers is motivating. If there was a single-gene cause for neurobehavior, either one per trait or a cross-trait "master gene" as described earlier, we could imagine how a therapeutic modification of gene expression – increase or decrease for deletions and duplication carriers respectively – could lead to a reduction in symptoms. However, the finding that traits are driven by pairs – if not more complex architectures – means that drug treatment of CNV carriers for neurobehavioral traits may require something more akin to combination therapies. Moreover, if the reality is broader, such as in cross-genome contributions to traits, we may have to concede that precision therapeutics may not be feasible for 16p11.2 and 22q11.2 related conditions. However, psychiatric drug development in general will benefit from a better understanding of important target genes throughout the genome.

Although copy number variants are well-appreciated as causes of neurobehavioral disorders, with an ever-growing phenotypic spectrum, the more subtle genetic mechanisms for their activity remain unknown. In my opinion, the contribution to the field of my thesis is our discovery that the 'one -gene-one-trait' model does not hold. To the best of my knowledge, these sorts of combinatorial and systematic studies have only been done in animal models thus far. We have shown that large cohorts of non-CNV-carriers are a useful way to understand the biology of CNV carriers. We have also shown ways to creatively use expression imputation – while this is a method used primarily for genome-wide screens, it is equally useful for smaller subsets of adjacent genes and even for pairwise gene effects. Finally, I would like to emphasize the utility of data integration: genetics, transcriptomics, and EHR – this sort of data integration will be vital for problems in psychiatric genetics in the coming years.

References

1. Marioni RE, Yang J, Dykiert D, Mõttus R, Campbell A, Davies G, et al. Assessing the genetic overlap between BMI and cognitive function. Mol Psychiatry. 2016;21:1477–82.

2. Tabriz AA, Sohrabi MR, Parsay S, Abadi A, Kiapour N, Aliyari M, et al. Relation of intelligence quotient and body mass index in preschool children: a community-based cross-sectional study. Nutr Diabetes. 2015;5:e176.

3. Lee SH, Ripke S, Neale BM, Faraone S V, Purcell SM, Perlis RH, et al. No Title. 2013;45.

4. Palmer DS, Howrigan DP, Chapman SB, Adolfsson R, Bass N, Blackwood D, et al. Exome sequencing in bipolar disorder reveals shared risk gene AKAP11 with schizophrenia. medRxiv. 2021;2021.03.09.21252930.

5. Gusev A, Mancuso N, Won H, Kousi M, Finucane HK, Reshef Y, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat Genet. 2018;50:538–48.

6. Walker RL, Ramaswami G, Hartl C, Mancuso N, Gandal MJ, de la Torre-Ubieta L, et al. Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. Cell. 2019;179:750-771.e22.

7. Iyer J, Singh MD, Jensen M, Patel P, Pizzo L, Huber E, et al. Pervasive genetic interactions modulate neurodevelopmental defects of the autism-Associated 16p11.2 deletion in Drosophila melanogaster. Nat Commun. 2018;9:1–19.

8. Blaker-Lee A, Gupta S, McCammon JM, De Rienzo G, Sive H. Zebrafish homologs of genes within 16p11.2, a genomic region associated with brain disorders, are active during brain development, and include two deletion dosage sensor genes. Dis Model Mech. 2012;5.

9. Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z, et al. Mirror

extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature. 2011;478:97–102.

10. Pucilowska J, Vithayathil J, Tavares EJ, Kelly C, Colleen Karlo J, Landreth GE. The 16p11.2 deletion mouse model of autism exhibits altered cortical progenitor proliferation and brain cytoarchitecture linked to the ERK MAPK pathway. J Neurosci. 2015;35:3190–200.

11. Duyzend MH, Nuttle X, Coe BP, Baker C, Nickerson DA, Bernier R, et al. Maternal Modifiers and Parent-of-Origin Bias of the Autism-Associated 16p11.2 CNV. Am J Hum Genet. 2016;98:45–57.

12. Du Q, de la Morena MT, van Oers NSC. The Genetics and Epigenetics of 22q11.2 Deletion Syndrome. Front Genet. 2020;10:1365.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Mikhail Vysotskiy
B88028162B53482...
Author Signature

8/30/2022
Date