Title

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

Journal

Genome Medicine, 13(1)

ISSN

1756-994X

Authors

Vysotskiy, Mikhail
Zhong, Xue
Miller-Fleming, Tyne W
et al.

Publication Date

2021-12-01

DOI

10.1186/s13073-021-00972-1

Peer reviewed

Vysotskiy et al. Genome Medicine (2021) 13:172
https://doi.org/10.1186/s13073-021-00972-1

Genome Medicine

RESEARCH Open Access

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

Mikhail Vysotskiy[1,2,3,4†], Xue Zhong[5,6†], Tyne W. Miller-Fleming[5,6], Dan Zhou[5,6], Autism Working Group of the Psychiatric Genomics Consortium^, Bipolar Disorder Working Group of the Psychiatric Genomics Consortium^, Schizophrenia Working Group of the Psychiatric Genomics Consortium^, Nancy J. Cox[5,6] and Lauren A. Weiss[1,2,3*]

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 1500 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 over-represented 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.

* Correspondence: lauren.weiss@ucsf.edu
†Mikhail Vysotskiy and Xue Zhong contributed equally to this work.
[1]Department of Psychiatry and Behavioral Sciences, University of California San Francisco, 513 Parnassus Ave., Health Sciences East 9th floor HSE901E, San Francisco, CA 94143, USA
[2]Institute for Human Genetics, University of California San Francisco, San Francisco, CA 94143, USA
Full list of author information is available at the end of the article

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.

Keywords: Copy number variants, Transcriptome imputation, Electronic health records, Psychiatric traits, Phenome-wide association studies

Background

Multi-gene copy number variants (CNVs), including a 600-kb region at 16p11.2 and a 3-Mb 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 [2]. 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, 15, 16]. 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 [17, 18]. 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 [19]. Duplications of *Rai1* appear to explain body weight and behavior abnormalities in mouse models of 17p11.2 duplications [20]. Another example is the Williams syndrome CNV at 7q11.23 [21, 22]. 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 [23, 24]. The second method, of finding a smaller "critical region," was used to fine-map the 17q21.31 CNV [25, 26]. 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 [26]. 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* [27]. 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 [28–30]. However, it is unclear that *TBX1* is sufficient to explain brain-related disorders in 22q11.2 carriers [31, 32].

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 [33]. 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 [34]. 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 [35]. 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 [36–41], but have not successfully mapped individual genes in these CNVs to brain-related traits [28–30]. 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 [37, 38, 42]. 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 [43]. 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 [44–49]. 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 [50–53]. 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 [54–56].

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 (Fig. 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 [57, 58]. 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 [59–63]. 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 [64]. 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 = 7386 cases, 8566 controls, downloaded May 2019) [60, 61, 65]. 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) [62, 63].

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 [66]. 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

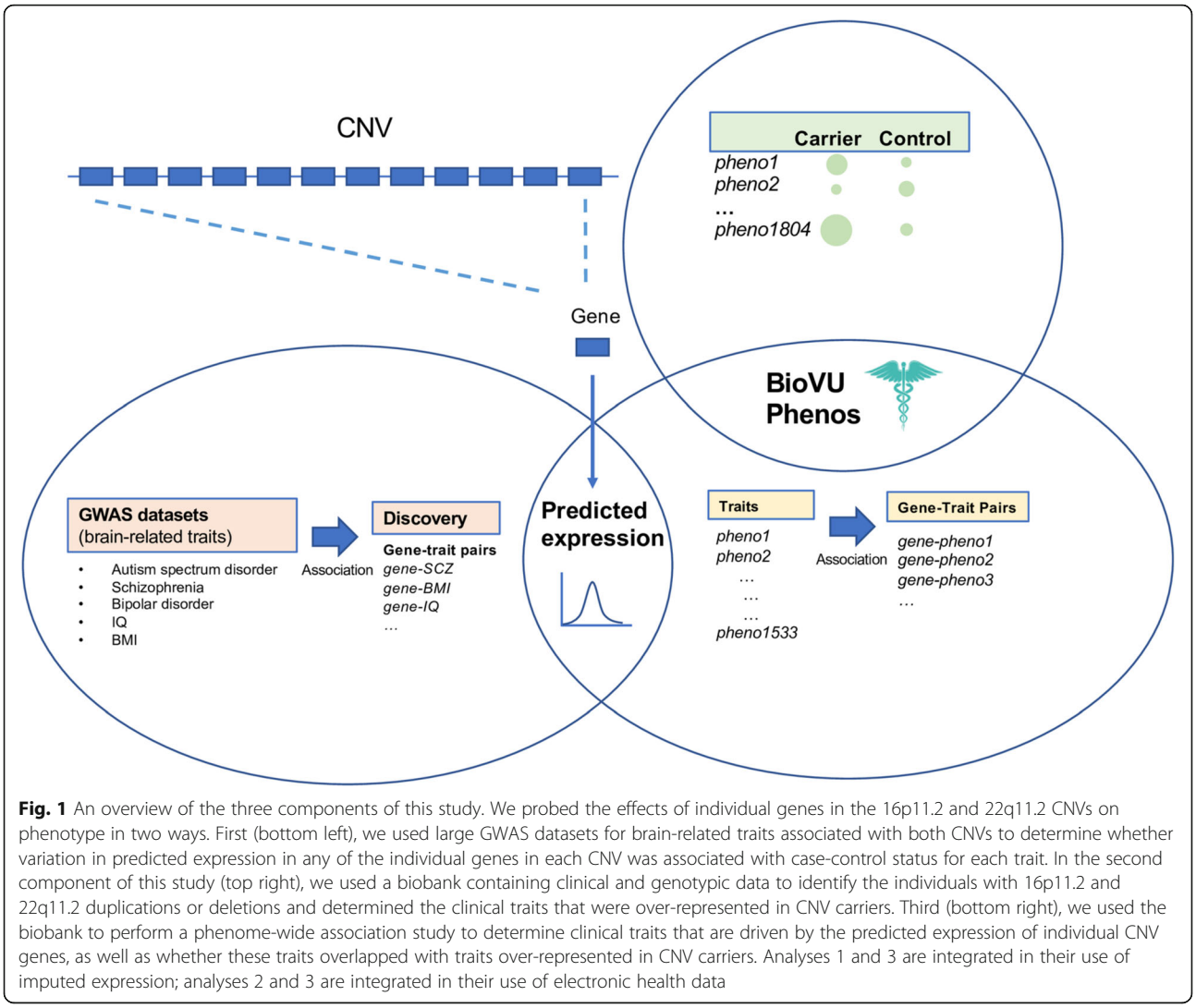

Fig. 1 An overview of the three components of this study. We probed the effects of individual genes in the 16p11.2 and 22q11.2 CNVs on phenotype in two ways. First (bottom left), we used large GWAS datasets for brain-related traits associated with both CNVs to determine whether variation in predicted 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 1 and 3 are integrated in their use of imputed expression; analyses 2 and 3 are integrated in their use of electronic health data

analysis because of previously observed high inflation of summary statistics [67, 68].

Expression prediction models

In order to impute gene expression, we obtained PrediXcan models for 48 tissues based on GTEx v7 Europeans [57, 58, 69]. 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 [57]. 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 200-kb 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 noncoding) at or near 16p11.2 lacked high-quality prediction models in every tissue. One hundred two genes (of which 90 were noncoding) 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 [70]. 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, readjusted 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 [59–61]. 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 [58]. 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 [71]. 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 multitissue 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 [72] and MR-JTI [73]) 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 multitissue 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. Linkage disequilibrium patterns in the region were visualized using LocusZoom [74].

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 [64]. 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 [64]. 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 [75] using the Michigan imputation server [76]. 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 [77, 78]. 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 (1531 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 [58]. 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 and 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 1531 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 = 1531). 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 [79]. "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 [80] 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 3 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 = 1795. 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. non-carrier 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 [77]. 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 1531 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 1795 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 [81]. The gene-trait pairs we selected for Tables 1 and 2 were used as inputs, with the − log10 p value of association used as the weighting to determine the edge width. For the

Table 1 Selected 16p11.2 gene associations with PheWAS traits

Gene	PheWAS trait	*P* value	Reason for inclusion
NPIPB11	Psychosis[a]	1.04×10^{-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]	3.03×10^{-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]	2.8×10^{-5}	PheWS
	Bariatric surgery	0.0019	Brain-related
PRRT2	Other specified nonpsychotic and/or transient mental disorders	0.0031	Brain-related
	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
	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

Table 1 Selected 16p11.2 gene associations with PheWAS traits *(Continued)*

Gene	PheWAS trait	*P* value	Reason for inclusion
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
	Dysphagia	0.0055	Del/dup
TBC1D10B	Schizophrenia and other psychotic disorders	0.0013	Drain-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

Possible reasons for inclusion are (1) del, dup, or del/dup: trait is over-represented in 16p11.2 deletion carriers, duplication carriers, or both ($P < 2.8 \times 10^{-5}$); (2) brain-related trait; (3) PheWS, phenome-wide significant

[a]Phenome-wide significant gene-trait pair ($P < 3.3 \times 10^{-5}$)

[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

22q11.2 circle plot, only associations with $P < 5 \times 10^{-3}$ were used in order to create a legible plot. Descriptions were cut off at 55 characters; to read the entire descriptions, see Tables 1 and 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 [44, 45, 82]. As flanking genes affected by copy number variation may be relevant to phenotype, we additionally considered genes 200 kb in each direction from each CNV [83]. 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 [58].

Two genes at 16p11.2 show predicted expression positively associated ($P < 7.9 \times 10^{-5}$) with schizophrenia (Fig. 2; Additional file 6: Table S5): *TMEM219* ($P = 1.5 \times 10^{-5}$) and *INO80E* ($P = 5.3 \times 10^{-10}$). 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 ($P = 4.9 \times 10^{-6}$). For IQ, there was one strong positive association at the 16p11.2 locus (Fig. 2; Additional file 6: Table S5): *SPN* ($P = 2.9 \times 10^{-22}$). 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 (Fig. 2;

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

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
	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]	1.3×10^{-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
	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	Drain-related
	Antisocial/borderline personality disorder	0.0028	Drain-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

Table 2 Selected 22q11.2 gene associations with PheWAS traits *(Continued)*

Gene	PheWAS trait	*P* value	Reason for inclusion
	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
	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
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

Table 2 Selected 22q11.2 gene associations with PheWAS traits *(Continued)*

Gene	PheWAS trait	*P* value	Reason for inclusion
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
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
	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

Table 2 Selected 22q11.2 gene associations with PheWAS traits *(Continued)*

Gene	PheWAS trait	*P* value	Reason for inclusion
	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

Possible reasons for inclusion are (1) del, dup, or del/dup: trait is over-represented in 16p11.2 deletion carriers, duplication carriers, or both ($P < 2.8 \times 10^{-5}$); (2) brain-related trait; (3) PheWS, phenome-wide significant

[a]Phenome-wide significant gene-trait pair ($P < 3.3 \times 10^{-5}$)

[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

Additional file 6: Table S5): *SPN* ($P = 6.2 \times 10^{-18}$), *TMEM219* ($P = 2.2 \times 10^{-5}$), *TAOK2* ($P = 8.5 \times 10^{-11}$), and *INO80E* ($P = 1.0 \times 10^{-7}$). We focused on genes with negative associations with BMI because, in humans, obesity is associated with deletions at 16p11.2 [1, 16]. Two additional genes, *KCTD13* ($P = 9.5 \times 10^{-6}$) and *MVP* ($P = 2.1 \times 10^{-5}$), 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) [66]. 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* ($P = 5.1 \times 10^{-11}$), and *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*

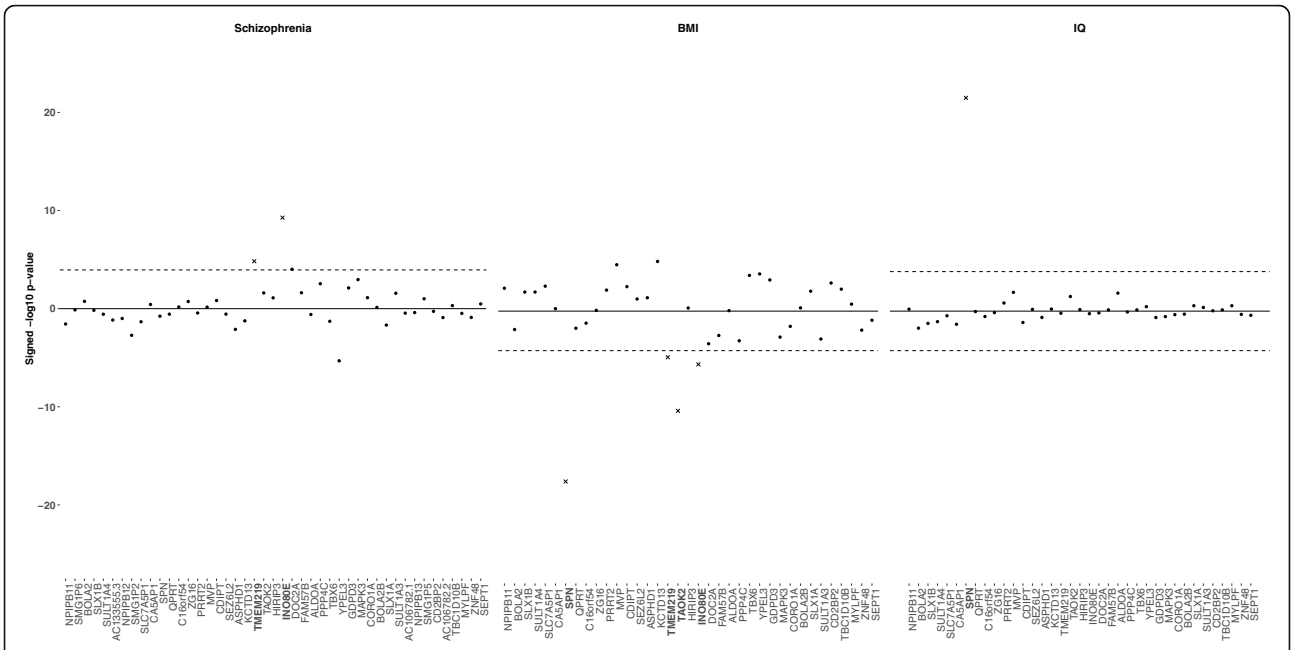

Fig. 2 Association between 16p11.2 genes and three brain-related traits. Association between predicted expression of 16p11.2 genes and schizophrenia (left), BMI (middle), IQ (right) using MultiXcan (schizophrenia) and S-MultiXcan (BMI, IQ). Genes are listed on the horizontal access in order of chromosomal 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

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 ($P = 2.3 \times 10^{-6}$). 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 [68]. 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* ($P_{UKBB} = 6.1 \times 10^{-23}$, $P_{cond} = 7.5 \times 10^{-21}$), *INO80E* ($P_{UKBB} = 1.9 \times 10^{-27}$, $P_{cond} = 2.8 \times 10^{-32}$), and *KCTD13* ($P_{UKBB} = 1.2 \times 10^{-30}$, $P_{cond} = 4 \times 10^{-27}$) were independently associated with BMI, while *TMEM219* ($P_{UKBB} = 7 \times 10^{-37}$, $P_{cond} = 2.3 \times 10^{-18}$), *TAOK2* ($P_{UKBB} = 4.2 \times 10^{-29}$, $P_{cond} = 2.3 \times 10^{-19}$), and *MVP* ($P_{UKBB} = 5.1 \times 10^{-11}$, $P_{cond} = 5 \times 10^{-6}$) 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 signals, 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 [64]. 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 1795 medical phenotype codes (Figs. 3 and 4) [77, 80]. Traits that were significantly over-represented in carriers ($P < 2.8 \times 10^{-5}$) 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.

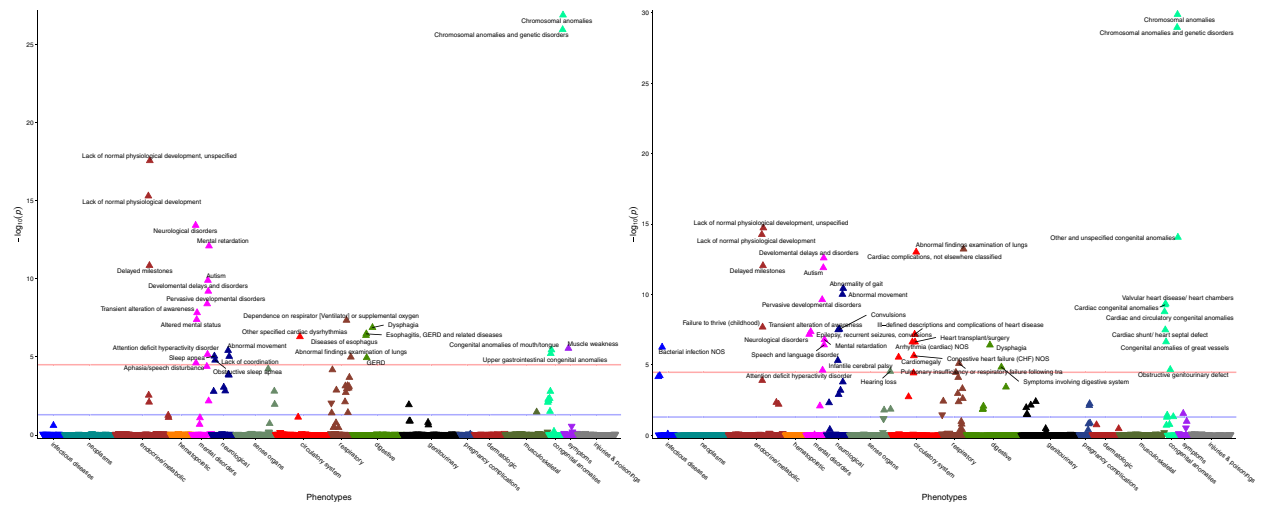

Fig. 3 Clinical traits over-represented in 16p11.2 deletion and duplication carriers. CNV carriers were identified in the EHR by keyword search and chart review (left, 16p11.2 deletions [$n = 48$], right, 16p11.2 duplications [$n = 48$], see "Methods"). Controls included all individuals without the CNV within the medical home population at Vanderbilt ($n \sim 707{,}000$). 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 ($p = 0.05/1{,}795 = 2.8 \times 10^{-5}$); 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

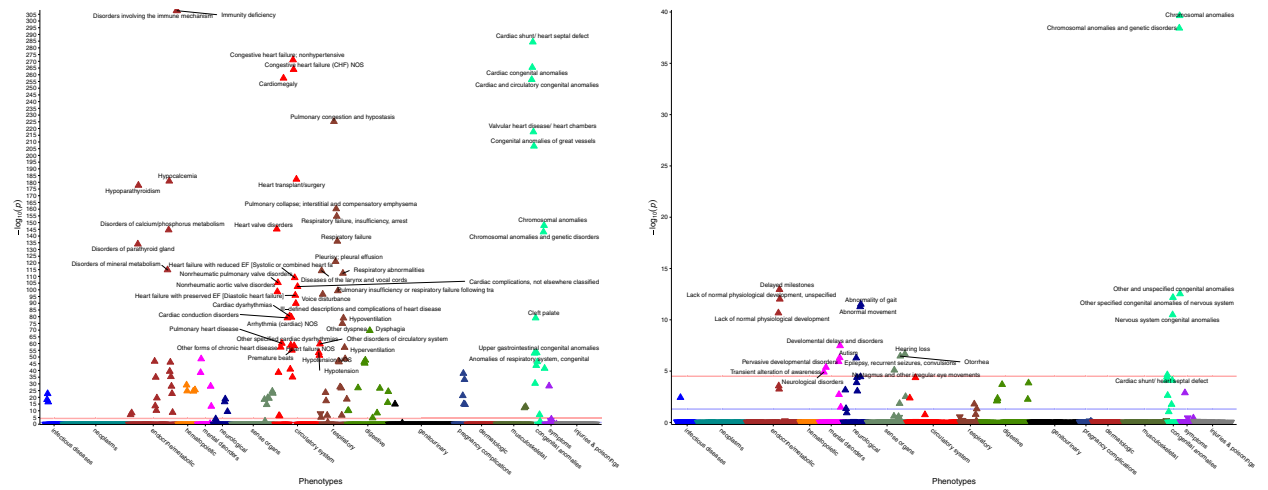

Fig. 4 Clinical traits over-represented in 22q11.2 deletion and duplication carriers. CNV carriers were identified in the EHR by keyword search and chart review (left, 22q11.2 deletions [n = 388], right, 22q11.2 duplications [n = 43], see "Methods"). Controls included all individuals without the CNV 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 ($p = 0.05/1{,}795 = 2.8 \times 10^{-5}$); 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 ($P < 1.0 \times 10^{-50}$) are labeled in the 22q11.2 deletion plot (left). Phenotypes reaching Bonferroni-corrected significance level are labeled in the 22q11.2 duplication plot (right)

16p11.2 deletion carrier status was associated with developmental diagnoses (Fig. 3): *lack of normal physiological development* ($P = 2.8 \times 10^{-18}$), *developmental delays and disorders* ($P = 6.3 \times 10^{-10}$), *delayed milestones* ($P = 1.4 \times 10^{-11}$) [3]. In addition, 16p11.2 deletion carrier status was associated with *autism* ($P = 1.3 \times 10^{-10}$) and *mental retardation* ($P = 7.9 \times 10^{-13}$) [5]. The digestive diagnosis of *GERD* ($P = 1.1 \times 10^{-5}$) has been previously observed in carriers but was unlikely to be a primary reason for genetic testing [84]. *GERD* was accompanied by other digestive diagnoses such as *dysphagia* ($P = 1.3 \times 10^{-7}$) and *diseases of esophagus* ($P = 4.3 \times 10^{-7}$). *Muscle weakness* ($P = 2.8 \times 10^{-6}$) and *abnormal movements* ($P = 3.9 \times 10^{-6}$) are consistent with neurological traits reported in 16p11.2 deletion carriers such as hypotonia and motor impairments [85]. *Sleep apnea* ($P = 8.9 \times 10^{-6}$) was a novel phenotype, potentially related to increased BMI in deletion carriers.

16p11.2 duplication carrier status was similarly associated with developmental diagnoses (Fig. 3): *lack of normal physiological development* ($P = 5.6 \times 10^{-15}$), *developmental delays and disorders* ($P = 2.5 \times 10^{-13}$), *delayed milestones* ($P = 9.0 \times 10^{-13}$), *autism* ($P = 1.3 \times 10^{-12}$), and *mental retardation* ($P = 1.6 \times 10^{-7}$) [3, 5]. 16p11.2 duplication carriers status was also associated with multiple heart defects, including *valvular heart disease/heart chambers* ($P = 4.6 \times 10^{-10}$) and *cardiac shunt/heart septal defect* ($P = 3.2 \times 10^{-8}$), both of which have been reported previously [86]. 16p11.2 duplications are known to be a risk factor for epilepsy and were associated with an epilepsy-related diagnosis of *convulsions* ($P = 2.9 \times 10^{-8}$) in the biobank [3, 87]. *Infantile cerebral palsy* ($P = 4.9 \times 10^{-6}$), 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* ($P = 5.5 \times 10^{-7}$) a novel finding.

For 22q11.2 deletion carriers, the canonical associated features were cardiac defects such as *cardiomegaly* ($P = 3.5 \times 10^{-258}$) and *cardiac shunt/heart septal defects* ($P = 4.7 \times 10^{-285}$) (Fig. 4) [6, 7]. Other highly associated diagnoses were developmental: *lack of normal physiological development* ($P = 1.7 \times 10^{-47}$), *developmental delays and disorders* ($P = 6.3 \times 10^{-29}$), *delayed milestones* ($P = 6.0 \times 10^{-11}$) [6, 7]. Congenital anomalies such as *cleft palate* ($P = 9.4 \times 10^{-80}$) 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* ($P = 1.4 \times 10^{-10}$).

22q11.2 duplication carrier status was also associated with developmental diagnoses (Fig. 4): *delayed milestones* ($P = 1.1 \times 10^{-13}$), *lack of normal physiological*

development ($P = 9.7 \times 10^{-13}$), *pervasive developmental disorders* ($P = 1.2 \times 10^{-6}$) [8]. 22q11.2 duplication status was associated with cardiac phenotypes such as *cardiac shunt/ heart septal defect* ($P = 2.3 \times 10^{-5}$). 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* ($P = 3.1 \times 10^{-12}$) and *hearing loss* ($P = 2.1 \times 10^{-7}$) have also been seen in 22q11.2, including as indications for genetic testing [88].

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 catalog 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 1531 medical phenotypic codes meeting frequency criteria ($n = 20$ cases) in this subset. There were six phenome-wide significant ($P < 3.3 \times 10^{-5}$) gene-trait associations at 16p11.2 including the following: *INO80E* with *skull and face fracture and other intercranial injury* ($P = 1.9 \times 10^{-15}$), *NPIPB11* with *psychosis* ($P = 1.0 \times 10^{-5}$), and *SLX1B* with *psychosis* ($P = 3.0 \times 10^{-5}$). There were eleven phenome-wide significant gene-trait associations at 22q11.2 including as follows: *AIFM3* with *renal failure* ($P = 2.3 \times 10^{-5}$), *LZTR1* with *malignant neoplasm, other* ($P = 1.4 \times 10^{-5}$), *SCARF2* with *mood disorders* ($P = 1.3 \times 10^{-5}$), *PI4KA* with *disorders of iris and ciliary body* ($P = 1.1 \times 10^{-7}$), and *disorders resulting from impaired renal function* ($P = 2.2 \times 10^{-5}$). 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 [89–91].

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 [28–31]. *TBX6* at 16p11.2, which has a role in bone development and scoliosis, has *pathologic fracture of vertebrae* in its top 1% phenome associations ($P = 0.0028$) [92–94]. *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 [95, 96]. We identified additional genes at 16p11.2 and 22q11.2 that are associated with Mendelian traits, using OMIM [79], 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 1, Fig. 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, Fig. 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 7: Table S6): *SEPT1* (*psychosis*—in UK Biobank schizophrenia 20002_1289 $P = 0.03$), *AIFM3* (*mood disorders*—in UK Biobank bipolar

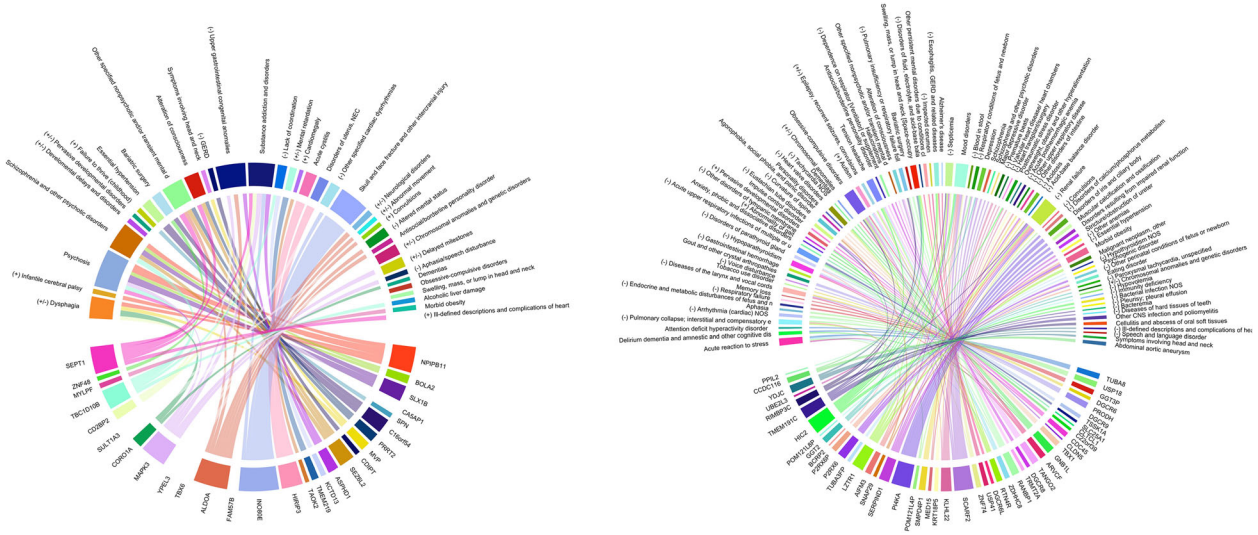

Fig. 5 Graphical summary of selected PheWAS results by gene. Each circle contains the CNV genes, in chromosomal order, on the bottom, and their associated PheWAS traits at the top. Genes are connected to their PheWAS-associated traits, with the width of the line proportional to the −log10 *p* value of 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 S9 in Additional file 10

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 3.7×10^{-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 8: Table S7). 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*, and *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 1, Fig. 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, Fig. 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 (2nd 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 over-represented in 16p11.2 and 22q11.2 carriers for comparison with individual gene results. Third, we used the same EHR system biobank containing over 1500 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 [97]. Mice heterozygous for this gene have shown abnormal locomotor activation [98]. Locomotor activity in mice is a frequently used proxy for brain-related disorders including schizophrenia [99]. Our results are consistent with a previous observation that eQTLs from dorsolateral prefrontal cortex for *INO80E* co-localize with schizophrenia GWAS SNPs [100]. In addition, an analogous imputed expression-based transcriptome-wide association study observed association between *INO80E* and schizophrenia using summary statistics [101]. A third transcriptomic association study using prenatal and adult brain tissues also pointed to *INO80E* as a risk gene for schizophrenia [102]. 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 [98].

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 [103, 104]. Recently, a large genome-wide analysis of rare CNVs fine-mapped *SPN* duplications as a driver of several phenotypic categories including *behavioral abnormality* [105]. 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 [106–109]. 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 [50]. 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 2000 genes [110]. 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 [43]. One set of genes may be the more influential determinant of the obesity trait in each organism.

Our PheWAS of traits over-represented 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 1500 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 [111]. *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) [89]. Model organisms with defects in *PI4KA*, associated with *disorders of iris and ciliary body* in our study, showed eye-related phenotypes [90, 91]. 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 [28–30, 92–94]. 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 [112]. 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 [43, 45]. *SCARF2* has recently been proposed as a driver of schizophrenia within a fine-mapping study within CNV carriers [105]. 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 1 MB of the gene [57]. 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 [113, 114]. 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 [115, 116]. 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 [58]. 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 [51]. 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 [117]. The GTEx correlation of eQTL effect sizes between brain and non-brain tissues is 0.499 (Spearman) [51]. 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 [39]. 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 [118, 119].

Because the CNV carrier individuals in our biobank are young (median age < 18), we do not 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 did not meet our frequency threshold, including the following: *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 helps 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 grows, 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.

Abbreviations

CNV: Copy number variant; GWAS: Genome-wide association study; PheWAS: Phenome-wide association study; PheWS: Phenome-wide significant; ASD: Autism spectrum disorder; BMI: Body mass index; IQ: Intelligence quotient; eQTL: Expression quantitative trait locus; EHR: Electronic health record; SD: Synthetic derivative

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-021-00972-1.

Additional file 1: Table S1. List of genotyped discovery and replication cohorts used in the study. List of datasets used for discovery and replication of association results with sample sizes. The specific cohorts from the Psychiatric Genomics Consortium that were used for this analysis are listed. All variables from the UK Biobank that were used for replication are shown.

Additional file 2: Table S2. List of genes at or near 16p11.2 and 22q11.2. List of coding and non-coding genes in the CNV region, as well as flanking genes 200 kb on either side. Genes for which PrediXcan models based on GTEx v7 were available and the range of model qualities (R^2) are noted, along with the number of tissues in which prediction models were available. Genes are annotated with their type (e.g., protein-coding, pseudogene, etc.), whether they are in the CNV or flanking, and any other names by which they may be referred in the literature.

Additional file 3: Fig. S1-S6. Supplementary figures.

Additional file 4: Table S3. Mendelian phenotypes annotated to 16p11.2 and 22q11.2 genes in PheWAS results. We compare Mendelian phenotypes annotated to 16p11.2 and 22q11.2 genes (as catalogued in

OMIM) with our imputed gene expression PheWAS results. For each of the Mendelian traits, we list one or more related traits that were tested in PheWAS along with the p-value, selecting the trait(s) with the best p-value to represent. Traits that are in the top 1% of associations for individual genes are marked. This table is a proof-of-concept that our PheWAS approach can pick up known gene-phenotype associations but has not been quantified for enrichment due to the subjective nature of identifying related traits.

Additional file 5: Table S4: Identifying 16p11.2 and 22q11.2 cases from electronic health records (EHR)**.** Keyword searches across all documents within the Vanderbilt EHR were performed to identify individuals carrying 16p11.2 or 22q11.2 CNVs. Individuals with documents containing matching keywords were reviewed manually to confirm the presence of 16p11.2 or 22q11.2 CNV. Individuals were excluded from case groups 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. Confirmed case numbers are listed, with the non-genotyped counts in parentheses. Non-genotyped individuals were used for downstream phenome-wide analyses.

Additional file 6: Table S5. Results of MultiXcan and S-MultiXcan associations between CNV genes and autism, schizophrenia, bipolar disorder, BMI, and IQ. For autism, bipolar disorder, and schizophrenia, z-scores and p-values come from a METAL meta-analysis across PGC cohorts. For BMI and IQ, mean z-scores and p-values come directly from S-MultiXcan output. Genes in each CNV are sorted by chromosomal position.

Additional file 7: Table S6. Comparison of association results to independent data. For each gene-trait pair, we list the original p-value, the GWAS trait(s) that we classified as most similar to a PheWAS trait, its best p-value in an independent dataset, the number of GWAS datasets that were used for this trait, and the rank of this gene within that dataset. For UK Biobank summary statistics, we have genome-wide data; for datasets with individual-level data, only 16p11.2 and 22q11.2 genes were calculated. See Table S2 for more information on datasets used.

Additional file 8: Table S7. Conditional analysis for independence of associations. Conditional analysis was performed on the PGC schizophrenia data, the UK Biobank BMI data, as well as two BioVU clinical trait associations (16p11.2 genes and *psychosis*, 22q11.2 genes and *morbid obesity*). For each trait, we performed MultiXcan first adjusting for a specific gene, then by leaving a gene in and adjusting all the other genes associated with that trait out. The P_{cond} reported in the text is the p-value of this gene-trait pair when adjusting for all other genes considered for conditioning for this trait, unless otherwise stated.

Additional file 9: Table S8. Traits over-represented in CNV carriers. The four categories of CNV carrier – 16p11.2 duplication, 16p11.2 deletion, 22q11.2 duplication, 22q11.2 deletion – were tested separately. The results for all clinical traits tested are provided. The number of cases and controls for each trait is given, as well as whether the p-value meets either Bonferroni or FDR correction. Traits in bold were represented in over 5% of carriers.

Additional file 10: Table S9. Top PheWAS associations of 16p11.2 and 22q11.2 genes. The top 15 associated traits for each gene, regardless of p-value, are shown. These represent the top 1% of associations among all traits tested. Genes are listed in alphabetical order, with each trait's sample size and phecode noted [80].

Additional file 11: Table S10. Enrichment of clinical categories among the top PheWAS associations. The top 15 traits (codes) for each gene analyzed (n = 1470 gene-trait pairs) were divided into 17 clinical categories (observed counts column). The values in the expected counts column are calculated as 1470 * {the proportion of traits of that category tested}. For example, 159 out of 1531 codes tested were from the "circulatory system" category, so the expected counts for "circulatory system" are calculated as 1470*159/1531. The last column contains the p-value from a binomial test comparing whether the observed proportion of clinical categories is more extreme than expected.

Additional file 12. Supplemental Acknowledgements. Members of the Psychiatric Genomics Consortium who contributed to this work.

Acknowledgements

We would like to acknowledge the consortia and individuals that provided the individual data we used in this study: BioVU at Vanderbilt University, the Genotype-Tissue Expression Consortium, and the Autism, Schizophrenia, and Bipolar working groups of the Psychiatric Genomics Consortium. PGC working group authors are listed in Additional file 12. We would also like to thank Megan Sheuy for providing cleaned BMI data for BioVU participants.

Authors' contributions

Study conception: NJC, LAW. Data preparation and cleaning: MV, DZ, XZ, TWM. GWAS data analysis: MV. EHR data review and analysis: TM. PheWAS: XZ, TWM. Writing and editing: all authors. All author(s) read and approved the final manuscript.

Funding

This work was sponsored by the National Institute of Mental Health [R01 MH107467 to LAW, R01 MH113362 to NJC]; as well as the National Human Genome Research Institute [U01HG009086 to NJC]. TWM was also supported by the National Human Genome Research Institute [T32 HG008341]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The BioVU projects at Vanderbilt University Medical Center are supported by numerous sources: institutional funding, private agencies, and federal grants. These include the NIH funded Shared Instrumentation Grant S10OD017985 and S10RR025141; CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975 from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the National Institutes of Health. Genomic data are also supported by investigator-led projects that include U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962, R01HD074711; and additional funding sources listed at https://victr.vumc.org/biovu-funding/.

Availability of data and materials

Individual-level genotypes for Psychiatric Genomics Consortium cohorts can be obtained by applying at www.med.unc.edu/pgc/. Individual-level UK Biobank data can be obtained by application at https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access. Summary-level genetic datasets used here are available to freely download from GIANT BMI (https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium) and CNCR IQ (https://ctg.cncr.nl/software/summary_statistics). PrediXcan models are available to download at predictdb.org. BioVU contains protected patient health records which are available only by application through Vanderbilt University. GTEx genotypes and phenotypes are listed on dbGAP (phs000424.v7.p2). The complete results of our PrediXcan and PheWAS analyses of these datasets are available in the supplementary tables of this article.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Psychiatry and Behavioral Sciences, University of California San Francisco, 513 Parnassus Ave., Health Sciences East 9th floor HSE901E, San Francisco, CA 94143, USA. [2]Institute for Human Genetics, University of California San Francisco, San Francisco, CA 94143, USA. [3]Weill Institute for Neurosciences, University of California San Francisco, San Francisco, CA 94143, USA. [4]Pharmaceutical Sciences and Pharmacogenomics Graduate Program, University of California San Francisco, San Francisco, CA 94143, USA.

[5]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA. [6]Vanderbilt Genetics Institute, Nashville, TN 37232, USA.

Received: 12 February 2021 Accepted: 16 September 2021
Published online: 29 October 2021

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(5):332–41. https://doi.org/10.1136/jmg.2009.073015.
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(4):628–38. https://doi.org/10.1093/hmg/ddm376.
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(7):667–75. https://doi.org/10.1056/NEJMoa075974.
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(4):307–13. https://doi.org/10.1002/ajmg.a.30984.
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(10):2058–69. https://doi.org/10.1002/ajmg.a.40637.
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(2):204–8. https://doi.org/10.1038/gim.2016.98.
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(4):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(10):868–80. https://doi.org/10.1097/GIM.0b013e3182217a06.
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(2):148–61. https://doi.org/10.1016/j.ajhg.2008.12.014.
14. Bijlsma EK, Gijsbers ACJ, Schuurs-Hoeijmakers JHM, van Haeringen A, van de Putte DE F, 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. 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(1):37–40. https://doi.org/10.1038/mp.2013.156.
16. 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(7281):671–5. https://doi.org/10.1038/nature08727.
17. 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.
18. 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(1):84–7. https://doi.org/10.1038/71743.
19. Slager RE, Newton TL, Vlangos CN, Finucane B, Elsea SH. Mutations in RAI1 associated with Smith–Magenis syndrome. Nat Genet. 2003;33(4):466–8. https://doi.org/10.1038/ng1126.
20. 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(11):3035–41. https://doi.org/10.1172/JCI28953.
21. Williams JCP, Barratt-Boyes BG, Lowe JB. Supravalvular aortic stenosis. Circulation. 1961;24(6):1311–8. https://doi.org/10.1161/01.CIR.24.6.1311.
22. Beuren AJ, Apitz J, Harmjanz D. Supravalvular aortic stenosis in association with mental retardation and a certain facial appearance. Circulation. 1962;26(6):1235–40. https://doi.org/10.1161/01.CIR.26.6.1235.
23. 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(1):159–68. https://doi.org/10.1016/0092-8674(93)90168-P.
24. 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(1):11–6. https://doi.org/10.1038/ng0993-11.
25. 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.
26. Koolen DA, Sharp AJ, Hurst JA, Firth HV, Knight SJL, Goldenberg A, et al. Clinical and molecular delineation of the 17q21.31 microdeletion syndrome. J Med Genet. 2008;45:710–20.
27. Koolen DA, Kramer JM, Neveling K, Nillesen WM, Moore-Barton HL, Elmslie FV, et al. Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. Nat Genet. 2012;44(6):639–41. https://doi.org/10.1038/ng.2262.
28. Jerome LA, Papaioannou VE. DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1. Nat Genet. 2001;27(3):286–91. https://doi.org/10.1038/85845.
29. 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(6824):97–101. https://doi.org/10.1038/35065105.
30. 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(4):619–29. https://doi.org/10.1016/S0092-8674(01)00247-1.
31. 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(20):7729–34. https://doi.org/10.1073/pnas.0600206103.
32. 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;425(3):146–50. https://doi.org/10.1016/j.neulet.2007.07.055.
33. 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. https://doi.org/10.1101/2020.09.12.20192922.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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;5:834–51.
39. 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.
40. 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(23):2645–50. https://doi.org/10.1093/hmg/10.23.2645.
41. 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.
42. 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(19):3699–712. https://doi.org/10.1093/hmg/ddx255.
43. 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.
44. 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. https://doi.org/10.1101/2020.02.09.940965.
45. 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(6):870–83. https://doi.org/10.1016/j.ajhg.2014.05.004.
46. 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(1):38–55. https://doi.org/10.1016/j.ajhg.2012.05.011.
47. 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(1):5356. https://doi.org/10.1038/s41467-018-07766-x.
48. 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.
49. 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(5):784–96. https://doi.org/10.1016/j.ajhg.2015.04.002.
50. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
51. 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(7675):204–13. https://doi.org/10.1038/nature24277.
52. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–5.
53. Ardlie KG, DeLuca DS, Segrè AV, 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.
54. 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. Am J Hum Genet. 2018;103:535–52.
55. Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA. Clan genomics and the complex architecture of human disease. Cell. 2011;147:32–43.
56. 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(1):70–80. https://doi.org/10.1016/j.cell.2013.08.030.
57. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091–8. https://doi.org/10.1038/ng.3367.
58. 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.
59. Schizophrenia Working Group of the Psychiatric Genomics Consortium S. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511(7510):421–7. https://doi.org/10.1038/nature13595.
60. 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(5):793–803. https://doi.org/10.1038/s41588-019-0397-8.
61. 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(3):431–44. https://doi.org/10.1038/s41588-019-0344-8.
62. 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(7538):197–206. https://doi.org/10.1038/nature14177.
63. 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(7):912–9. https://doi.org/10.1038/s41588-018-0152-6.
64. 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(3):362–9. https://doi.org/10.1038/clpt.2008.89.
65. Schizophrenia Working Group of the Psychiatric Genomics Consortium {fname}. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421–7.
66. UK Biobank — Neale lab. [cited 2020 Mar 28]. Available from: http://www.nealelab.is/uk-biobank
67. 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(20):3641–9. https://doi.org/10.1093/hmg/ddy271.
68. 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.
69. 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(1):1825. https://doi.org/10.1038/s41467-018-03621-1.
70. 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(7675):204–13. https://doi.org/10.1038/nature24277.
71. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190–1. https://doi.org/10.1093/bioinformatics/btq340.
72. 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(4):675–82. https://doi.org/10.1038/s41588-019-0367-1.
73. 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(11):1239–46. https://doi.org/10.1038/s41588-020-0706-2.
74. 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(18):2336–7. https://doi.org/10.1093/bioinformatics/btq419.
75. 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(10):1279–83. https://doi.org/10.1038/ng.3643.
76. 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(10):1284–7. https://doi.org/10.1038/ng.3656.
77. 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;26:1205–10.
78. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. Bioinformatics. 2014;30:2375–6.
79. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). 2021. World Wide Web URL: https://www.omim.org/help/faq#1_814.
80. 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(12):1102–11. https://doi.org/10.1038/nbt.2749.
81. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics. 2014;30(19):2811–2. https://doi.org/10.1093/bioinformatics/btu393.
82. 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.
83. 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(2):332–41. https://doi.org/10.1086/506371.
84. 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. https://doi.org/10.1101/2020.04.22.055731.
85. 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.
86. 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.
87. 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.
88. 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.
89. 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(2):182–7. https://doi.org/10.1038/ng.2855.
90. 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(23):4303–10. https://doi.org/10.1242/jcs.057646.
91. 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(9):6120–32. https://doi.org/10.1074/jbc.M113.531426.
92. 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(35):57430–41. https://doi.org/10.18632/oncotarget.10619.
93. 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(7):1548–58. https://doi.org/10.1038/s41436-018-0377-x.
94. Watabe-Rudolph M, Schlautmann N, Papaioannou VE, Gossler A. The mouse rib-vertebrae mutation is a hypomorphic Tbx6 allele. Mech Dev. 2002;119(2):251–6. https://doi.org/10.1016/S0925-4773(02)00394-5.
95. 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(3):601–7. https://doi.org/10.1038/s41436-018-0137-y.
96. 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(2):347–57. https://doi.org/10.1016/j.ajhg.2015.12.008.
97. 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(7701):391–5. https://doi.org/10.1038/s41586-018-0021-6.
98. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE. Mouse Genome Database (MGD) 2019. Nucleic Acids Res. 2018;47(D1):D801–6. https://doi.org/10.1093/nar/gky1056.
99. Powell CM, Miyakawa T. Schizophrenia-relevant behavioral testing in rodent models: a uniquely human disorder? Biol Psychiatry. 2006;59(12):1198–207. https://doi.org/10.1016/j.biopsych.2006.05.008.
100. 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(6):1169–84. https://doi.org/10.1016/j.ajhg.2018.04.011.
101. 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(4):538–48. https://doi.org/10.1038/s41588-018-0092-1.
102. 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.
103. 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(4):1328–32. https://doi.org/10.1073/pnas.86.4.1328.
104. 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(6320):706–9. https://doi.org/10.1038/350706a0.
105. 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. https://doi.org/10.1101/2021.01.26.21250098.
106. 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(10):1477–82. https://doi.org/10.1038/mp.2015.205.
107. 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(2):601–7. https://doi.org/10.3945/ajcn.2008.26482.
108. 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(10):1745–8. https://doi.org/10.1017/S0033291718000685.
109. 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(11):1236–41. https://doi.org/10.1038/ng.3406.
110. Mizuno A, Okada Y. Biological characterization of expression quantitative trait loci (eQTLs) showing tissue-specific opposite directional effects. Eur J Hum Genet. 2019;27(11):1745–56. https://doi.org/10.1038/s41431-019-0468-4.
111. 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;376:742–54.
112. 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(4):341–50. https://doi.org/10.1056/NEJMoa1406829.
113. 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;22;836–49.
114. Bachmann-Gagescu R, Mefford HC, Cowan C, Glew GM, Hing AV, 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.
115. 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.
116. 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.
117. Consortium TGte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318–30. https://doi.org/10.1126/science.aaz1776.
118. 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(18):3561–81. https://doi.org/10.1523/JNEUROSCI.1983-18.2019.
119. Jensen M, Girirajan S. An interaction-based model for neuropsychiatric features of copy-number variants. Jensen M, Girirajan S. An interaction-based model for neuropsychiatric features of copy-number variants. PLoS Genet. 2019;15(1):e1007879.

Publisher's Note