Combinatorial patterns of gene expression changes contribute to variable expressivity of the developmental delay-associated 16p12.1 deletion

Matthew Jensen,[1,2,7] Anastasia Tyryshkina,[1,3,7] Lucilla Pizzo,[1] Corrine Smolen,[1,2] Maitreya Das,[1] Emily Huber,[1] Arjun Krishnan,[4,5] and Santhosh Girirajan[1,2,3,6]*

1. Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA
2. Bioinformatics and Genomics Program, Huck Institute of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA
3. Neuroscience Program, Huck Institute of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA
4. Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI 48824, USA
5. Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA
6. Department of Anthropology, Pennsylvania State University, University Park, PA 16802, USA
7. These authors contributed equally to this work

*** Correspondence to**:
Santhosh Girirajan
205A Life Sciences Building
Pennsylvania State University
University Park, PA 16802
E-mail: sxg47@psu.edu
Twitter: @girirajan16

ABSTRACT

Recent studies have suggested that individual variants do not sufficiently explain the variable expressivity of phenotypes observed in complex disorders. For example, the 16p12.1 deletion is associated with developmental delay and neuropsychiatric features in affected individuals, but is inherited in >90% of cases from a mildly-affected parent. While children with the deletion are more likely to carry additional "second-hit" variants than their parents, the mechanisms for how these variants contribute to phenotypic variability are unknown. We performed detailed clinical assessments, whole-genome sequencing, and RNA sequencing of lymphoblastoid cell lines for 32 individuals in five large families with multiple members carrying the 16p12.1 deletion. We found that the deletion dysregulates multiple autism and brain development genes such as *FOXP1*, *ANK3*, and *MEF2*. Carrier children also showed expression changes that were inherited as well as *de novo* compared with their parents, which matched with 39/47 observed developmental phenotypes. We identified significant enrichments for 13/25 classes of "second-hit" variants in genes with expression changes, where 7/25 variant classes were only enriched when inherited from the non-carrier parent, including missense SNVs and large deletions. In 11 instances, including for *ZEB2* and *SYNJ1*, gene expression was synergistically altered by both the deletion and inherited "second-hits" in carrier children. Finally, brain-specific interaction network analysis showed strong connectivity between genes carrying "second-hits" and genes with transcriptome alterations, including differential expression, alternative splicing, and allele-specific expression. Our study shows that family-based assessments of transcriptome data are highly relevant towards understanding the genetic mechanisms associated with complex disorders.

INTRODUCTION

Complex genetic disorders, such as autism, intellectual disability/developmental delay (ID), epilepsy, and schizophrenia, have been attributed to rare copy-number variants (CNVs), or deletions and duplications encompassing multiple genes, as well as single nucleotide variants (SNVs)[1,3,5,7]. Despite advances in high-throughput sequencing methods and quantitative assessments of large cohorts, individual variants implicated for these disorders do not sufficiently explain the variable expressivity of clinical features often observed in affected individuals[9,11,13]. An example is the 520-kbp deletion at chromosome 16p12.1 [MIM: 136570], which was originally described in children with developmental delay[14,15] but was subsequently found to confer increased risk for schizophrenia[16], epilepsy[17], and cognitive defects in control populations[18]. Unlike syndromic CNVs such as Smith-Magenis syndrome [MIM: 182290] that primarily occur *de novo*[19], the 16p12.1 deletion is inherited in over 90% of affected children from carrier parents who manifest subclinical or mild cognitive and neuropsychiatric features[2,14]. In fact, we recently found that children with the deletion were more likely to carry an additional burden of rare CNVs[15] and deleterious variants in genes intolerant to variation[2], or "second-hit" variants, compared to their carrier parents, making the deletion an ideal model for assessing the combined effects of multiple variants towards variable clinical outcomes.

While dissecting the pathogenicity of complex disorders has been challenging, cohort and family-based studies that integrate multiple variants with different effect sizes or functional outcomes have provided insights into how the genetic architecture contributes to changes in penetrance, severity, and complexity of phenotypes. In particular, analysis of gene expression patterns in human cells allows for dissecting the direct and indirect effects of genomic variants towards biological functions in complex disorders. *For example*, Merla and colleagues assessed gene expression in skin fibroblasts and lymphoblastoid cell lines (LCLs) from individuals with the 7q11.23 deletion, associated with Williams syndrome [MIM: 194050], and found that several genes adjacent to the deletion region were also downregulated compared to controls[20]. Similarly, expression changes due to the autism-associated 16p11.2 deletion [MIM: 611913] correlated with changes in head circumference phenotypes[21] and converged on several neurodevelopmental pathways, including synaptic function and chromatin modification[22]. Other studies have used expression data to identify pathogenic variants potentially missed by genome sequencing studies[23–25]. *For example*, Frésard and colleagues identified novel causal variants for 6/80

individuals with rare undiagnosed diseases through paired analysis of whole-blood transcriptomes and genomes[23]. Additionally, several recent studies have used family-based approaches to study the effects of rare inherited variants towards gene expression. *For example,* Pala and colleagues found that rare inherited variants in both coding and non-coding regions increased the likelihood of gene expression changes among 61 families in the bottlenecked Sardinia population, indicating the importance of such variants towards disease risk[26]. While these studies have shown the utility of assessing transcriptomic consequences of individual causal variants, they were focused on either control populations *or* relatively invariable disorders, and did not examine the simultaneous effects of multiple variants with different effect sizes towards changes in gene expression within major biological pathways.

Here, we integrated whole-genome sequencing and transcriptome data of LCLs from 32 individuals in five large 16p12.1 deletion families who manifested variable phenotypes, in order to investigate how the combined effects of the deletion and "second-hits" perturb transcriptional networks and biological functions. We found that the 16p12.1 deletion disrupts expression of genes involved in neuronal and developmental functions, such as signal transduction and cell proliferation, as well as genes preferentially expressed in the fetal and adult brain. We further identified significant contributions of several classes of rare "second-hit" coding and non-coding variants towards changes in gene expression among carrier children compared with their parents, especially when the variants were inherited from the noncarrier parent. In fact, we found 11 instances of genes in carrier children whose expression was synergistically altered by the combined effects of the 16p12.1 deletion and "second-hit" variants inherited from the noncarrier parent. Although a relatively small sample size precluded global analyses between these expression changes and developmental phenotypes, we found that specific expression changes contributed towards distinct clinical features of affected children through disruption of biological functions related to neurodevelopment. Our results suggest that the 16p12.1 deletion and "second-hit" variants jointly disrupt the developmental transcriptome through shared pathways to contribute towards developmental phenotypes, emphasizing the importance of family-based transcriptome studies for complex disorders.

MATERIALS AND METHODS

Patient recruitment and clinical phenotype analysis

We obtained clinical data and whole-blood DNA from 32 individuals in five families with the 16p12.1 deletion. Carrier children and their family members provided informed consent according to a protocol reviewed and approved by the Pennsylvania State University Institutional Review Board (IRB #STUDY00000278). Among the recruited individuals were 10 children with the deletion ("carrier children"), six sets of carrier and noncarrier parents, three sets of carrier and noncarrier grandparents, and four noncarrier siblings (**Figure S1; Table S1**). Affected children and family members were identified as carriers of the deletion through prior clinical diagnostic tests, which we confirmed using SNP microarray analysis[2]. Phenotypic information from the five families was collected through self-reported questionnaires, medical records, and interviews with family members. These data represent comprehensive medical history of affected children and their family members, including neuropsychiatric and developmental features (including cognitive, behavioral, and psychiatric diagnoses), anthropomorphic measures, abnormalities across multiple organ systems (nervous, craniofacial, musculoskeletal, cardiac, hearing/vision, digestive, and urinary systems), and family history of medical or psychiatric disorders.

DNA extraction, whole-genome sequencing, and variant identification

We identified a full range of rare deleterious variants from whole-genome sequencing and SNP microarray for each of the 32 family members in our cohort. Genomic DNA was extracted from peripheral blood using QIAamp DNA Blood Maxi extraction kit (Qiagen, Hilden, Germany) and treated with RNAse. DNA levels were then quantified using Quant-iT™ PicoGreen™ dsDNA assay methods (Thermo Fisher Scientific, Waltham, MA, USA), and sample integrity was assessed in agarose gel. After constructing Illumina TruSeq DNA PCR-free libraries (San Diego, CA, USA), whole-genome sequencing was performed on each sample by Macrogen Labs (Rockville, MD, USA) using an Illumina HiSeq X sequencer to obtain an average coverage of 34.5X. Raw sequencing data were processed for quality control using Trimmomatic[27] with leading:5, trailing:5, and slidingwindow:4:20 parameters, aligned to the human hg19 reference genome using BWA[28] v.0.7.13, and sorted and indexed using Samtools[29] v.1.9.

The GATK Best Practices pipeline v.3.8 (HaplotypeCaller) and v.4.0.11 (GenotypeGVCFs)[30] was used to identify SNVs and small indels from WGS data. In short, duplicate reads were marked and removed using PicardTools, and after calibration of base-pair quality scores, GATK HaplotypeCaller was used to identify variants in each sample. Variant calls were then pooled for joint genotyping and calibration of variant quality scores. Custom-built pipelines using Annovar[31] v.2016Feb01 applied a total of 430 annotation classes for variant function, population frequency, conservation, genomic region, and predicted pathogenicity. Variants were filtered based on the following quality metrics[32]: QUAL >50, read depth >8, allele balance between 0.25-0.75 (or >0.9 for homozygous variants), and quality depth (QUAL/reads with alternate allele) <1.5. Rare variants were defined as variants with frequency ≤0.001 in the gnomAD v.2.1.1 genome database[33], and present in <10 samples in our in-house WGS cohort of 125 families (335 individuals) with rare CNVs. We finally classified rare SNVs and small indels for downstream analysis as follows: Rare missense and loss-of-function (including frameshift and stopgain) variants within protein coding regions, as well as variants in the 5' and 3' UTR or within 1 kbp of the transcription start (TSS) or end sites (upstream and downstream), were classified based on their RefSeq-defined genomic locations in Annovar (**Figure S2**). Splice-site variants were identified based on MutationTaster annotations[34] for disease-causing ("D") or disease-causing automatic ("A") variants. Rare non-coding regulatory variants within 50 kbp of TSS for protein-coding genes were classified according to ChromHMM chromatin state segmentation data for GM12878 lymphoblastoid cells[35], available from the ENCODE Project, into promoters (chromosome states 1-3), enhancers (states 4-7), or silencers (state 12). With the exception of loss-of-function and splice-site variants, all coding and non-coding variants were filtered for CADD Phred-like pathogenicity scores[4] ≥10. Inheritance patterns of these variants were determined using in-house pipelines.

Copy-number variants and structural variants were identified using a combination of WGS data for all samples and SNP microarray data for 25/32 samples, previously described in [2]. To identify variants from microarray data, Illumina Omni 2.5 BeadChip microarray experiments were performed for each sample at either the HudsonAlpha Institute for Biotechnology (Huntsville, AL, USA; n=18), Yale Center for Genome Analysis (New Haven, CT, USA; n=5), or the Department of Genome Sciences at the University of Washington (Seattle, WA, USA; n=2). CNV calls for each sample were generated using PennCNV[36] v.1.0.3, and were filtered for

≥50 kbp in length and ≥5 target probes. CNVs and SVs were also detected from aligned WGS data using a combination of four pipelines: CNVNator[37] v.0.4.1 (bin size of 200), DELLY[38] v.0.8.2, LUMPY-sv[39] v.0.2.13 with Smoove v.0.2.5, and Manta[40] v.1.6.0. In both WGS and microarray-derived datasets, adjacent CNVs were merged if they overlapped or had a gap <20% of CNV length and <50 kbp. We then integrated the CNV and SV calls from each of the datasets as follows: For smaller CNVs and SVs <50 kbp, any duplication or deletion called by at least two of the four WGS-based callers were considered for downstream analysis, with the minimum intersected regions defining the new breakpoints. For larger CNVs and SVs >50 kbp, the union of CNVNator and PennCNV calls were considered for downstream analysis. Integrated calls were based on 50% reciprocal overlap among the callers. After integration, variants were filtered for presence in <10 individuals in our in-house WGS cohort, as determined by 50% reciprocal overlap. Finally, RefSeq gene-coding regions spanned by SVs were categorized as follows: encapsulating variants which span the entire gene, interstitial variants that are contained within a gene, and 5' and 3' UTR variants that overhang the gene on either end (**Figure S2**). Inheritance patterns of CNV and SV calls were determined if calls in the child and parent had >50% reciprocal overlap.

Short tandem repeats were called from aligned WGS data with GangSTR[41] v.2.4, using the GangSTR hg19 reference file v.13.1. The calls were filtered and analyzed using three tools from the STR analysis toolkit TRTools[42]. First, dumpSTR was used to filter for quality of calls using the following parameters: read depth >20, read depth <1000, QUAL >0.9, spanbound only (calls that are spanned by reads), and filter bad confidence intervals (filtered calls whose maximum likelihood estimates were not within the confidence interval). The reads were then merged with mergeSTR, and basic statistics were calculated using statSTR. In addition to the dumpSTR filters, we applied the following filters to our call set: >95% of samples called for the STR location, variance at location >0, and overlap of the STR location with a RefSeq defined protein-coding gene. STR expansions were defined as any call for which the deviation of the repeat length was greater than the mean length plus three standard deviations among all individuals in our cohort. Finally, STR variants were categorized according to their genomic location, including exonic, intronic, 5' or 3' UTR, upstream, and downstream (**Figure S2**). Inheritance patterns of STRs were determined by matching the number of repeats in the child to their parents.

Generation of lymphoblastoid cell lines and RNA-sequencing

Peripheral blood samples for all 32 individuals in our cohort were submitted to the Coriell Institute for Medical Research (Camden, NJ, USA) for generation of lymphoblastoid cell lines through Epstein-Barr virus transformation of B-lymphocytes (**Table S2**). After receiving the LCL samples, cells were grown at 5% CO_2 and a concentration of 1X 106 cells/mL under L-glutamine-supplemented RPMI 1640 medium (11875-119, Thermo Fisher Scientific) containing 15% fetal bovine serum (35-010-CV, Corning Life Sciences, Tewksbury, MA, USA) and Cytiva HyClone™ Penicillin Streptomycin solution (SV30010, Thermo Fisher Scientific). Total RNA was isolated from three biological replicates of P6-P7 cells per sample using TRIzol Reagent (Thermo Fisher Scientific) and PureLink RNA Mini Kit (12183018A, Thermo Fisher Scientific), and subsequently treated with DNA-free DNA Removal Kit (AM1906, Thermo Fisher Scientific). RNA integrity number scores (RIN) were assessed using Agilent Bioanalyzer 2100 (**Figure S3A**), and replicates with RIN scores >8.5 were sequenced. Paired-end 50 bp libraries for each replicate were generated using Illumina TruSeq Stranded mRNA kit, and were sequenced using Illumina NovaSeq at the Penn State College of Medicine Genome Sciences Facility (Hershey, PA, USA). Two sequencing runs of 48 replicates were performed, with the biological replicates of each sample split among the two runs to mitigate batch effects, to generate a total of 43.5 million reads/replicate.

Quantification of gene expression and coverage of disease genes

Sequenced RNA reads were filtered using Trimmomatic[27] v.0.36 to remove reads <30 bp long. Following the GTEx Consortium RNA-Seq pipeline[6], the filtered reads were aligned to the human genome (GENCODE v.19) using STAR[43] v.2.4.2a, and sorted and indexed using Samtools[29] v.1.9. Duplicates reads were marked with PicardTools v.2.9.0. We assessed the quality of the aligned reads with transcript integrity scores[44], which moderately correlated ($r=0.38$, $p=1.0\times10^{-4}$, Pearson correlation) with the RIN scores for each sample (**Figure S3A**). Isoform counts for GENCODE 19 genes were quantified using RSEM[45] v.1.2.22. A collapsed gene coordinate GTF file was generated using the GENCODE 19 gene coordinates and the GTEX collapse_annotation script. Gene-level counts and transcripts per million read (TPM) values were quantified using RNASeQC[46] v.1.1.8, using strict mode and the collapsed gene coordinates.

After filtering for transcripts where all three replicates of at least one sample showed >0.2 TPM, we obtained a total of 24,340 expressed transcripts across our cohort, representing 43.3% of all GENCODE transcripts. We further compared our set of expressed LCL genes to disease gene databases[12,47–50] and genes expressed in the adult brain from GTEx consortium RNA-Seq data[6]. We defined expressed genes in GTEx tissues if they showed >0.5 TPM in 80% of samples for a particular tissue. The expressed LCL genes covered >70% of each of these gene sets, including 83% of genes expressed in GTEx brain tissues (**Figure S4A-B**). These data are in concordance with gene expression data from GTEx, where gene expression values in LCLs and brain tissues showed an average Spearman correlation of 0.84 (**Figure S4C**). These findings indicate that our LCL data would be able to identify changes in expression patterns for most genes related to neurodevelopmental disease.

Differential expression and outlier expression analysis

We performed differential expression analysis between all 16p12.1 deletion carriers and noncarriers, as well as between parents and offspring across the five families, using edgeR[51] v.3.30.0 on gene-level counts to create generalized linear models and perform quasi-likelihood F-tests with Benjamini-Hochberg correction. For testing differences between all deletion carriers (n=19) and noncarriers (n=13), we included family as a covariate in the linear model, used default filtering for low-expressed genes, and removed genes with sex-specific differences in GTEx LCL samples as well as genes on the X and Y chromosomes (due to unequal sex ratios in deletion carriers and noncarriers). To control for expression outliers, we iteratively identified sets of differentially expressed genes, defined using an FDR<0.05 threshold (Benjamini-Hochberg correction), between deletion carriers and noncarriers after removing one sample at a time. We then took the intersection of the resulting 32 sets of differentially expressed genes, and obtained a total of 1,569 transcripts differentially expressed in the deletion carriers (**Table S3**). We also performed differential expression analysis using PQLseq v.1.2 to account for gene expression similarity due to relatedness[52]. We first generated input files from unfiltered WGS SNV data using PLINK[53] v.1.9, and used GEMMA v.0.98.3, which calculates kinship between two individuals based on genotype similarity[54], to generate a kinship matrix for our cohort. This matrix was used as input for PQLseq along with gene-level counts from RNA-Seq data, after removing the same sex-specific genes as for the edgeR analysis.

We next performed family-based analysis on 13 separate trios identified across the five families (nine carrier children compared to parents and four carrier parents compared to grandparents), which are listed in **Table S1**. For example, we separately analyzed two trios in family GL_001 (**Figure S1**). For comparison, we analyzed an additional four trios with noncarrier children compared to carrier and noncarrier parents (**Figure S5C**). For each trio, we first used an edgeR workflow without covariates to identify differentially expressed genes between the offspring and carrier parent (|logFC|>0.5, FDR<0.05, Benjamini-Hochberg correction), and separately assessed expression changes between the offspring and noncarrier parent. Genes with low expression (expressed in <25% of all replicates) and sex-specific genes were removed from edgeR analysis. We then overlapped the two sets of differentially expressed genes to classify expression changes by their inheritance patterns as follows: "*de novo*" if the gene was differentially expressed in the offspring compared with both parents; "inherited from the carrier parent" if the gene was only differentially expressed compared with the noncarrier parent; and "inherited from the noncarrier parent" if the gene was only differentially expressed compared with the carrier parent (**Figure S5A**).

To identify genes with outlier expression in our cohort, we calculated z-scores of gene expression values in each individual for 14,212 protein coding genes expressed in the LCL samples. We normalized the expression values in each person by calculating the median TPM expression across the three replicates for each gene, transformed the values using $\log_2(x + 1)$, and calculated z-scores for each log-transformed TPM compared with all samples in our cohort. As principal component analysis (PCA) showed clustering of samples by family (**Figure S6A**), we used PEER[55] v.1.0 to correct the z-scores using one PEER principal component (**Figure S6B**). After correction, we further assessed for clustering of samples and replicates, and found strong Spearman correlations among replicates derived from the same sample (**Figure S3B**). We defined outlier genes as any gene with |z-score| >2 (**Figure S6C**), in line with recent studies utilizing outlier expression values[23].

Enrichment analysis for biological function, brain expression, and disease relevance

Enrichment analysis for sets of differentially expressed genes was performed using goseq[56] v.3.12, which tests for overrepresentation of gene categories in RNA-seq data. Goseq controls for selection bias in RNA-seq datasets by modeling the distribution of transcript lengths of

differentially expressed genes. We assessed for enriched biological processes using the Gene Ontology database[57], as well as genes expressed in specific adult brain tissues from GTEx[6] and developing brain tissues from the BrainSpan Atlas[58]. We defined preferentially-expressed genes in GTEx and BrainSpan tissues as expression >2 standard deviations higher than the median expression across all tissues for that gene. We further assessed for enrichment of differentially expressed gene sets for candidate neurodevelopmental disease genes (DBD Gene Database)[47], as well as specific gene sets for autism (SFARI Gene database)[48], intellectual disability (DDD and DDG2P databases)[49,50], and schizophrenia[12]. Finally, we annotated sets of genes with altered expression for two common measures of intolerance to variation, RVIS[8] and pLI[10], and used genes considered to be intolerant to variation (RVIS <20th percentile or pLI score >0.9) for downstream analysis.

Weighted gene correlation network analysis

We performed weighted gene correlation network analysis (WGCNA) to identify modules of genes that were co-expressed among samples in our cohort[59]. We used the R package tximport[60] to import RSEM-derived gene expression counts, filtered genes for >10 counts/replicate in at least one sample, and used DESeq2[61] to generate variance-stabilized expression counts for each gene. To detect co-expression patterns specific to deletion carriers, we used ComBat[62] within sva v.3.12 to perform batch correction with family as a covariate. We detected 35 co-expression modules in our samples using WGCNA[59] v.1.69, with the following parameters: power threshold = 8, signed hybrid network, unsigned topological overlap matrix, bi-weight mid-correlation, module size = 30-30000, and merge cut height = 0.25. Two modules showed strong sex-specific gene expression and were excluded from further analysis. The average gene expression values in each module were compared between carriers and noncarriers using two-tailed t-tests, and genes in each of these modules were tested for enrichment of Gene Ontology terms using goseq.

Integration of gene expression and genomic variant data

We calculated the effect size of different classes of rare “second-hit” variants towards gene expression changes, stratified by sample type and inheritance patterns. We compared 25 classes of rare variants identified from WGS data (**Figure S2**) towards differentially expressed genes in family trios as well as outlier expression genes in all individuals. For all comparisons, we

calculated odds ratios and 95% confidence intervals for each variant class towards changes in expression using Fisher's exact and Wald tests, respectively (**Table S4**). We note that we considered each variant class independently, so that dysregulated genes with multiple types of disrupting variants were counted within multiple variant classes. For the differential expression analysis, we first assessed variants in the 13 trios with carrier offspring for genes with differential expression (**Table S1**), and then determined the effects of variants in carrier children (n=9 trios) inherited from carrier or noncarrier parents towards expression changes that were inherited from the same parent. For the outlier expression analysis, we first assessed variants for outlier expression genes in all individuals. We then stratified these data by sample type (carrier child, carrier parent, and noncarrier parent), and compared variants in carrier children that were inherited from their carrier or noncarrier parent. To identify synergistic effects between the 16p12.1 deletion and "second-hit" variants, we identified a subset of genes with outlier expression in deletion carriers that were also differentially expressed in the global comparison of carriers and noncarriers, and then identified those genes which also had "second-hit" variants inherited from the noncarrier parent.

Alternative splicing analysis

To assess alternative splicing events from RNA-sequencing data, we used DESeq2[61] to detect differential expression of isoforms. After importing isoform-level expression counts from RSEM using tximport[60], we filtered for genes with >2 counts across all samples, and performed pairwise comparisons between carrier offspring and their parents in the 13 trios listed in **Table S1**, plus the four trios with noncarrier children for comparisons. We then repeated the DESeq2 analysis for gene expression counts, and only included differentially expressed isoforms within genes that did not show overall differential expression, to specifically account for isoform changes due to alternative splicing. Similar to the family-based differential expression analysis, we assigned inheritance patterns to each alternative splicing event observed in offspring based on the pairwise comparisons to each parent. We further compared alternate isoforms identified by DESeq2 to those in GTEx LCL data[6] to identify unique isoforms in our cohort. Finally, we integrated these data with 12 classes of putative splice-site disrupting variants identified from WGS data, and calculated odds ratios as described above.

Allele-specific expression analysis

We used the phASER[63] v.1.1.1 pipeline to identify allele-specific expression events in our cohort. We first used whatshap[64] v.0.18 to perform read-backed and pedigree-informed phasing of our WGS samples, and then merged the three replicate BAM files of aligned RNA-Seq reads for each sample together using Samtools, We then used phASER, which uses phased WGS data to infer phasing of RNA-Seq samples, to phase the RNA-Seq alignments and to count the number of reads per haplotype block. We ran phASER with the parameters --mapq 255 and --baseq 10, and used the recommended blacklist to remove HLA genes. Finally, we quantified log-fold changes for allelic counts in each protein-coding gene with >10 read counts using phASER Gene AE, and identified ASE for genes with FDR>0.05 using binomial tests with Benjamini-Hochberg correction. For each identified ASE event, we examined the over-expressed haplotype for presence of a deleterious rare coding variant identified from WGS, which would potentially indicate pathogenic effects of the ASE event. Finally, we determined inheritance patterns of ASE genes based on the presence of ASE in parents of offspring.

eQTL discovery and analysis

We used QTLTools[65] v.1.2 to identify eQTLs in our cohort. Because we had three replicates per participant, we first calculated the median TPM values for all transcripts in an individual. Genes were filtered for >0.1 median TPM in more than 50% of samples. Principal components for gene expression (from RNA-Seq data) and genotype (from whatshap-phased WGS data) were then computed using QTLtools. The top three genotype and the top two gene expression principal components were used as covariates for the linear model, in addition to three explicit covariates (family, sex, and carrier status). QTLtools cis-permutation tests (n=1000 replicates) were then used to discover eGenes and associated variants in our samples. We performed multiple testing correction with the QTLtools script runFDR_cis.R. Finally, we annotated significant eQTL variants (FDR<0.05) associated with protein-coding genes for presence in GTEx LCL data, genomic location, population frequency, and biological functions using the WGS Annovar-based pipeline[31].

Brain-specific network analysis

We assessed the connectivity patterns of genes with "second-hit" variants and changes in expression in the context of a brain-specific interaction network. This network was previously built using a Bayesian classifier trained on hundreds of gene co-expression, protein-protein interaction, and regulatory-sequence datasets, in order to predict the likelihood of interactions between any two pairs of brain-expressed genes[66,67]. To create a network containing only the highest probability predicted gene interactions, we extracted all pairs of genes with weighted probabilities >2.0, representing the top ~0.5% of pairwise interactions (217,975,718 pairs of genes). We then calculated the weighted shortest path lengths for all pairs of genes in the network, using the inverse of the probabilities as weights for each edge. Finally, we created sub-networks that contained genes with "second-hit" protein-coding variants (LOF, missense, splice-site, exonic STR, or encapsulated deletion or duplication) or expression changes (differential expression, outlier expression, alternative splicing, ASE, or eQTL minor allele) for each carrier offspring from the 13 trios (**Table S1**). For each trio, we calculated the average shortest paths between all pairs of genes with expression changes and genes with "second-hit" coding variants, and then compared these distances to average shortest paths calculated from 100 permuted network replicates, where genes were randomly reassigned to different nodes. Network analysis was performed using the NetworkX package in Python[68].

Statistical analysis

All genomic and statistical analyses were conducted using either Python v.3.7.3, with packages numpy[69] v.1.16.2, scipy[70] v.1.1.0, and pandas[71] v.1.0.0, or using R v.3.5.1 (R Foundation for Statistical Computing, Vienna, Austria). Details of all statistical tests, including summary statistics, test statistics, p-values, and confidence intervals, are provided in **Table S4**.

RESULTS

The 16p12.1 deletion leads to pervasive disruption of genes involved in neurodevelopment

We performed RNA-sequencing on LCL samples from 19 deletion carriers and 13 noncarriers from five large families with multiple affected members (**Figure 1; Figure S1; Table S1**), and identified 1,569 transcripts that were differentially expressed (FDR<0.05) in deletion carriers compared with noncarriers (**Figures 2A-B; Table S3**). Application of additional corrections for relatedness among the samples[52] (see Methods) yielded 1,044 differentially expressed transcripts, of which 840 (80.5%) were also identified in the main analysis (**Figure S7A; Table S3**). We first confirmed that each of the seven protein-coding genes in the deletion region were downregulated in deletion carriers (**Figure 2C**). Interestingly, 21 protein-coding genes adjacent to the 16p12.1 region (between chromosomal bands 16p11.2 and 16p12.3) also showed differential gene expression in carriers, 18 of which were under-expressed in the deletion carriers. *For example*, two genes within the autism-associated 16p11.2 region, *SEZ6L2* and *DOC2A*, were downregulated in carriers of the 16p12.1 deletion. As none of the carriers harbored an atypical deletion, it is possible that these adjacent genes could be affected by haploinsufficiency of regulatory elements located within the deletion region. In fact, three downregulated genes adjacent to the deletion, *DNAH3*, *OTOA*, and *NPIPB4*, exhibited chromatin interactions with enhancer elements within the deletion region, detected using published Hi-C data of LCL samples[72] (**Figure 2C**).

We found that differentially expressed genes in deletion carriers were enriched (FDR<0.05) for multiple biological functions, including biological adhesion and cell proliferation regulation for relatedness-corrected genes, and signal transduction and locomotion for genes without relatedness correction (**Table S3**). Additionally, we observed an enrichment (FDR=0.015) for candidate autism genes[48], including *FOXP1*, *CUL7*, *ANK3*, and *EP300*, among the differentially-expressed genes (**Figure 2B; Table S3**). Weighted-gene correlation network analysis further identified several modules of genes with significant expression changes in deletion carriers (p<0.05, two-tailed t-test), including downregulated genes enriched for cell signaling and adhesion, and upregulated genes enriched for neurogenesis, nervous system development, and MAPK and Notch signaling (**Figure S8; Table S5**). Differentially expressed genes in deletion carriers were further enriched (FDR<0.05) for genes preferentially expressed in the hippocampus and basal ganglia of the adult brain[6] (**Figure S7B; Table S3**), as well as in the

striatum, thalamus, and frontal cortex during late fetal and adolescent/young adulthood timepoints[58], which are critical transition periods for expression of neurodevelopmental genes[73–75] (**Figure 2D; Table S3**). Overall, our data suggest that the 16p12.1 deletion leads to pervasive transcriptomic changes across multiple biological and neuronal processes in the developing brain.

Family-specific patterns of gene expression changes influence developmental phenotypes

We next investigated how gene expression patterns segregated within 13 complete trios with carrier offspring extracted from the five families, including carrier children compared to their parents as well as carrier parents compared with grandparents (**Table S1**). For each trio, we identified differentially expressed genes for offspring-carrier parent and offspring-noncarrier parent pairs (see Methods), and overlapped the two sets of differentially expressed genes to categorize expression changes based on their inheritance pattern (**Figure S5A; Tables S6, S7**). We found no differences (p=0.735, two-tailed paired Mann-Whitney test) in the proportion of differentially expressed genes inherited from the carrier (avg. 2,223 genes/offspring) or noncarrier parent (avg. 1,908 genes/offspring; **Figure 3A**). This suggests that "second-hit" variants from the noncarrier parent contribute equally to gene expression changes, and therefore to disease pathogenicity, as the combined effects of the deletion and any "second-hit" variants from the carrier parent, an observation that corresponds with our recent findings of increased burden of "second-hits" transmitted to the child from noncarrier parents[2] (**Figure S5B**). Interestingly, we also observed an average of 1,192 genes/offspring that were differentially expressed compared with both parents (**Figure 3A**), such as *SHANK2*, *FOXP1*, and *CACNA1D*. These expression changes potentially represent effects of *de novo* variants or combinatorial effects of variants inherited from both parents, which could explain the increased phenotypic severity observed in the carrier children. However, the trends in expression patterns widely varied across families, which in some cases could be explained by family history of neuropsychiatric disease (**Figure S5C**). For example, we found that children within family GL_004, whose parents were unaffected or presented with mild depression, had the lowest number of gene expression changes among any carrier children in the cohort. Meanwhile, children in families GL_001 and GL_052, whose carrier parents manifested multiple overt

cognitive and neuropsychiatric features, had higher proportions of inherited expression changes from their carrier parents than their noncarrier parents (**Figure S5C**).

We next assessed the dysregulated biological functions in each trio (**Figure S9; Table S7**), and found that *de novo* or inherited differentially expressed genes in carrier children were enriched for biological processes (FDR<0.05) that could be related to 39 out of 47 (83.0%) specific developmental phenotypes observed in the affected children (**Figure 3B**). For example, inherited gene expression changes for carrier child P1C_01 in family GL_001 were enriched for nervous system development, neurotransmitter metabolism, neuron projection, and synaptic transmission functions, while their *de novo* expression changes were enriched for genes involved in skeletal and muscular development. The inherited changes in neuronal genes could contribute to the intellectual disability/developmental delay (ID/DD) and speech delay phenotypes observed in the child, as both parents also had several psychiatric features, while the *de novo* changes in developmental genes could be related to hypotonia, growth delay, and craniofacial features uniquely observed in the child. Overall, these results suggest that expression changes of neurodevelopmental-related genes could account for phenotypic differences among carriers of the 16p12.1 deletion.

"Second-hit" variants and the 16p12.1 deletion show synergistic effects towards gene expression

We next investigated whether changes in gene expression could be attributed to "second-hits", or rare genetic modifiers elsewhere in the genome. Rare variants disrupting protein-coding regions and nearby regulatory elements have been previously linked to gene expression changes in both control populations[26,76–78] and disease cohorts, where causal genes may be missed by DNA sequencing methods[21,23–25]. We hypothesized that "second-hits" by themselves or in combination with the deletion could contribute to the observed gene expression changes in affected children. We therefore identified 25 classes of rare gene-disruptive "second-hit" variants from WGS data for each individual (**Figure S2; Tables S8, S9**), including SNVs and indels in coding and non-coding regulatory regions (UTRs, introns, and putative promoter, enhancer, and silencer elements within 50 kbp of a gene) with Phred-like CADD scores[4] >10, and CNVs and short tandem repeats (STRs) that spanned gene-coding regions. We then calculated the likelihood that these "second-hit" variants are associated with changes in expression of a proximal gene, as

determined by either differential expression analysis between carrier offspring and their parents in the 13 trios, or outlier expression analysis among all individuals in the cohort[21,76] (see Methods). While family-based differential expression analysis detects all expression changes between affected children and their parents[79], including those due to the downstream effects of the deletion, outlier analysis more robustly identifies specific effects of "second-hits" towards larger changes in expression, including synergistic effects in combination with the deletion. Overall, we observed an average of 285 outlier genes (|z-score| >2) per individual, including candidate neurodevelopmental genes[47] such as *CTNNB1*, *FOXG1*, *DISC1*, and *ZNF804A* (**Figure S6; Tables S6, S10**). We found that 10.8% of outlier genes (avg. 31/286 per person) and 11.2% of differentially expressed genes (avg. 310/2,774 per carrier offspring) were potentially disrupted by a rare coding or non-coding variant (**Table S9**). Altered expression of genes without such variants could be due to several factors, such as common variants, DNA methylation events, downstream effects of other dysregulated genes, or environmental factors.

In agreement with previous studies[26,76], we found that genes with outlier expression were enriched ($p<0.05$, Fisher's exact test) for 6/25 classes of rare variants that directly affected gene-coding regions, including loss-of-function (LOF), missense, and splice-site SNVs, 5' UTR overhanging deletions, and 5' UTR overhanging and gene-encapsulating duplications (**Figure 4A**). Similarly, we found that 10/25 variant classes were significantly associated with differentially expressed genes in carrier offspring for the 13 trios ($p<0.05$, Fisher's exact test), including coding missense SNVs, duplications overhanging the 5' UTR, and encapsulated deletions (**Figure S10A**). We further found that outlier genes had significantly higher burdens of rare variants in aggregate ($p=1.01\times10^{-3}$, one-tailed t-test) and for 7/25 individual classes ($p<0.05$) compared with non-outlier genes (**Figures S11A-B**). Interestingly, we also found that outlier genes that were intolerant to variation (pLI score >0.9 or RVIS percentile <20; $p<3.26\times10^{-4}$, two-tailed t-test) or preferentially expressed in the brain ($p=0.011$) had a higher burden of nearby rare variants than other outlier genes (**Figure S11C**). We next assessed the effect size of "second-hits" towards outlier gene expression among carriers and noncarriers of the deletion, and found significant enrichments ($p<0.05$, Fisher's exact test) of LOF variants, 5' UTR overhanging duplications, and 5' UTR-disrupting SNVs towards outlier expression in carrier children but not in carrier parents (**Figure 4B**). We observed similar findings among differentially expressed genes, where missense SNVs, upstream SNVs, encapsulated deletions, and 5' UTR overhanging

duplications were only likely to alter gene expression in carrier children (**Figure S10B**). Notably, LOF variants and 5' UTR overhanging duplications were also enriched for outlier expression in noncarrier parents, suggesting that these classes of "second-hit" variants were more likely to be deleterious in carrier children when inherited from noncarrier parents (**Figure 4B**). In fact, we observed five classes of "second-hit" variants in carrier children, including gene-encapsulating and 3' UTR overhanging deletions, 5' UTR overhanging duplications, and missense and upstream SNVs, that had significant effects towards outlier expression ($p<0.05$, Fisher's exact test) when inherited from the noncarrier parent but not from the carrier parent (**Figure 4C**). Similarly, we found that LOF variants and SNVs in silencer regions correlated with differential gene expression in carrier children only when inherited from the noncarrier parent ($p<0.05$, Fisher's exact test) (**Figure S10C**). For example, a carrier child in family GL_007, who exhibited hypotonia and muscle weakness, inherited a deleterious variant from their noncarrier parent in the 5' UTR of *EIF2AK1* [MIM: 613635], associated with motor dysfunction[80], that potentially led to downregulation of that gene (**Figure 4D**). Overall, our findings showed that distinct classes of "second-hit" variants differentially contribute towards changes in gene expression when inherited in a complex manner from either the carrier or noncarrier parent.

We next investigated whether "second-hit" variants showed synergistic effects towards expression changes in genes also dysregulated by the 16p12.1 deletion. We found 11 instances of genes, such as *ANK3, DOCK10,* and *SLC26A1*, that were differentially expressed in all deletion carriers, showed outlier expression in an individual deletion carrier, and had a nearby variant (two coding and nine non-coding) inherited from the noncarrier parent (**Figure 4E; Figure S12; Table S11**). For example, two carrier children in family GL_005 inherited an intronic variant within *ZEB2* [MIM: 605802] from their noncarrier parent, whose altered dosage is associated with Mowat-Wilson syndrome[81,82] [MIM: 235730]. While both the 16p12.1 deletion and the "second-hit" variant individually corresponded with increased *ZEB2* expression, the presence of both variants in the carrier children resulted in even stronger overexpression of the gene compared to those with either individual variant (**Figure 4E**). Overexpression of *ZEB2* could contribute to the Mowat-Wilson like features observed in the carrier child P1C_05, including ID/DD, seizures, hypotonia, and digestive abnormalities. Similarly, a carrier child in GL_005 inherited a rare variant in a promoter region upstream of the mRNA splicing-associated[83] gene *SNRNP27* from their noncarrier parent. *SNRNP27* is over-expressed in deletion carriers but

under-expressed in both the carrier child and the noncarrier parent, highlighting a case where a "second-hit" variant reverses an expression change caused by the deletion (**Figure 4E**). Furthermore, a carrier child in GL_004 shared an intronic variant with two noncarrier relatives in the gene *SYNJ1* [MIM: 604297], which is associated with synaptic transmission[84] and is under-expressed in carriers of the deletion. While other individuals with the same variant had normal *SYNJ1* expression, the carrier child exhibited under-expression of the gene compared to both carriers and noncarriers of the deletion, suggesting that the variant may alter *SYNJ1* expression only in the presence of the deletion (**Figure 4E**). While it is possible that other variants elsewhere in the genome could also influence expression levels of these genes, these examples highlight putative synergistic effects between the 16p12.1 deletion and "second-hit" variants towards gene expression, where the "second-hit" variants may amplify or reduce the effects of the CNV.

A broad range of transcriptomic alterations contribute to phenotypic variability of the 16p12.1 deletion

To identify a complete spectrum of gene expression alterations in each individual, we next evaluated alternative gene splicing, allele-specific expression (ASE), and expression quantitative trait loci (eQTL) among individuals in our cohort. We first identified an average of 3,267 alternative isoforms present in carrier offspring of the 13 trios compared to their parents (**Tables S6, S12**), including for several neurodevelopmental-associated genes[47] such as *KANSL1*, *SHANK2*, and *SYNGAP1*. After categorizing splicing events by inheritance, we found no differences between splicing events inherited from carriers (average=1,307) or noncarriers (average=1,392; p=0.635, two-tailed paired Mann-Whitney test), with fewer changes occurring *de novo* (average=568; $p=2.44\times10^{-4}$; **Figure S13A**). We next found enrichments ($p<0.05$, Fisher's exact test) for alternative splicing in genes disrupted by "second-hit" splice-site, intronic, or missense SNVs, interstitial and 3' UTR overhanging deletions, and intronic STRs (**Figure S13B; Table S9**). We also found that intronic SNVs were more likely to disrupt splicing in carrier children if they were inherited from the noncarrier parent (p=0.033, Fisher's exact test) than the carrier parent (p=0.132, **Figure 5A**), while missense SNVs were more likely to lead to alternative splicing when present in carrier children (p=0.034) than in carrier parents (p=0.168) (**Figure S13C**). For example, a deleterious splice-site variant in the transcriptional activator

TADA2A [MIM: 602276] led to an alternate isoform (*TADA2A*-003) in multiple family members of GL_007 that was not observed in GTEx LCL data (**Figure 5B**). *TADA2A* is a candidate gene within the schizophrenia-associated 17q12 deletion[85] [MIM: 614527], and five out of six family members with the splicing variant have schizophrenia-like clinical features (i.e. hallucinations or delusions), including four deletion carriers and one noncarrier child.

Next, we identified an average of 285 genes with ASE per individual in our cohort (**Tables S6, S13**, including for the neurodevelopmental-associated genes[47] *DNMT3A*, *NSUN2*, and *HDAC8*. ASE events in offspring of the 13 trios were more likely to occur *de novo* than inherited ($p=2.44\times10^{-4}$, two-tailed paired Mann-Whitney test), in contrast to differential expression and alternative splicing events (**Figure S14A**). Genes with ASE have previously been shown to have a higher burden of nearby rare deleterious variants[76], and the pathogenicity of a gene with ASE increases with the presence of a deleterious variant on the overexpressed allele[78]. In our cohort, we found five ASE events in carrier children that led to overexpression of a deleterious "second-hit" coding variant (**Figure S14B**). For example, two carrier children with autism in family GL_007 showed overexpression of a deleterious "second-hit" missense variant in the candidate autism gene[86] *CARD11* [MIM: 607210], which was inherited from their noncarrier parent (**Figure 5C**).

We further performed eQTL discovery analysis to identify variants statistically correlated with expression changes in our cohort, agnostic to variant pathogenicity or population frequency. We identified 21 eQTLs which modulated the expression of 23 eGenes (**Figure S15A; Table S14**), of which 19 were not present in GTEx LCL data, representing unique loci in our cohort. Carrier children showed a trend (p=0.107, two-tailed Mann-Whitney test) towards carrying a higher number of minor eQTLs alleles (average=4.3/person) than their carrier parents (average=3.2/person) (**Figure S15B; Table S6**). Furthermore, several eGenes had biological functions related to neuronal processes[87–89] (**Table S14**), including *SERPINF1*, *BEGAIN*, and *ARFGEF2*. For example, we identified a relatively rare eQTL (allele frequency = 0.015) for overexpression of *ARHGAP39* [MIM: 615880], a key regulator of neurogenesis and dendrite morphology associated with learning and memory[90] (**Figure 5D**). The eQTL minor allele, located in a transcription factor binding cluster, was only observed in a carrier child and their noncarrier parent within GL_007, who both presented with neuropsychiatric phenotypes.

To assess the joint contributions of each type of expression change among the individuals in our cohort, we assessed correlations between the numbers of gene expression changes assessed in our study by inheritance pattern (**Figure 5E**). We observed nine significant positive correlations ($p<0.05$, Pearson correlation) between pairs of gene expression changes in each person, which often shared the same inheritance patterns. For example, the number of genes with differential expression strongly correlated with the number of genes with alternative splicing when inherited from the carrier parent ($r=0.93$, $p=4.41\times10^{-6}$) or non-carrier parent ($r=0.91$, $p=1.37\times10^{-5}$). Together, the correlations between transcriptomic alterations suggest that different types of gene expression changes could be inherited together, potentially due to the same "second-hit" variants disrupting expression of similar genes and biological pathways, as is observed for signals in genome-wide association studies[91].

Genes with "second-hit" variants and expression changes show strong connectivity in a brain-specific network

Finally, to determine whether associations between transcriptomic changes and "second-hit" variants in LCL samples were also relevant in the nervous system, we assessed connectivity patterns of genes with "second-hit" variants and altered gene expression using a brain-specific interaction network[66,67]. We generated individual networks for carrier offspring in the 13 trios, and calculated shortest distances between genes with protein-coding "second-hit" variants and genes with observed expression changes in each offspring (see Methods). We found that the average shortest distances between genes with "second-hits" and expression changes were significantly smaller in 7/13 offspring than those derived from permuted networks ($p<0.05$, one-tailed z-test with $n=100$ permutations). In fact, networks for offspring in aggregate had significantly smaller shortest distances than the shortest distances from the sets of permuted networks ($p=4.88\times10^{-4}$, two-tailed paired Mann-Whitney test) (**Figure 5F**). These data indicate that "second-hit" variants closely interact with genes with expression changes in the brain, suggesting a potential mechanism for how gene expression changes that underlie developmental phenotypes can be influenced by "second-hit" variants in the genome.

DISCUSSION

We previously described a two-hit model for the 16p12.1 deletion, where the presence of both the deletion and "second-hit" variants determine the phenotypic trajectory of affected children[2,15]. Here, we propose a mechanism for how the deletion and "second-hits" jointly interact to alter clinical phenotypes by way of the transcriptome. We found that the 16p12.1 deletion itself disrupts the expression of genes across the genome through direct effects, such as chromatin interactions, and through indirect effects, such as downstream genetic interactions (**Figure 6**). For example, chromatin interactions were observed between regions within the 16p12.1 deletion and flanking genes such as *DNAH3*, and 1,493 genes outside of chromosome 16 were also dysregulated in deletion carriers. Similarly, we found that "second-hits" disrupt gene expression through both direct and indirect mechanisms. Genes with nearby "second-hit" variants were more likely to exhibit outlier expression and alternative splicing, and genes with "second-hits" were more closely connected to genes with expression changes in a brain-specific network than random sets of genes in permuted networks. In fact, we observed 11 examples of combined effects of the deletion and "second-hit" variants towards expression in our cohort, including the candidate neurodevelopmental genes *SYNJ1* and *ZEB2*. These synergistic effects towards gene expression are similar to those previously observed for eQTLs[92] and HLA alleles[93], except that these effects are potentially due to the combined effects of rare deleterious variants. These results suggest that the 16p12.1 deletion and the "second-hit" variants jointly mold the shape of the transcriptome, resulting in strong dysregulation of developmental genes and contributing to neuropsychiatric features in 16p12.1 deletion carriers.

Results from our study align with previous studies, which found that rare variants of different classes have varying effect sizes towards gene expression[26,76]. Our study extends this paradigm by identifying classes of rare "second-hit" variants whose contributions to gene expression changes differ by inheritance pattern. We found that high-effect variants, such as duplications and loss-of-function variants, always cause expression changes regardless of parent-of-origin, and splice-site variants almost always lead to changes in isoform expression. In contrast, medium-effect variants, including missense, intronic, and upstream SNVs, were more strongly associated with gene expression changes when inherited from the noncarrier parent than the carrier parent. These findings indicate that the noncarrier parents are more likely to pass gene expression-altering "second-hit" variants down to their carrier children, providing an explanation

for why children with the deletion have more severe phenotypic manifestations than their carrier parents (**Figure 6**). Future family-based transcriptome studies with larger sample sizes may be able to pinpoint specific rare variants within dysregulated genes that are associated with distinct phenotypes in the carrier children.

We also identified putative biological and developmental pathways disrupted by both the deletion and "second-hit" variants. For example, we found that genes differentially expressed by the deletion were preferentially expressed in multiple brain tissues during development, and were enriched for core signaling and developmental pathways. In fact, knockdown of individual homologs of 16p12.1 genes in *Drosophila melanogaster* models showed neuronal phenotypes and transcriptome disruptions, suggesting that the individual effects of multiple genes in the deletion sensitize the genome for neuropsychiatric phenotypes[94]. Interestingly, we found several examples of biological functions and mechanisms that were simultaneously dysregulated by both the deletion and "second-hit" variants in the carrier children. For example, most carrier children inherited differential expression in genes enriched for nervous system development, cell adhesion, signaling, and locomotion from both their carrier and noncarrier parents. These results provide insights into how the deletion and "second-hit" variants synergistically dysregulate genes and pathways related to development, ultimately contributing towards a wide range of developmental phenotypes observed in children with the deletion.

Some limitations can be noted in the context of our study. *First*, we investigated gene expression changes within patient-derived LCL samples, which have reduced relevance for brain expression. However, over 80% of genes expressed in GTEx brain samples, as well as over 70% of neurodevelopmental disease genes, were expressed in our LCL samples (**Figures S4A-B**). Nevertheless, repeating the study in tissues that are implicated in neurodevelopmental disorders, potentially using patient-derived reprogrammed neuronal progenitor cells, would verify the associations between variants, expression changes, and clinical features. *Second*, we have a relatively small cohort of 32 individuals within five families. It would be useful to determine whether the identified associations are strengthened in a larger cohort. Phenotypically more diverse cohorts would also allow for performing additional correlations between gene expression changes and specific clinical features, such as whether more outlier genes are present among families with stronger histories of neuropsychiatric disease. Overall, our work supports a model for complex disorders, where combinatorial effects of multiple variants with different effect sizes

affect expression of genes in developmental pathways, which further influence the expressivity of clinical features. These results exemplify that family-based transcriptome studies, similar to family-based genome studies, can help explain changes in phenotypes from parents to children and between siblings, especially in complex disorders with a high degree of intra- and inter-familial variability.

SUPPLEMENTAL DATA

Supplemental Data include 15 figures and 14 tables in Excel format.

ACKNOWLEDGEMENTS

We are grateful to all members of the five families who participated in this study. We also thank Sherryann Wert, Melissa Berkowitz, and Tara Schmidlen (Coriell Institute) for their assistance in the generation of stable LCL lines; Craig Praul and Yuka Imamura (Penn State Genomics Core Facility) for assistance with designing the RNA-sequencing strategy; Casey Brown (UPenn), Istvan Albert (Penn State), and Qunhua Li (Penn State) for assistance with statistical analysis of transcriptome data; and Jesse Gillis (CSHL), Dajiang Liu (Penn State), and members of the Girirajan lab for useful discussions and critical reading of the manuscript. We appreciate access to data from the Genotype-Tissue Expression (GTEx) Project, which was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal in January 2020. This work was supported by NIH R01-GM121907 and resources from the Huck Institutes of the Life Sciences to S.G. M.J. was supported by NIH T32-GM102057. A.T. was supported by NIH T32-LM012415. L.P. was supported by Fulbright Commission Uruguay-ANII. A.K. was supported by NIH R35-GM128765. The funding bodies had no role in data collection, analysis, or interpretation.

WEB RESOURCES

BrainSpan Atlas, https://www.brainspan.org/

CADD, https://cadd.gs.washington.edu/

ChromHMM, https://www.encodeproject.org/software/chromhmm/

DBD Gene Database, https://dbd.geisingeradmi.org/

DDD, https://www.ddduk.org/

DDG2P, https://decipher.sanger.ac.uk/ddd/ddgenes

Gene Ontology, http://geneontology.org/

gnomAD/pLI, https://gnomad.broadinstitute.org/

GTEx, https://www.gtexportal.org/home/

MutationTaster, http://www.mutationtaster.org/

OMIM, https://www.omim.org/

RVIS, http://genic-intolerance.org/

SFARI Gene, https://gene.sfari.org/

DATA AND CODE AVAILABILITY

Patient-derived LCL samples generated in this study are available at the NIGMS Human Genetic Cell Repository at the Coriell Institute (https://www.coriell.org/1/NIGMS). Whole genome sequencing and RNA-sequencing data generated in this study will be available at the NCBI dbGaP database (https://www.ncbi.nlm.nih.gov/gap/). Accession numbers for LCL samples and sequencing data are provided in **Table S2**. All code generated for this project, including pipelines for running bioinformatic software and custom analysis scripts, is available at https://github.com/girirajanlab/16p12.1_rnaseq_analysis and https://github.com/girirajanlab/16p12.1_WGS_analysis.

REFERENCES

1. Eichler, E.E. (2019). Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. N. Engl. J. Med. *381*, 64–74.

2. Pizzo, L., Jensen, M., Polyak, A., Rosenfeld, J.A., Mannik, K., Krishnan, A., McCready, E., Pichon, O., Le Caignec, C., Van Dijck, A., et al. (2019). Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. Genet. Med. *21*, 816–825.

3. Berkovic, S.F., Scheffer, I.E., Petrou, S., Delanty, N., Dixon-Salazar, T.J., Dlugos, D.J., Helbig, I., Frankel, W.N., Goldstein, D.B., Heinzen, E.L., et al. (2015). A roadmap for precision medicine in the epilepsies. Lancet Neurol. *14*, 1219–1228.

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

5. Schaaf, C.P., Betancur, C., Yuen, R.K.C., Parr, J.R., Skuse, D.H., Gallagher, L., Bernier, R.A., Buchanan, J.A., Buxbaum, J.D., Chen, C.A., et al. (2020). A framework for an evidence-based gene list relevant to autism spectrum disorder. Nat. Rev. Genet. *21*, 367–376.

6. Ardlie, K.G., DeLuca, D.S., Segrè, A. V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660.

7. Kavanagh, D.H., Tansey, K.E., O'Donovan, M.C., and Owen, M.J. (2015). Schizophrenia genetics: Emerging themes for a complex disorder. Mol. Psychiatry *20*, 72–76.

8. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genet. *9*, e1003709.

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

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

11. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177–1186.

12. Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature *506*, 185–190.

13. Coe, B.P., Girirajan, S., and Eichler, E.E. (2012). A genetic model for neurodevelopmental disease. Curr. Opin. Neurobiol. *22*, 829–836.

14. Girirajan, S., Rosenfeld, J.A., Cooper, G.M., Antonacci, F., Siswara, P., Itsara, A., Vives, L.,

Walsh, T., McCarthy, S.E., Baker, C., et al. (2010). A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat. Genet. *42*, 203–209.

15. Girirajan, S., Rosenfeld, J.A., Coe, B.P., Parikh, S., Friedman, N., Goldstein, A., Filipink, R.A., McConnell, J.S., Angle, B., Meschino, W.S., et al. (2012). Phenotypic Heterogeneity of Genomic Disorders and Rare Copy-Number Variants. N. Engl. J. Med. *367*, 1321–1331.

16. Rees, E., Kendall, K., Pardiñas, A.F., Legge, S.E., Pocklington, A., Escott-Price, V., MacCabe, J.H., Collier, D.A., Holmans, P., O'Donovan, M.C., et al. (2016). Analysis of intellectual disability copy number variants for association with schizophrenia. JAMA Psychiatry *73*, 963–969.

17. Mefford, H.C., Muhle, H., Ostertag, P., von Spiczak, S., Buysse, K., Baker, C., Franke, A., Malafosse, A., Genton, P., Thomas, P., et al. (2010). Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. PLoS Genet. *6*, e1000962.

18. Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K., Arnarsdottir, S., Bjornsdottir, G., Walters, G.B., Jonsdottir, G.A., Doyle, O.M., et al. (2014). CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature *505*, 361–366.

19. Girirajan, S., and Eichler, E.E. (2010). Phenotypic variability and genetic susceptibility to genomic disorders. Hum. Mol. Genet. *19*, R176-87.

20. Merla, G., Howald, C., Henrichsen, C.N., Lyle, R., Wyss, C., Zabot, M.T., Antonarakis, S.E., and Reymond, A. (2006). Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. Am. J. Hum. Genet. *79*, 332–341.

21. Luo, R., Sanders, S.J., Tian, Y., Voineagu, I., Huang, N., Chu, S.H., Klei, L., Cai, C., Ou, J., Lowe, J.K., et al. (2012). Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. Am. J. Hum. Genet. *91*, 38–55.

22. Blumenthal, I., Ragavendran, A., Erdin, S., Klei, L., Sugathan, A., Guide, J.R., Manavalan, P., Zhou, J.Q., Wheeler, V.C., Levin, J.Z., et al. (2014). Transcriptional consequences of 16p11.2 deletion and duplication in mouse cortex and multiplex autism families. Am. J. Hum. Genet. *94*, 870–883.

23. Frésard, L., Smail, C., Ferraro, N.M., Teran, N.A., Li, X., Smith, K.S., Bonner, D., Kernohan, K.D., Marwaha, S., Zappala, Z., et al. (2019). Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat. Med. *25*, 911–919.

24. Gonorazky, H.D., Naumenko, S., Ramani, A.K., Nelakuditi, V., Mashouri, P., Wang, P., Kao, D., Ohri, K., Viththiyapaskaran, S., Tarnopolsky, M.A., et al. (2019). Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. Am. J. Hum. Genet. *104*, 466–483.

25. Mohammadi, P., Castel, S.E., Cummings, B.B., Einson, J., Sousa, C., Hoffman, P., Donkervoort, S., Jiang, Z., Mohassel, P., Reghan Foley, A., et al. (2019). Genetic regulatory variation in populations informs transcriptome analysis in rare disease. Science *366*, 351–356.

26. Pala, M., Zappala, Z., Marongiu, M., Li, X., Davis, J.R., Cusano, R., Crobu, F., Kukurba, K.R., Gloudemans, M.J., Reinier, F., et al. (2017). Population- and individual-specific regulatory variation in Sardinia. Nat. Genet. *49*, 700–707.

27. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

28. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

29. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

30. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Auwera, G.A. Van der, Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv 201178.

31. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

32. Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.X., et al. (2015). Excess of rare, inherited truncating mutations in autism. Nat. Genet. *47*, 582–588.

33. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

34. Schwarz, J.M., Cooper, D.N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: Mutation prediction for the deep-sequencing age. Nat. Methods *11*, 361–362.

35. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

36. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674.

37. Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974–984.

38. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, 333–339.

39. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: A probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.

40. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222.

41. Mousavi, N., Shleizer-Burko, S., Yanicky, R., and Gymrek, M. (2019). Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. *47*, e90.

42. Mousavi, N., Margoliash, J., Pusarla, N., Saini, S., Yanicky, R., and Gymrek, M. (2020). TRTools: a toolkit for genome-wide analysis of tandem repeats. Bioinformatics ePub.

43. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

44. Wang, L., Nie, J., Sicotte, H., Li, Y., Eckel-Passow, J.E., Dasari, S., Vedell, P.T., Barman, P., Wang, L., Weinshiboum, R., et al. (2016). Measure transcript integrity using RNA-seq data. BMC Bioinformatics *17*, 58.

45. Li, B., and Dewey, C.N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

46. Deluca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.D., Williams, C., Reich, M., Winckler, W., and Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics *28*, 1530–1532.

47. Gonzalez-Mantilla, A.J., Moreno-De-Luca, A., Ledbetter, D.H., and Martin, C.L. (2016). A cross-disorder method to identify novel candidate genes for developmental brain disorders. JAMA Psychiatry *73*, 275–283.

48. Abrahams, B.S., Arking, D.E., Campbell, D.B., Mefford, H.C., Morrow, E.M., Weiss, L.A., Menashe, I., Wadkins, T., Banerjee-Basu, S., and Packer, A. (2013). SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol. Autism *4*, 36.

49. Firth, H. V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Vooren, S. Van, Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am. J. Hum. Genet. *84*, 524–533.

50. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., Van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of genome-wide research data. Lancet *385*, 1305–1314.

51. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

52. Sun, S., Zhu, J., Mozaffari, S., Ober, C., Chen, M., and Zhou, X. (2019). Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. Bioinformatics *35*, 487–496.

53. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A tool set for whole-genome

association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

54. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.

55. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.

56. Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. *11*, R14.

57. The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. *47*, D330–D338.

58. Miller, J.A., Ding, S.-L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., Aiona, K., et al. (2014). Transcriptional landscape of the prenatal human brain. Nature *508*, 199–206.

59. Langfelder, P., and Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

60. Soneson, C., Love, M.I., and Robinson, M.D. (2016). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. F1000Res. *4*, 1521.

61. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

62. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

63. Castel, S.E., Mohammadi, P., Chung, W.K., Shen, Y., and Lappalainen, T. (2016). Rare variant phasing and haplotypic expression from RNA sequencing with phASER. Nat. Commun. *7*, 12817.

64. Martin, M., Patterson, M., Garg, S., O Fischer, S., Pisanti, N., Klau, G., Schöenhuth, A., and Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. BioRxiv 085050.

65. Delaneau, O., Ongen, H., Brown, A.A., Fort, A., Panousis, N.I., and Dermitzakis, E.T. (2017). A complete tool set for molecular QTL discovery and analysis. Nat. Commun. *8*, 15452.

66. Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. *47*, 569–576.

67. Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., and Troyanskaya, O.G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat. Neurosci. *19*, 1454–1462.

68. Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008). Exploring network structure, dynamics, and function using NetworkX. In 7th Python in Science Conference (SciPy 2008), pp. 11–15.

69. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D.,

Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362.

70. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

71. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, pp. 56–61.

72. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

73. Werling, D.M., Pochareddy, S., Choi, J., An, J.Y., Sheppard, B., Peng, M., Li, Z., Dastmalchi, C., Santpere, G., Sousa, A.M.M., et al. (2020). Whole-Genome and RNA Sequencing Reveal Variation and Transcriptomic Coordination in the Developing Human Prefrontal Cortex. Cell Rep. *31*, 107489.

74. Courchesne, E., Gazestani, V.H., and Lewis, N.E. (2020). Prenatal Origins of ASD: The When, What, and How of ASD Development. Trends Neurosci. *43*, 326–342.

75. Skene, N.G., Roy, M., and Grant, S.G. (2017). A genomic lifespan program that reorganises the young adult brain is targeted in schizophrenia. Elife *6*, e17915.

76. Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J., et al. (2017). The impact of rare variation on gene expression across tissues. Nature *550*, 239–243.

77. Zhao, J., Akinsanmi, I., Arafat, D., Cradick, T.J., Lee, C.M., Banskota, S., Marigorta, U.M., Bao, G., and Gibson, G. (2016). A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. Am. J. Hum. Genet. *98*, 299–309.

78. Castel, S.E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., Guigo, R., Iossifov, I., Vasileva, A., and Lappalainen, T. (2018). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat. Genet. *50*, 1327–1334.

79. Ballouz, S., Dörfel, M., Crow, M., Crain, J., Faivre, L., Keegan, C.E., Kitsiou-Tzeli, S., Tzetis, M., Lyon, G.J., and Gillis, J. (2017). Not by systems alone: Replicability assessment of disease expression signals. BioRxiv 128439.

80. Mao, D., Reuter, C.M., Ruzhnikov, M.R.Z., Beck, A.E., Farrow, E.G., Emrick, L.T., Rosenfeld, J.A., Mackenzie, K.M., Robak, L., Wheeler, M.T., et al. (2020). De novo EIF2AK1 and EIF2AK2 Variants Are Associated with Developmental Delay, Leukoencephalopathy, and Neurologic Decompensation. Am. J. Hum. Genet. *106*, 570–583.

81. Yuan, H., Zhang, L., Chen, M., Zhu, J., Meng, Z., and Liang, L. (2015). A de novo triplication on 2q22.3 including the entire ZEB2 gene associated with global developmental delay, multiple congenital anomalies and behavioral abnormalities. Mol. Cytogenet. *8*, 99.

82. Baxter, A.L., Vivian, J.L., Hagelstrom, R.T., Hossain, W., Golden, W.L., Wassman, E.R., Vanzo, R.J., and Butler, M.G. (2017). A Novel Partial Duplication of ZEB2 and Review of

ZEB2 Involvement in Mowat-Wilson Syndrome. Mol. Syndromol. *8*, 211–218.

83. Zahler, A.M., Rogel, L.E., Glover, M.L., Yitiz, S., Ragle, J.M., and Katzman, S. (2018). SNRP-27, the C. Elegans homolog of the tri-snRNP 27K protein, has a role in 5′ splice site positioning in the spliceosome. RNA *24*, 1314–1325.

84. Saito, T., Guan, F., Papolos, D.F., Lau, S., Klein, M., Fann, C.S.J., and Lachman, H.M. (2001). Mutation analysis of SYNJ1: A possible candidate gene for chromosome 21q22-linked bipolar disorder. Mol. Psychiatry *6*, 387–395.

85. Mitchel, M.W., Moreno-De-Luca, D., Myers, S.M., Finucane, B., Ledbetter, D.H., and Martin, C.L. (2016). 17q12 Recurrent Deletion Syndrome. In GeneReviews®, M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J. Bean, K. Stephens, and A. Amemiya, eds. (University of Washington, Seattle).

86. Lim, E.T., Uddin, M., De Rubeis, S., Chan, Y., Kamumbu, A.S., Zhang, X., D'Gama, A.M., Kim, S.N., Hill, R.S., Goldberg, A.P., et al. (2017). Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. Nat. Neurosci. *20*, 1217–1224.

87. Ramírez-Castillejo, C., Sánchez-Sánchez, F., Andreu-Agulló, C., Ferrón, S.R., Aroca-Aguilar, J.D., Sánchez, P., Mira, H., Escribano, J., and Fariñas, I. (2006). Pigment epithelium-derived factor is a niche signal for neural stem cell renewal. Nat. Neurosci. *9*, 331–339.

88. Yao, I., Iida, J., Nishimura, W., and Hata, Y. (2002). Synaptic and nuclear localization of brain-enriched guanylate kinase-associated protein. J. Neurosci. *22*, 5354–5364.

89. De Wit, M.C.Y., De Coo, I.F.M., Halley, D.J.J., Lequin, M.H., and Mancini, G.M.S. (2009). Movement disorder and neuronal migration disorder due to ARFGEF2 mutation. Neurogenetics *10*, 333–336.

90. Nowak, F. V. (2018). Porf-2 = arhgap39 = vilse: A pivotal role in neurodevelopment, learning and memory. ENeuro *5*, e0082-18.2018.

91. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D.B. (2010). Rare Variants Create Synthetic Genome-Wide Associations. PLoS Biol. *8*, e1000294.

92. Powell, J.E., Henders, A.K., McRae, A.F., Kim, J., Hemani, G., Martin, N.G., Dermitzakis, E.T., Gibson, G., Montgomery, G.W., and Visscher, P.M. (2013). Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data. PLoS Genet. *9*, e1003502.

93. Lenz, T.L., Deutsch, A.J., Han, B., Hu, X., Okada, Y., Eyre, S., Knapp, M., Zhernakova, A., Huizinga, T.W.J., Abecasis, G., et al. (2015). Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat. Genet. *47*, 1085–1090.

94. Pizzo, L., Lasser, M., Yusuff, T., Jensen, M., Ingraham, P., Huber, E., Singh, M.D., Monahan, C., Iyer, J., Desai, I., et al. (2021). Functional assessment of the "two-hit" model for neurodevelopmental defects in Drosophila and X. laevis. BioRxiv 2020.09.14.295923.

95. Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M.N.K., Li, Y., Hu, M., et al. (2018). The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. *19*, 151.

FIGURE TITLES AND LEGENDS

Figure 1. Overview of experimental design. We performed whole genome sequencing, RNA sequencing, and clinical phenotyping on five large families (32 total individuals) with the 16p12.1 deletion, indicated with red asterisks in the pedigrees. Children (green) and adults (blue) in the pedigrees are shaded by phenotypic severity, with white indicating no clinical features, lighter shades indicating mild features, darker shades indicating severe features, and grey indicating no phenotypic data available, as detailed in **Figure S1** and **Table S1**. We then performed multiple analyses to assess the role of the deletion and rare "second-hit" variants towards the observed transcriptomic changes and developmental phenotypes, including differential expression between carriers and noncarriers of the deletion, differential expression between parents and carrier offspring in 13 trios from the five families, outlier gene expression among all individuals, identification of additional transcriptomic alterations such as alternative splicing and allele-specific expression, and gene interaction patterns in the context of a brain-specific network.

Figure 2. Differential expression of genes between carriers and noncarriers of the 16p12.1 deletion. (A) Volcano plot showing downstream (non-16p12.1 deletion) genes differentially expressed (FDR<0.05) between carriers (n=19) and noncarriers (n=13) of the deletion. Red circles indicate genes preferentially expressed in GTEx cerebral cortex tissues. **(B)** Scatter plot showing all genes differentially expressed between carriers and noncarriers of the deletion by chromosome, excluding genes on sex chromosomes. Genes are colored by FDR of differential expression. Labeled genes indicate candidate autism genes with differential expression. **(C)** Expression changes and chromatin connectivity of genes within the 16p12.1 region. The top plot shows pairwise chromatin interactions within the 3.5 Mbp 16p12.1 region, with red lines representing stronger Hi-C intensity, while the bottom plot shows $\log_2$-fold change of expression in deletion carriers of genes within and adjacent to the 16p12.1 deletion. The Hi-C data is from previously reported Hi-C experiments of LCL samples[72], and the heatmap was generated using the 3D Genome Browser[95]. **(D)** Line plot shows enrichment ($\log_{10}$ FDR) of differentially expressed genes in deletion carriers for genes preferentially expressed in six select BrainSpan tissues across 11 developmental timepoints.

Figure 3. Differential expression of genes between offspring and carrier and noncarrier parents. (A) Boxplot shows the proportion of differentially expressed genes in carrier offspring of 13 trios (**Table S1**) that either occurred *de novo* or were inherited from their carrier or noncarrier parents (*p<0.05, two-tailed paired Mann-Whitney test). Boxplot indicates median (center line), 25th and 75th percentiles (bounds of box), and minimum and maximum (whiskers). **(B)** Table shows observed clinical features in eight carrier children with overt developmental phenotypes, as well as enrichments (FDR<0.05) of differentially expressed genes in each carrier child for biological functions related to each clinical feature. Cells are colored according to the inheritance pattern of differentially expressed genes for each enriched biological process.

Figure 4. Enrichment of "second-hit" variants near genes with outlier expression. (A) Forest plot shows enrichment (*p<0.05, Fisher's exact test) of genes with outlier expression in all individuals in the cohort (n=32) for rare proximal coding and non-coding variants, including single-nucleotide variants (SNVs) and insertions/deletions (indels) with CADD scores >10 [4], structural variants (SVs), and short tandem repeats (STRs). **(B)** Forest plot shows classes of "second-hit" variants with significant enrichment (*p<0.05, Fisher's exact test) towards genes with outlier expression in carrier children (n=10), carrier parents (n=6), or noncarrier parents (n=6). **(C)** Forest plot shows classes of "second-hit" variants with significant enrichment (*p<0.05, Fisher's exact test) towards genes with outlier expression in carrier children (n=9) that are inherited from either carrier or noncarrier parents. All forest plots show log-odds ratios (dots) and 95% confidence intervals (whiskers). **(D)** Scatter plot shows expression values (transcripts per million, or TPM) for *EIF2AK1* in LCL replicates for all individuals in our cohort. Samples in blue have outlier expression of *EIF2AK1* (z-score <-2) and carry a deleterious "second-hit" variant in the 5' UTR of the gene. **(E)** Scatter plots show expression values (TPM) of genes with synergistic effects due to the 16p12.1 deletion and inherited "second-hit" variants. Blue circles indicate expression values for samples from carrier children and family members with the rare "second-hit" variants, orange circles indicate expression values for samples from family members without the "second-hit" variant, and green circles indicate expression values of samples from other deletion carriers and noncarriers in the cohort. Black lines denote mean gene expression for LCL replicates of each individual.

Figure 5. Alternative splicing, allele specific expression, eQTL, and network analysis. (A) Forest plot shows classes of rare variants with significant enrichment (*$p<0.05$, Fisher's exact test) towards genes with alternative splicing in carrier children (n=9) that are inherited from either carrier or noncarrier parents. Forest plot shows log-odds ratios (dots) and 95% confidence intervals (whiskers). (**B)** Scatter plot shows isoform usage percentage for the isoform *TADA2A*-003 in LCL replicates for individuals in family GL_007. Samples in blue carry a "second-hit" splice-site variant in *TADA2A* and exhibit a higher frequency of the alternative isoform. **(C)** Scatter plot shows allele frequencies for the autism-associated gene *CARD11* in carrier child P2C_07, noncarrier parent F2NC_07, and carrier parent M2C_07 in GL_007. Blue circles indicate allele frequency for haplotypes carrying a "second-hit" coding variant disrupting *CARD11*. **(D)** Scatter plot shows z-scores for expression values of *ARHGAP39* for all individuals with available genotypes for the gene. Two individuals who carry the minor allele for the *ARHGAP39* eQTL (blue dots) have higher expression of the gene than the rest of the cohort (orange dots). **(E)** Plot shows correlations among the numbers of gene expression alterations in carrier offspring for the 13 trios assessed in our study. The color and size of the circles are proportional to the correlation coefficients between classes of gene expression changes, where blue indicates a positive correlation and red indicates a negative correlation. Asterisks denote significant correlations ($p<0.05$, Pearson correlation). **(F)** Boxplot shows the average shortest distances for carrier offspring in the 13 trios between pairs of genes with "second-hit" coding variants and genes with identified expression changes in a brain-specific interaction network. Genes with expression changes were more strongly connected to genes with "second-hit" variants in the brain-specific network than the average distances for genes within 100 permuted brain-specific networks per sample ($p=4.88\times10^{-4}$, two-tailed paired Mann-Whitney test). Boxplot indicates median (center line), 25th and 75th percentiles (bounds of box), and minimum and maximum (whiskers).

Figure 6. Genetic and transcriptomic mechanisms for phenotypic variability in 16p12.1 deletion families. Affected children inherit the 16p12.1 deletion (red) and a smaller number of rare "second-hit" variants (green) from a carrier parent, and a larger number of "second-hit" variants from the noncarrier parent (blue). Altered expression of genes due to these "second-hit" variants affects nearby downstream connected genes in an interaction network (grey), causing

additional transcriptomic perturbation. Because of this, carrier children have numerous gene expression changes compared with their carrier parents, including genes showing synergistic effects of the deletion and “second-hit” variants (orange), potentially accounting for more severe developmental phenotypes observed in the children.

Figure 1

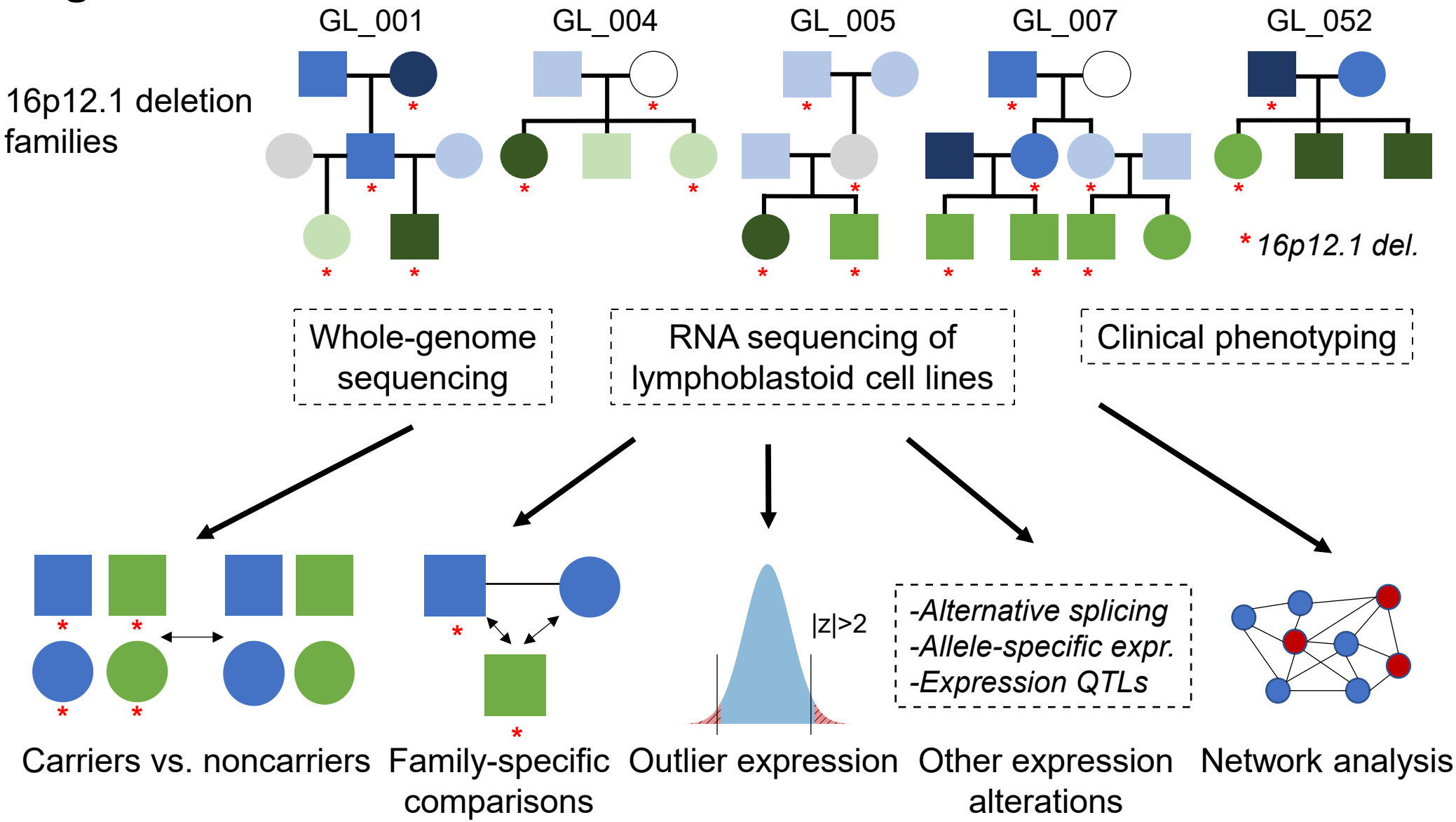

16p12.1 deletion families
GL_001
GL_004
GL_005
GL_007
GL_052
* 16p12.1 del.
Whole-genome sequencing
RNA sequencing of lymphoblastoid cell lines
Clinical phenotyping
|z|>2
-Alternative splicing
-Allele-specific expr.
-Expression QTLs
Carriers vs. noncarriers
Family-specific comparisons
Outlier expression
Other expression alterations
Network analysis

Figure 2

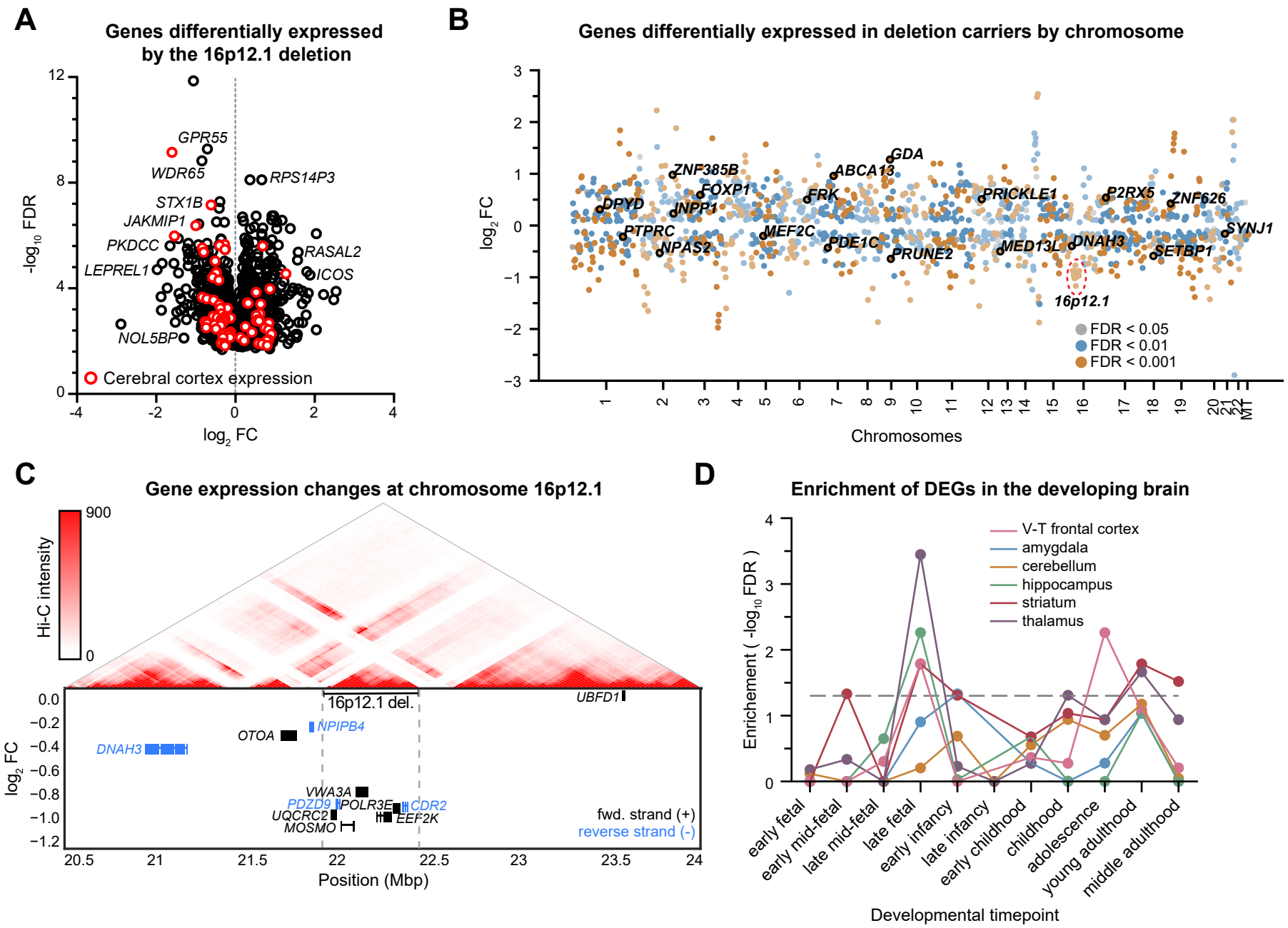
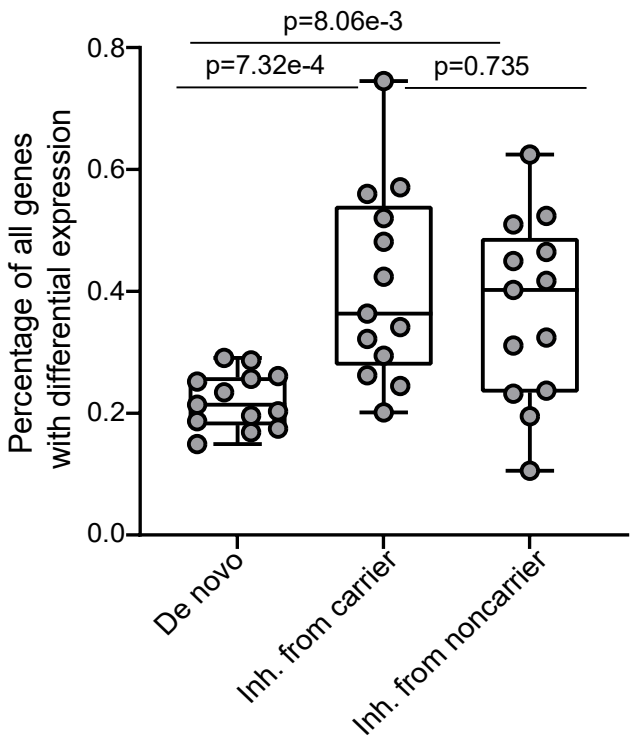

A
Genes differentially expressed by the 16p12.1 deletion
-log10 FDR
log2 FC
GPR55
WDR65
RPS14P3
STX1B
JAKMIP1
PKDCC
LEPREL1
RASAL2
ICOS
NOL5BP
Cerebral cortex expression
B
Genes differentially expressed in deletion carriers by chromosome
log2 FC
Chromosomes
GDA
ZNF385B
ABCA13
FOXP1
FRK
PRICKLE1
P2RX5
ZNF626
DPYD
NPP1
PTPRC
MEF2C
NPAS2
PDE1C
PRUNE2
MED13L
DNAH3
SETBP1
SYNJ1
16p12.1
FDR < 0.05
FDR < 0.01
FDR < 0.001
C
Gene expression changes at chromosome 16p12.1
Hi-C intensity
900
0
16p12.1 del.
UBFD1
NPIPB4
OTOA
DNAH3
VWA3A
PDZD9
POLR3E
CDR2
UQCRC2
EEF2K
MOSMO
fwd. strand (+)
reverse strand (-)
log2 FC
Position (Mbp)
D
Enrichment of DEGs in the developing brain
Enrichement (-log10 FDR)
V-T frontal cortex
amygdala
cerebellum
hippocampus
striatum
thalamus
early fetal
early mid-fetal
late mid-fetal
late fetal
early infancy
late infancy
early childhood
childhood
adolescence
young adulthood
middle adulthood
Developmental timepoint

Figure 3

A

Family-based diff. expression



B

Biological function enrichments for DEGs of carrier childen

Carrier child	P1C_01							P1C_04	P1C_05		P2C_05	P1C_07		P2C_07			P3C_07		P2C_52
Clinical phenotype	Craniofacial features	Micro/macrocephaly	Neuropsych. features	Hypotonia	Brain abnormalities	Kidney/urinary defects	Musculoskeletal defects	Neuropsych. features	Neuropsych. features	Epilepsy	Neuropsych. features	Neuropsych. features	Epilepsy	Neuropsych. features	Hypotonia	Musculoskeletal features	Neuropsych. features	Musculoskeletal features	Neuropsych. features
Nervous system dev.																			
Neurotransmitter metab.																			
Neuron projection																			
Synapse organization																			
Synaptic transmission																			
Axon formation																			
Axon outgrowth																			
Skeletal development																			
Kidney development																			
Muscle development																			

De novo Inh. carrier Inh. noncarrier Inh. and *de novo*

Figure 4

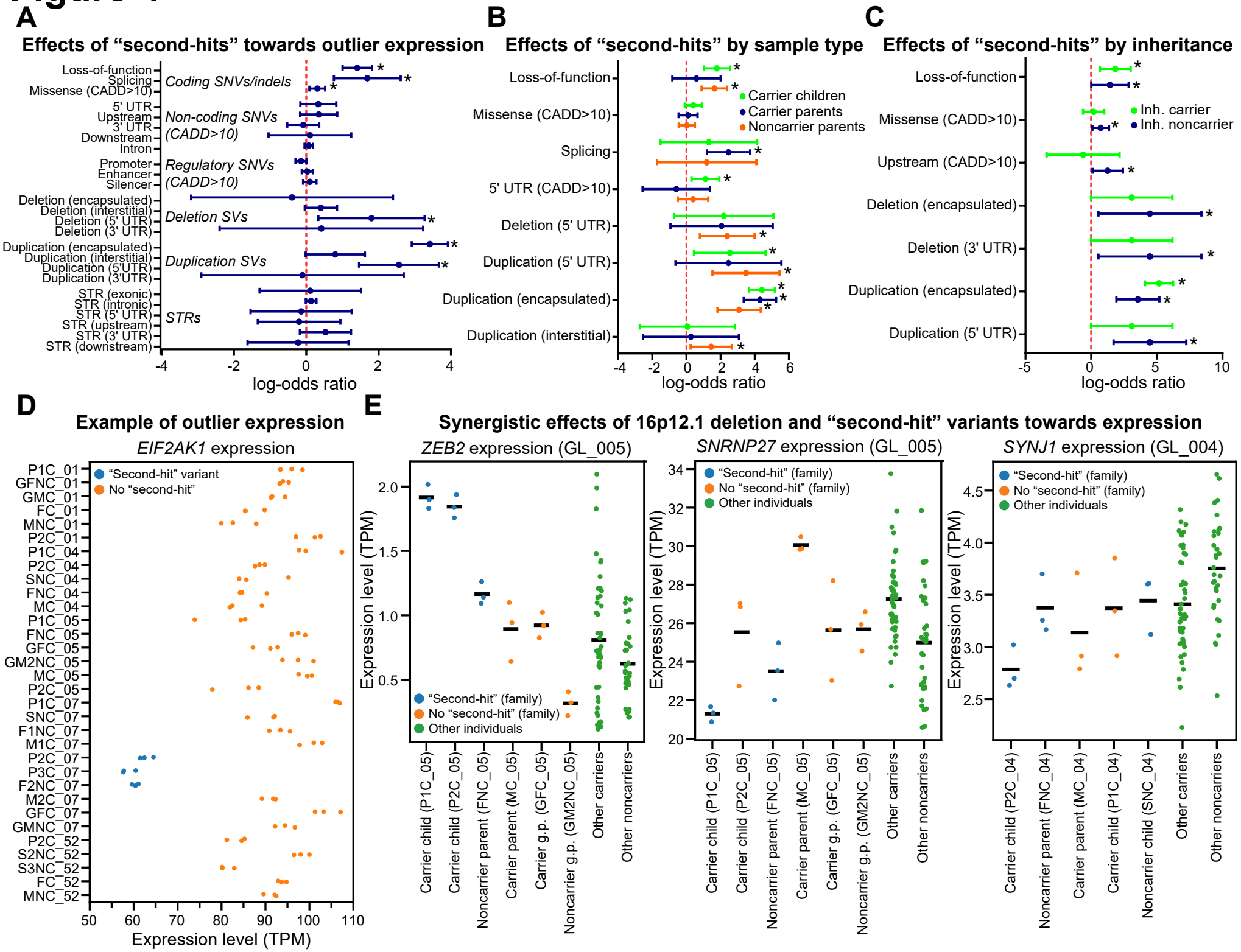

A
Effects of "second-hits" towards outlier expression
Coding SNVs/indels
Loss-of-function
Splicing
Missense (CADD>10)
Non-coding SNVs (CADD>10)
5' UTR
Upstream
3' UTR
Downstream
Intron
Regulatory SNVs (CADD>10)
Promoter
Enhancer
Silencer
Deletion SVs
Deletion (encapsulated)
Deletion (interstitial)
Deletion (5' UTR)
Deletion (3' UTR)
Duplication SVs
Duplication (encapsulated)
Duplication (interstitial)
Duplication (5'UTR)
Duplication (3'UTR)
STRs
STR (exonic)
STR (intronic)
STR (5' UTR)
STR (upstream)
STR (3' UTR)
STR (downstream)
log-odds ratio
B
Effects of "second-hits" by sample type
Carrier children
Carrier parents
Noncarrier parents
Loss-of-function
Missense (CADD>10)
Splicing
5' UTR (CADD>10)
Deletion (5' UTR)
Duplication (5' UTR)
Duplication (encapsulated)
Duplication (interstitial)
log-odds ratio
C
Effects of "second-hits" by inheritance
Inh. carrier
Inh. noncarrier
Loss-of-function
Missense (CADD>10)
Upstream (CADD>10)
Deletion (encapsulated)
Deletion (3' UTR)
Duplication (encapsulated)
Duplication (5' UTR)
log-odds ratio
D
Example of outlier expression
EIF2AK1 expression
"Second-hit" variant
No "second-hit"
Expression level (TPM)
E
Synergistic effects of 16p12.1 deletion and "second-hit" variants towards expression
ZEB2 expression (GL_005)
SNRNP27 expression (GL_005)
SYNJ1 expression (GL_004)
"Second-hit" (family)
No "second-hit" (family)
Other individuals
Expression level (TPM)
Carrier child (P1C_05)
Carrier child (P2C_05)
Noncarrier parent (FNC_05)
Carrier parent (MC_05)
Carrier g.p. (GFC_05)
Noncarrier g.p. (GM2NC_05)
Other carriers
Other noncarriers
Carrier child (P2C_04)
Noncarrier parent (FNC_04)
Carrier parent (MC_04)
Carrier child (P1C_04)
Noncarrier child (SNC_04)

Figure 5

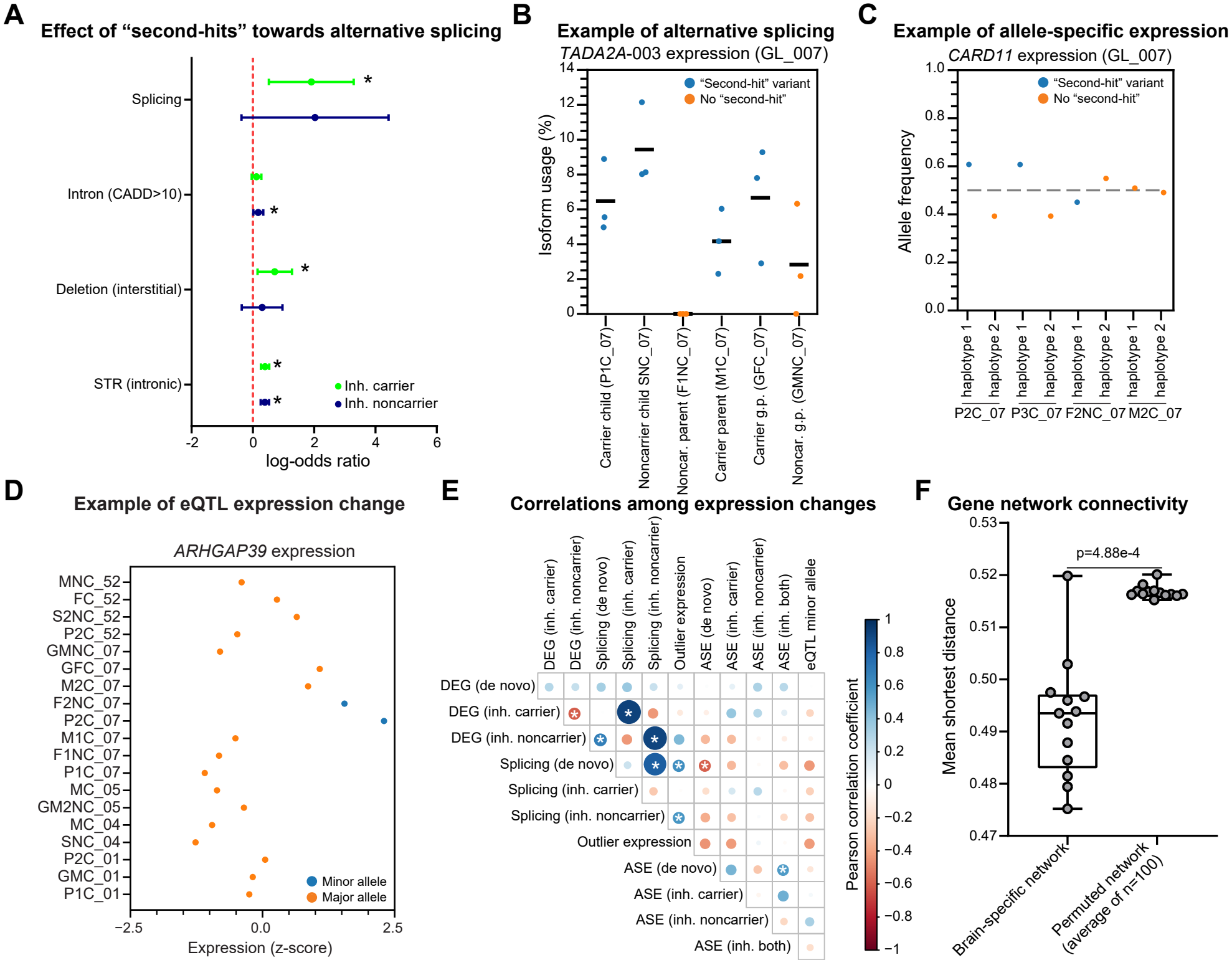
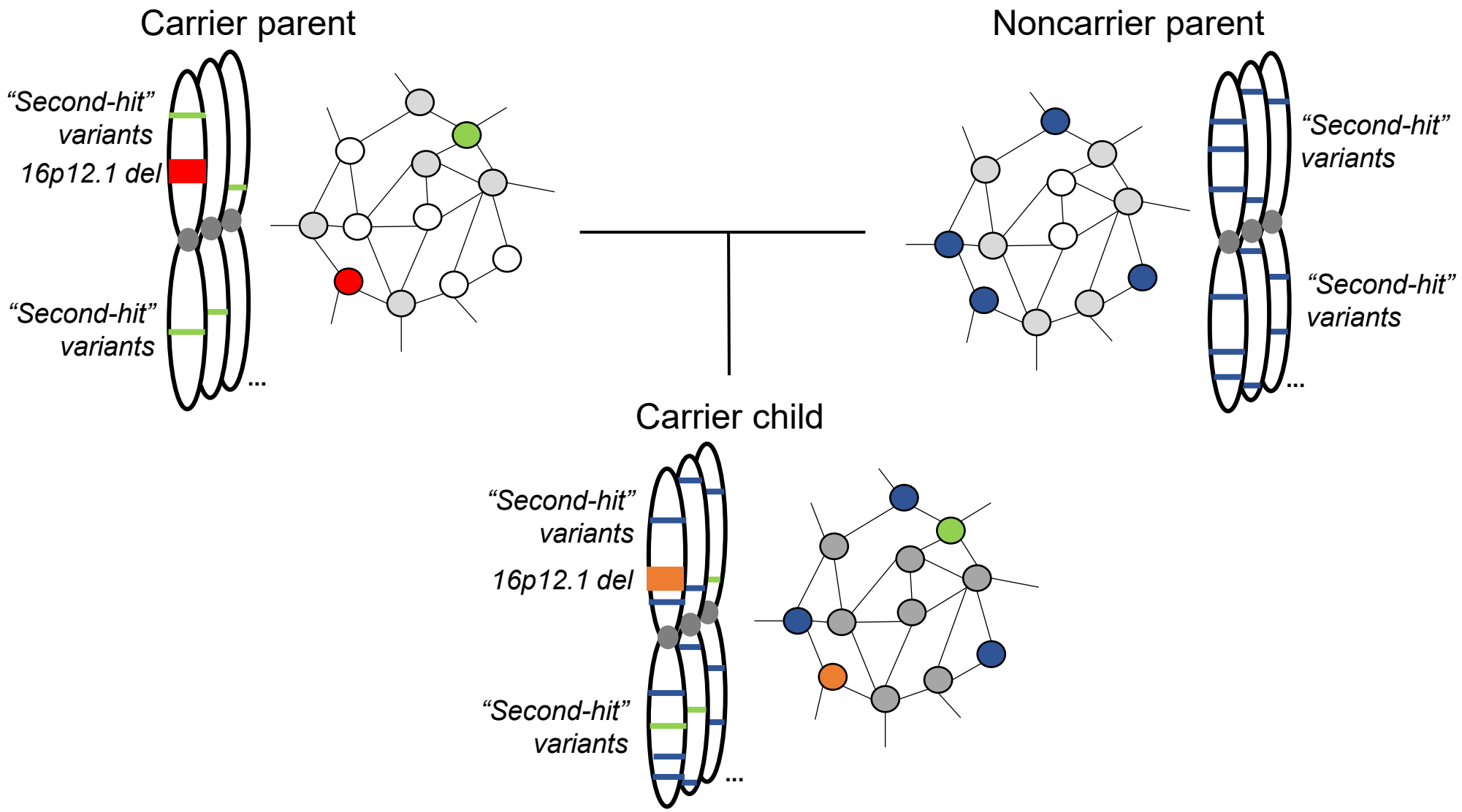

A
Effect of "second-hits" towards alternative splicing
Splicing
Intron (CADD>10)
Deletion (interstitial)
STR (intronic)
Inh. carrier
Inh. noncarrier
log-odds ratio
B
Example of alternative splicing
TADA2A-003 expression (GL_007)
"Second-hit" variant
No "second-hit"
Isoform usage (%)
Carrier child (P1C_07)
Noncarrier child SNC_07)
Noncar. parent (F1NC_07)
Carrier parent (M1C_07)
Carrier g.p. (GFC_07)
Noncar. g.p. (GMNC_07)
C
Example of allele-specific expression
CARD11 expression (GL_007)
"Second-hit" variant
No "second-hit"
Allele frequency
haplotype 1
haplotype 2
P2C_07
P3C_07
F2NC_07
M2C_07
D
Example of eQTL expression change
ARHGAP39 expression
Minor allele
Major allele
Expression (z-score)
E
Correlations among expression changes
DEG (inh. carrier)
DEG (inh. noncarrier)
Splicing (de novo)
Splicing (inh. carrier)
Splicing (inh. noncarrier)
Outlier expression
ASE (de novo)
ASE (inh. carrier)
ASE (inh. noncarrier)
ASE (inh. both)
eQTL minor allele
Pearson correlation coefficient
F
Gene network connectivity
p=4.88e-4
Mean shortest distance
Brain-specific network
Permuted network (average of n=100)

Figure 6


Carrier parent
"Second-hit" variants
16p12.1 del
"Second-hit" variants
...
Noncarrier parent
"Second-hit" variants
"Second-hit" variants
...
Carrier child
"Second-hit" variants
16p12.1 del
"Second-hit" variants
...