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

Matthew Jensen,[1,2,*] Anastasia Tyryshkina,[1,3,*] 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

*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

ABSTRACT

Background: 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.

Methods: 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 identified contributions of the 16p12.1 deletion and "second-hit" variants towards a range of expression changes in deletion carriers and their family members, including differential expression, outlier expression, alternative splicing, allele-specific expression, and expression-quantitative trait loci analyses.

Results: We found that the deletion dysregulates multiple autism and brain development genes such as *FOXP1*, *ANK3*, and *MEF2*. Carrier children also showed an average of 5,323 gene expression changes compared with one or both parents, which matched with 33/39 observed developmental phenotypes. We identified significant enrichments for 13/25 classes of "second-hit" variants in genes with expression changes, where 4/25 variant classes were only enriched when inherited from the non-carrier parent, including loss-of-function SNVs and large duplications. 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 in deletion carriers.

Conclusions: Our results suggest a potential mechanism for how "second-hit" variants modulate expressivity of complex disorders such as the 16p12.1 deletion through transcriptomic perturbation of gene networks important for early development. Our work further shows that family-based assessments of transcriptome data are highly relevant towards understanding the genetic mechanisms associated with complex disorders.

Keywords: Copy-number variant, RNA sequencing, complex disorders, whole-genome sequencing, inherited variants, developmental disorders

BACKGROUND

Complex disorders, such as autism, intellectual disability/developmental delay (ID/DD), epilepsy, and schizophrenia, have been attributed to rare copy-number variants (CNVs), or deletions and duplications encompassing multiple genes, as well as individual rare single nucleotide variants (SNVs) and the combined effects of common variants (1–6). 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 and pleiotropy of clinical features often observed in affected individuals (7–9). An example is the 520-kbp deletion at chromosome 16p12.1 (OMIM: 136570), which was originally described in children with developmental delay (10,11) but was subsequently found to confer increased risk for schizophrenia (12), epilepsy (13), and cognitive defects in control populations (14). Unlike syndromic CNVs such as Smith-Magenis syndrome that primarily occur *de novo* (15), the 16p12.1 deletion is inherited in over 90% of affected children from carrier parents who manifest subclinical or mild cognitive and neuropsychiatric features (10,16). In fact, we recently found that children with the deletion were more likely to carry an additional burden of rare CNVs (11) and deleterious variants in genes intolerant to variation (16), 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, and found that several genes adjacent to the deletion region were also downregulated compared to controls (17). Similarly, expression changes due to the autism-associated 16p11.2 deletion correlated with changes in head circumference phenotypes (18) and converged on several neurodevelopmental pathways, including synaptic function and chromatin modification (19). Other studies have used expression data to identify pathogenic variants potentially missed by genome sequencing studies (20–22).

90 *For example*, Frésard and colleagues identified novel causal variants for 6/80 individuals with
91 rare undiagnosed diseases through paired analysis of whole-blood transcriptomes and genomes
92 (20). Additionally, several recent studies have used family-based approaches to study the effects
93 of rare inherited variants towards gene expression. *For example,* Pala and colleagues found that
94 rare inherited variants in both coding and non-coding regions increased the likelihood of gene
95 expression changes among 61 families in the bottlenecked Sardinia population, indicating the
96 importance of such variants towards disease risk (23). While these studies have shown the utility
97 of assessing transcriptomic consequences of individual causal variants, they were focused on
98 either control populations *or* relatively invariable disorders, and did not examine the
99 simultaneous effects of multiple variants with different effect sizes towards changes in gene
100 expression within major biological pathways.

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

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. Among the recruited individuals were 10 children with the deletion ("carrier children"), six sets of carrier and noncarrier parents (including one family with two pairs of 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 (16). We collected phenotypic information from the five families using two standardized clinical questionnaires: one assessing developmental phenotypes in children, and the other assessing psychiatric features in adults. 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. Family members first submitted completed checklists eliciting major phenotypes and medical history, which were then integrated with detailed medical records for each person. A follow-up phone interview was then conducted with family members to fill in any missing information on the clinical questionnaire. Using this method, we assessed clinical data on 31/32 individuals in the cohort. Summarized clinical features for children and adults in this study are listed in **Table S1**. We note that all families had self-reported European or Caucasian ancestry. Based on the curated phenotypic data, we calculated quantitative scores for children using a modified de Vries scoring rubric, as described previously (16), which represents the diversity and severity of phenotypic features in affected children (24). We similarly summed the number of neuropsychiatric features to generate phenotypic scores in adults. Phenotypic scores for all individuals in the cohort are listed in **Table S1**.

DNA extraction, whole-genome sequencing, and variant identification

We identified 25 classes of rare deleterious variants from whole-genome sequencing (WGS) and SNP microarray for each of the 32 family members in our cohort. The 25 rare variant classes identified in this study are displayed in **Figure S2** and listed in **Table S2**. Genomic DNA was

150 extracted from peripheral blood using QIAamp DNA Blood Maxi extraction kit (Qiagen, Hilden,
151 Germany) and treated with RNAse. DNA levels were then quantified using Quant-iT™
152 PicoGreen™ dsDNA assay methods (Thermo Fisher Scientific, Waltham, MA, USA), and
153 sample integrity was assessed in agarose gel. After constructing Illumina TruSeq DNA PCR-free
154 libraries (San Diego, CA, USA), whole-genome sequencing was performed on each sample by
155 Macrogen Labs (Rockville, MD, USA) using an Illumina HiSeq X sequencer to obtain an
156 average coverage of 34.5X. Raw sequencing data were processed for quality control using
157 Trimmomatic (25) with leading:5, trailing:5, and slidingwindow:4:20 parameters, aligned to the
158 human hg19 reference genome using BWAv.0.7.13 (26), and sorted and indexed using Samtools
159 v.1.9 (27).
160 The GATK Best Practices pipeline v.3.8 (HaplotypeCaller) and v.4.0.11
161 (GenotypeGVCFs) (28) was used to identify SNVs and small indels from WGS data. In short,
162 duplicate reads were marked and removed using PicardTools, and after calibration of base-pair
163 quality scores, GATK HaplotypeCaller was used to identify variants in each sample. Variant
164 calls were then pooled for joint genotyping and calibration of variant quality scores. Custom-
165 built pipelines using Annovar v.2016Feb01 (29) applied a total of 430 annotation classes for
166 variant function, population frequency, conservation, genomic region, and predicted
167 pathogenicity. Variants were filtered based on the following quality metrics (30): QUAL >50,
168 read depth >8, allele balance between 0.25-0.75 (or >0.9 for homozygous variants), and quality
169 depth (QUAL/reads with alternate allele) <1.5. Rare variants were defined as variants with
170 frequency ≤0.001 in the gnomAD v.2.1.1 genome database (31), and present in <10 samples in
171 our in-house WGS cohort of 125 families (335 individuals) with rare CNVs, in order to remove
172 technical artifacts that may be missed by gnomAD. We finally classified rare SNVs and small
173 indels for downstream analysis as follows: Rare missense and loss-of-function (LOF, including
174 frameshift and stopgain) variants within protein coding regions, as well as variants in the 5' and
175 3' untranslated region (UTR) or within 1 kbp of the transcription start (TSS) or end sites
176 (upstream and downstream), were classified based on their RefSeq-defined genomic locations in
177 Annovar (**Figure S2**). Splice-site variants were identified based on MutationTaster annotations
178 (32) for disease-causing ("D") or disease-causing automatic ("A") variants. Rare non-coding
179 regulatory variants within 50 kbp of TSS for protein-coding genes were classified according to
180 ChromHMM chromatin state segmentation data for GM12878 lymphoblastoid cells (33),

181 available from the ENCODE Project, into promoters (chromosome states 1-3), enhancers (states
182 4-7), or silencers (state 12). With the exception of loss-of-function and splice-site variants, all
183 coding and non-coding variants were filtered for CADD Phred-like pathogenicity scores ≥10
184 (34). Inheritance patterns of these variants were determined using in-house pipelines.
185 Copy-number variants and structural variants were identified using a combination of
186 WGS data for all samples and SNP microarray data for 25/32 samples, previously described in
187 (16). To identify variants from microarray data, Illumina Omni 2.5 BeadChip microarray
188 experiments were performed for each sample at either the HudsonAlpha Institute for
189 Biotechnology (Huntsville, AL, USA; n=18), Yale Center for Genome Analysis (New Haven,
190 CT, USA; n=5), or the Department of Genome Sciences at the University of Washington
191 (Seattle, WA, USA; n=2). CNV calls for each sample were generated using PennCNV v.1.0.3
192 (35), and were filtered for ≥50 kbp in length and ≥5 target probes. CNVs and SVs were also
193 detected from aligned WGS data using a combination of four pipelines: CNVNator v.0.4.1 (36)
194 (bin size of 200), DELLYv.0.8.2 (37), LUMPY-sv v.0.2.13 with Smoove v.0.2.5 (38), and Manta
195 v.1.6.0 (39). In both WGS and microarray-derived datasets, adjacent CNVs were merged if they
196 overlapped or had a gap <20% of CNV length and <50 kbp. We then integrated the CNV and SV
197 calls from each of the datasets as follows: For smaller CNVs and SVs <50 kbp, any duplication
198 or deletion called by at least two of the four WGS-based callers were considered for downstream
199 analysis, with the minimum intersected regions defining the new breakpoints. For larger CNVs
200 and SVs >50 kbp, the union of CNVNator and PennCNV calls were considered for downstream
201 analysis. Integrated calls were based on 50% reciprocal overlap among the callers. As our SV
202 call-set had a low overlap with SV call-sets from control populations, likely due to different SV
203 calling methods used in the control cohorts (40,41), integrated variants were filtered for presence
204 in <10 individuals in our in-house WGS cohort, as determined by 50% reciprocal overlap.
205 Finally, RefSeq gene-coding regions spanned by SVs were categorized as follows: encapsulating
206 variants which span the entire gene, interstitial variants that are contained within a gene, and 5'
207 and 3' UTR variants that overhang the gene on either end (**Figure S2**). Inheritance patterns of
208 CNV and SV calls were determined if calls in the child and parent had >50% reciprocal overlap.
209 Short tandem repeats (STRs) were called from aligned WGS data with GangSTR v.2.4
210 (42), using the GangSTR hg19 reference file v.13.1. The calls were filtered and analyzed using
211 three tools from the STR analysis toolkit TRTools (43). First, dumpSTR was used to filter for

212 quality of calls using the following parameters: read depth >20, read depth <1000, QUAL >0.9,
213 spanbound only (calls that are spanned by reads), and filter bad confidence intervals (filtered
214 calls whose maximum likelihood estimates were not within the confidence interval). The reads
215 were then merged with mergeSTR, and basic statistics were calculated using statSTR. In addition
216 to the dumpSTR filters, we applied the following filters to our call set: >95% of samples called
217 for the STR location, variance at location >0, and overlap of the STR location with a RefSeq
218 defined protein-coding gene. STR expansions were defined as any call for which the deviation of
219 the repeat length was greater than the mean length plus three standard deviations among all
220 individuals in our cohort. Finally, STR variants were categorized according to their genomic
221 location, including exonic, intronic, 5' or 3' UTR, upstream, and downstream (**Figure S2**).
222 Inheritance patterns of STRs were determined by matching the number of repeats in the child to
223 their parents.
224

225 **Generation of lymphoblastoid cell lines and RNA-sequencing**

226 Peripheral blood samples for all 32 individuals in our cohort were submitted to the Coriell
227 Institute for Medical Research (Camden, NJ, USA) for generation of lymphoblastoid cell lines
228 through Epstein-Barr virus transformation of B-lymphocytes (**Table S3**). After receiving the
229 LCL samples, cells were grown at 5% CO_2 and a concentration of 1X 106 cells/mL under L-
230 glutamine-supplemented RPMI 1640 medium (11875-119, Thermo Fisher Scientific) containing
231 15% fetal bovine serum (35-010-CV, Corning Life Sciences, Tewksbury, MA, USA) and Cytiva
232 HyClone™ Penicillin Streptomycin solution (SV30010, Thermo Fisher Scientific). Total RNA
233 was isolated from three biological replicates of P6-P7 cells per sample using TRIzol Reagent
234 (Thermo Fisher Scientific) and PureLink RNA Mini Kit (12183018A, Thermo Fisher Scientific),
235 and subsequently treated with DNA-free DNA Removal Kit (AM1906, Thermo Fisher
236 Scientific). RNA integrity number scores (RIN) were assessed using Agilent Bioanalyzer 2100
237 (**Figure S3A**), and replicates with RIN scores >8.5 were sequenced. Paired-end 50 bp libraries
238 for each replicate were generated using Illumina TruSeq Stranded mRNA kit, and were
239 sequenced using Illumina NovaSeq at the Penn State College of Medicine Genome Sciences
240 Facility (Hershey, PA, USA). Two sequencing runs of 48 replicates were performed, with the
241 biological replicates of each sample split among the two runs to mitigate batch effects, to
242 generate a total of 43.5 million reads/replicate.

243

244 **Quantification of gene expression and coverage of disease genes**

245 Sequenced RNA reads were filtered using Trimmomatic v.0.36 (25) to remove reads <30 bp
246 long. Following the GTEx Consortium RNA-Seq pipeline (44), the filtered reads were aligned to
247 the human genome (GENCODE v.19) using STAR v.2.4.2a (45), and sorted and indexed using
248 Samtools v.1.9 (27). Duplicates reads were marked with PicardTools v.2.9.0. We assessed the
249 quality of the aligned reads with transcript integrity scores(46), which moderately correlated
250 ($r=0.38$, $p=1.0\times10^{-4}$, Pearson correlation) with the RIN scores for each sample (**Figure S3A**).
251 Isoform counts for GENCODE 19 genes were quantified using RSEM v.1.2.22 (47). A collapsed
252 gene coordinate GTF file was generated using the GENCODE 19 gene coordinates and the
253 GTEX collapse_annotation script. Gene-level counts and transcripts per million read (TPM)
254 values were quantified using RNASeQC v.1.1.8 (48), using strict mode and the collapsed gene
255 coordinates.

256 After filtering for transcripts where all three replicates of at least one sample showed >0.2
257 TPM, we obtained a total of 24,340 expressed transcripts across our cohort, representing 43.3%
258 of all GENCODE transcripts. We further compared our set of expressed LCL genes to disease
259 gene databases (49–53) and genes expressed in the adult brain from GTEx consortium RNA-Seq
260 data (44). We defined expressed genes in GTEx tissues if they showed >0.5 TPM in 80% of
261 samples for a particular tissue. The expressed LCL genes covered >70% of each of these gene
262 sets, including 83% of genes expressed in GTEx brain tissues (**Figure S4A-B**). These data are in
263 concordance with gene expression data from GTEx, where gene expression values in LCLs and
264 brain tissues showed an average Spearman correlation of 0.84 (**Figure S4C**). These findings
265 indicate that our LCL data would be able to identify changes in expression patterns for most
266 genes related to neurodevelopmental disease.

267

268 **Differential expression and outlier expression analysis**

269 We performed differential expression analysis between all 16p12.1 deletion carriers and
270 noncarriers, as well as between parents and offspring across the five families, using edgeR(54)
271 v.3.30.0 on gene-level counts to create generalized linear models and perform quasi-likelihood
272 F-tests with Benjamini-Hochberg correction. For testing differences between all deletion carriers
273 (n=19) and noncarriers (n=13), we included family as a covariate in the linear model, used

274 default filtering for low-expressed genes, and removed genes with sex-specific differences in
275 GTEx LCL samples as well as genes on the X and Y chromosomes (due to unequal sex ratios in
276 deletion carriers and noncarriers). To control for expression outliers, we iteratively identified sets
277 of differentially expressed genes, defined using an FDR<0.05 threshold (Benjamini-Hochberg
278 correction), between deletion carriers and noncarriers after removing one sample at a time. We
279 then took the intersection of the resulting 32 sets of differentially expressed genes, and obtained
280 a total of 1,569 transcripts differentially expressed in the deletion carriers (**Table S4**). We also
281 performed differential expression analysis using PQLseq v.1.2 to account for gene expression
282 similarity due to relatedness (55). We first generated input files from unfiltered WGS SNV data
283 using PLINK v.1.9 (56), and used GEMMA v.0.98.3, which calculates kinship between two
284 individuals based on genotype similarity (57), to generate a kinship matrix for our cohort. This
285 matrix was used as input for PQLseq along with gene-level counts from RNA-Seq data, after
286 removing the same sex-specific genes as for the edgeR analysis.
287 We next performed family-based analysis on 13 separate trios identified across the five
288 families (nine carrier children compared to parents and four carrier parents compared to
289 grandparents), which are listed in **Table S1**. For example, we separately analyzed two trios in
290 family GL_001 (**Figure S1**). For comparison, we analyzed an additional four trios with
291 noncarrier children compared to carrier and noncarrier parents (**Figure S5C**). For each trio, we
292 first used an edgeR workflow without covariates to identify differentially expressed genes
293 between the offspring and carrier parent (|logFC|>0.5, FDR<0.05, Benjamini-Hochberg
294 correction), and separately assessed expression changes between the offspring and noncarrier
295 parent. Genes with low expression (expressed in <25% of all replicates) and sex-specific genes
296 were removed from edgeR analysis. We then overlapped the two sets of differentially expressed
297 genes to classify expression changes by family-specific patterns as follows: "unique" if the gene
298 was differentially expressed in the offspring compared with both parents; "shared with the carrier
299 parent" if the gene was only differentially expressed compared with the noncarrier parent; and
300 "shared with the noncarrier parent" if the gene was only differentially expressed compared with
301 the carrier parent (**Figure S5A**).
302 To identify genes with outlier expression in our cohort, we calculated z-scores of gene
303 expression values for each individual for 14,212 protein coding genes expressed in the LCL
304 samples. We normalized the expression values in each person by calculating the median TPM

305 expression across the three replicates for each gene, transformed the values using $\log_2(x + 1)$,
306 and calculated z-scores for each log-transformed TPM compared with all samples in our cohort.
307 As principal component analysis showed clustering of samples by family (**Figure S6A**), we used
308 PEER v.1.0 (58) to correct the z-scores using one PEER principal component (**Figure S6B**).
309 After correction, we further assessed for clustering of samples and replicates, and found strong
310 Spearman correlations among replicates derived from the same sample (**Figure S3B**). We
311 defined outlier genes as any gene with |z-score| >2 (**Figure S6C**), in line with recent studies
312 utilizing outlier expression values (20).
313

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

315 Enrichment analysis for sets of differentially expressed genes was performed using goseq v.3.12
316 (59), which tests for overrepresentation of gene categories in RNA-seq data. Goseq controls for
317 selection bias in RNA-seq datasets by modeling the distribution of transcript lengths of
318 differentially expressed genes. We assessed for enriched biological processes using the Gene
319 Ontology database (60), as well as genes expressed in specific adult brain tissues from GTEx
320 (44) and developing brain tissues from the BrainSpan Atlas (61). We defined preferentially-
321 expressed genes in GTEx and BrainSpan tissues as expression >2 standard deviations higher than
322 the median expression across all tissues for that gene. We further assessed for enrichment of
323 differentially expressed gene sets for candidate neurodevelopmental disease genes (DBD Gene
324 Database) (49), as well as specific gene sets for autism (SFARI Gene database) (50), intellectual
325 disability (DDD and DDG2P databases) (52,53), and schizophrenia (51). Finally, we annotated
326 sets of genes with altered expression for two common measures of intolerance to variation, RVIS
327 (62) and pLI (63), and used genes considered to be intolerant to variation (RVIS <20th percentile
328 or pLI score >0.9) for downstream analysis. All gene sets used for enrichment analyses were
329 filtered for genes with transcripts that are expressed in our LCL samples (>0.2 TPM in all three
330 replicates of at least one sample).
331

332 **PAGE and WGCNA analysis in deletion carriers and noncarriers**

333 We performed parametric analysis of gene set enrichment (PAGE) on genes that were
334 differentially expressed between carriers and non-carriers of the 16p12.1 deletion (64). PAGE is
335 a gene set analysis method that considers the direction of the expression log fold change to

336 discover sets of genes that are enriched among up- or down-regulated genes. For this analysis,
337 we included the log-fold change of 26,861 transcripts that were not filtered out by edgeR's
338 default filtering of low-expressed transcripts. We searched for significant up- or down-regulation
339 of genes within terms from the Gene Ontology database (60), using two-tailed z-tests with
340 Benjamini-Hochberg correction.

341 We further performed weighted gene correlation network analysis (WGCNA) to identify
342 modules of genes that were co-expressed among samples in our cohort (65). We used the R
343 package tximport (66) to import RSEM-derived gene expression counts, filtered genes for >10
344 counts/replicate in at least one sample, and used DESeq2 (67) to generate variance-stabilized
345 expression counts for each gene. To detect co-expression patterns specific to deletion carriers,
346 we used ComBat (68) within sva v.3.12 to perform batch correction with family as a covariate.
347 We detected 35 co-expression modules in our samples using WGCNA v.1.69 (65), with the
348 following parameters: power threshold = 8, signed hybrid network, unsigned topological overlap
349 matrix, bi-weight mid-correlation, module size = 30-30000, and merge cut height = 0.25. Two
350 modules showed strong sex-specific gene expression and were excluded from further analysis.
351 The average gene expression values in each module were compared between carriers and
352 noncarriers using two-tailed t-tests, and genes in each of these modules were tested for
353 enrichment of Gene Ontology terms using goseq.
354

355 **Integration of gene expression and genomic variant data**

356 We calculated the effect size of different classes of rare "second-hit" variants towards gene
357 expression changes, stratified by sample type and family-specific patterns. We compared 25
358 classes of rare variants identified from WGS data (**Figure S2; Table S2**) towards differentially
359 expressed genes in family trios as well as outlier expression genes in all individuals. For all
360 comparisons, we calculated odds ratios and 95% confidence intervals for each variant class
361 towards changes in expression using Fisher's exact and Wald tests, respectively; uncorrected p-
362 values and Benjamini-Hochberg corrected FDR values were reported for each comparison
363 (**Table S5**). We note that we considered each variant class independently, so that dysregulated
364 genes with multiple types of disrupting variants were counted within multiple variant classes. For
365 the differential expression analysis, we first assessed variants in the 13 trios with carrier
366 offspring for genes with differential expression (**Table S1**), and then determined the effects of

367 variants in carrier children (n=9 trios) inherited from carrier or noncarrier parents towards
368 expression changes shared with the same parent. For the outlier expression analysis, we first
369 assessed variants for outlier expression genes in all individuals. We then stratified these data by
370 sample type (carrier child, carrier parent, and noncarrier parent), and compared variants in carrier
371 children that were inherited from their carrier or noncarrier parent. To identify synergistic effects
372 between the 16p12.1 deletion and "second-hit" variants, we identified a subset of genes with
373 outlier expression in deletion carriers that were also differentially expressed in the global
374 comparison of carriers and noncarriers, and then identified those genes which also had "second-
375 hit" variants inherited from the noncarrier parent.
376

377 **Alternative splicing analysis**

378 To assess alternative splicing events from RNA-sequencing data, we used DESeq2 (67) to detect
379 differential expression of isoforms. After importing isoform-level expression counts from RSEM
380 using tximport (66), we filtered for genes with >2 counts across all samples, and performed
381 pairwise comparisons between carrier offspring and their parents in the 13 trios listed in **Table**
382 **S1**, plus the four trios with noncarrier children for comparisons. We then repeated the DESeq2
383 analysis for gene expression counts, and only included differentially expressed isoforms within
384 genes that did not show overall differential expression, to specifically account for isoform
385 changes due to alternative splicing. Similar to the family-based differential expression analysis,
386 we assigned family-specific patterns to each alternative splicing event observed in offspring
387 based on the pairwise comparisons to each parent. We further compared alternate isoforms
388 identified by DESeq2 to those in GTEx LCL data (44) to identify unique isoforms in our cohort.
389 Finally, we integrated these data with 12 classes of putative splice-site disrupting variants
390 identified from WGS data, and calculated odds ratios as described above.
391

392 **Allele-specific expression analysis**

393 We used the phASER v.1.1.1 (69) pipeline to identify allele-specific expression events in our
394 cohort. We first used whatshap v.0.18 (70) to perform read-backed and pedigree-informed
395 phasing of our WGS samples, and then merged the three replicate BAM files of aligned RNA-
396 Seq reads for each sample together using Samtools, We then used phASER, which uses phased
397 WGS data to infer phasing of RNA-Seq samples, to phase the RNA-Seq alignments and to count

398 the number of reads per haplotype block. We ran phASER with the parameters --mapq 255 and -
399 -baseq 10, and used the recommended blacklist to remove HLA genes. Finally, we quantified
400 log-fold changes for allelic counts in each protein-coding gene with >10 read counts using
401 phASER Gene AE, and identified ASE for genes with FDR>0.05 using binomial tests with
402 Benjamini-Hochberg correction. For each identified ASE event, we examined the over-expressed
403 haplotype for presence of a deleterious rare coding variant identified from WGS, which would
404 potentially indicate pathogenic effects of the ASE event. Finally, we determined family-specific
405 patterns of ASE genes based on the presence of ASE in parents of offspring.
406

407 **eQTL discovery and analysis**

408 We used QTLTools v.1.2 (71) to identify eQTLs in our cohort. Because we had three replicates
409 per participant, we first calculated the median TPM values for all transcripts in an individual.
410 Genes were filtered for >0.1 median TPM in more than 50% of samples. Principal components
411 for gene expression (from RNA-Seq data) and genotype (from whatshap-phased WGS data) were
412 then computed using QTLtools. The top three genotype and the top two gene expression
413 principal components were used as covariates for the linear model, in addition to three explicit
414 covariates (family, sex, and carrier status). QTLtools cis-permutation tests (n=1000 replicates)
415 were then used to discover eGenes, or genes whose expression are significantly correlated with
416 eQTLs, and associated variants in our samples. We performed multiple testing correction with
417 the QTLtools script runFDR_cis.R. Finally, we annotated significant eQTL variants (FDR<0.05)
418 associated with protein-coding genes for presence in GTEx LCL data, genomic location,
419 population frequency, and biological functions using the WGS Annovar-based pipeline (29).
420

421 **Brain-specific network analysis**

422 We assessed the connectivity patterns of genes with "second-hit" variants and changes in
423 expression in the context of a brain-specific interaction network. The network contains brain-
424 specific pairwise interactions for 14,763 genes expressed in the brain, of which 11,978 (81.1%)
425 are also expressed in the LCL samples. This network was previously built using a Bayesian
426 classifier trained on hundreds of gene co-expression, protein-protein interaction, and regulatory-
427 sequence datasets, in order to predict the likelihood of interactions between any two pairs of
428 brain-expressed genes (72,73). To create a network containing only the highest probability

429 predicted gene interactions, we extracted all pairs of genes with weighted probabilities >2.0,
430 representing the top ~0.5% of pairwise interactions (217,975,718 pairs of genes). We then
431 calculated the weighted shortest path lengths for all pairs of genes in the network, using the
432 inverse of the probabilities as weights for each edge. Finally, we created sub-networks that
433 contained genes with "second-hit" protein-coding variants (loss-of-function or LOF, missense,
434 splice-site, exonic STR, or encapsulated deletion or duplication) or expression changes
435 (differential expression, outlier expression, alternative splicing, ASE, or eQTL minor allele) for
436 each carrier offspring from the 13 trios (**Table S1**). For each trio, we calculated the average
437 shortest paths between all pairs of genes with expression changes and genes with "second-hit"
438 coding variants, and then compared these distances to average shortest paths calculated from 100
439 permuted network replicates, where genes were randomly reassigned to different nodes in
440 networks with otherwise identical topology. Network analysis was performed using the
441 NetworkX package in Python (74).
442
443 **Statistical analysis**
444 All genomic and statistical analyses were conducted using either Python v.3.7.3, with packages
445 numpy v.1.16.2 (75), scipy v.1.1.0 (76), and pandas v.1.0.0 (77), or using R v.3.5.1 (R
446 Foundation for Statistical Computing, Vienna, Austria). Details of all statistical tests, including
447 summary statistics, test statistics, odds ratios, confidence intervals, p-values, and Benjamini-
448 Hochberg corrected FDR values, are provided in **Table S5**.

449 **RESULTS**

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

451 We performed RNA-sequencing on LCL samples from 19 deletion carriers and 13 noncarriers
452 from five large families with multiple affected members (**Figure 1; Figure S1; Table S1**), and
453 identified 1,569 transcripts that were differentially expressed (FDR<0.05) in deletion carriers
454 compared with noncarriers (**Figures 2A-B; Table S4**). Application of additional corrections for
455 relatedness among the samples (55) (see Methods) yielded 1,044 differentially expressed
456 transcripts, of which 840 (80.5%) were also identified in the main analysis (**Figure S7A; Table**
457 **S4**). We first confirmed that each of the seven protein-coding genes in the deletion region were
458 downregulated in deletion carriers (**Figure 2C**). Interestingly, 13 protein-coding genes adjacent
459 to the 16p12.1 region (between chromosomal bands 16p11.2 and 16p12.3) also showed
460 differential gene expression in carriers, 10 of which were under-expressed in the deletion
461 carriers. *For example*, two genes within the autism-associated 16p11.2 region, *SEZ6L2* and
462 *DOC2A*, as well as the febrile seizure-associated gene *STX1B* (78), were downregulated in
463 carriers of the 16p12.1 deletion. As none of the carriers harbored an atypical deletion, it is
464 possible that these adjacent genes could be affected by disruption of regulatory elements located
465 within the deletion region. In fact, three downregulated genes adjacent to the deletion, *DNAH3*,
466 *OTOA*, and *NPIPB4*, exhibited chromatin interactions with enhancer elements within the
467 deletion region, detected using published Hi-C data of LCL samples (79) (**Figure 2C**).

468 We found that differentially expressed genes in deletion carriers were enriched
469 (FDR<0.05) for multiple biological functions, including biological adhesion and cell
470 proliferation regulation for relatedness-corrected genes, and signal transduction and locomotion
471 for genes without relatedness correction (**Table S4**). Additionally, we observed an enrichment
472 (FDR=0.015) for candidate autism genes (50), including *FOXP1*, *CUL7*, *ANK3*, and *EP300*,
473 among the differentially-expressed genes (**Figure 2B; Table S4**). Parametric Analysis of Gene
474 Set Expression (PAGE) showed that genes related to neuronal and muscular growth functions
475 were significantly upregulated in deletion carriers (FDR<0.05), while genes involved in
476 behavioral responses and learning were downregulated (**Figure S7B; Table S6**). Weighted-gene
477 correlation network analysis similarly identified several modules of genes with significant
478 expression changes in deletion carriers (p<0.05, two-tailed t-test), including downregulated
479 genes enriched for cell signaling and adhesion, and upregulated genes enriched for neurogenesis,

480 nervous system development, and MAPK and Notch signaling (**Figure S8; Table S7**).
481 Differentially expressed genes in deletion carriers were further enriched (FDR<0.05) for genes
482 preferentially expressed in the hippocampus and basal ganglia of the adult brain (44) (**Figure**
483 **S7C; Table S4**), as well as in the striatum, thalamus, and frontal cortex during late fetal and
484 adolescent/young adulthood timepoints (61), which are critical transition periods for expression
485 of neurodevelopmental genes (80–82) (**Figure 2D; Table S4**). Overall, our data suggest that the
486 16p12.1 deletion leads to pervasive transcriptomic changes across multiple biological and
487 neuronal processes in the developing brain. We note that because these results are based on
488 expression data from LCL samples, they should be followed up in neuronal models to delineate
489 any tissue-specific differences in gene expression.
490

491 **Family-specific patterns of gene expression changes influence developmental phenotypes**

492 We next investigated how gene expression patterns segregated within 13 complete trios with
493 carrier offspring extracted from the five families, including carrier children compared to their
494 parents as well as carrier parents compared with grandparents (**Table S1**). For each trio, we
495 identified differentially expressed genes for offspring-carrier parent and offspring-noncarrier
496 parent pairs (see Methods), and found an average of 5,323 total gene expression changes in
497 offspring compared to their parents (**Tables S8, S9**). We then overlapped the two sets of
498 differentially expressed genes to categorize expression changes based on their family-specific
499 pattern (**Figure S5A**). We found no significant differences (p=0.735, two-tailed paired Mann-
500 Whitney test) in the proportion of differentially expressed genes in offspring that were shared
501 with either the carrier (avg. 2,223 genes/offspring) or noncarrier parent (avg. 1,908
502 genes/offspring; **Figure 3A**). This may suggest that "second-hit" variants from the noncarrier
503 parent could contribute equally to gene expression changes, and therefore to disease
504 pathogenicity, as the combined effects of the deletion and any "second-hit" variants from the
505 carrier parent, an observation that corresponds with our recent findings of increased burden of
506 "second-hits" transmitted to the child from noncarrier parents (16) (**Figure S5B**). However, we
507 also note that this study may be under-powered to detect smaller differences in the proportion of
508 gene expression changes shared between offspring and their carrier and non-carrier parents.
509 Interestingly, we also observed an average of 1,192 genes/offspring that were differentially
510 expressed compared with both parents (**Figure 3A**), such as *SHANK2*, *FOXP1*, and *CACNA1D*.

511 These expression changes potentially represent effects of *de novo* variants or combinatorial
512 effects of variants inherited from both parents, which could explain the increased phenotypic
513 severity observed in the carrier children. However, the trends in expression patterns widely
514 varied across families, which in some cases could be explained by family history of
515 neuropsychiatric disease (**Figure S5C**). For example, we found that children within family
516 GL_004, whose parents were unaffected or presented with mild depression, had the lowest
517 number of gene expression changes among any carrier children in the cohort. Meanwhile,
518 children in families GL_001 and GL_052, whose carrier parents manifested multiple overt
519 cognitive and neuropsychiatric features, had higher proportions of expression changes shared
520 with their carrier parents compared to their noncarrier parents (**Figure S5C**).

521 We next assessed the dysregulated biological functions in each trio (**Figure S9; Table**
522 **S9**), and found that unique or shared differentially expressed genes in carrier children were
523 enriched for biological processes (FDR<0.05) that could be related to 33 out of 41 (80.5%)
524 developmental phenotypes observed in the affected children (**Figure 3B**). For example, shared
525 gene expression changes for carrier child P1C_01 in family GL_001 were enriched for nervous
526 system development, neurotransmitter metabolism, neuron projection, and synaptic transmission
527 functions, while their unique expression changes were enriched for genes involved in skeletal
528 and muscular development. The shared changes in neuronal genes could contribute to the ID/DD
529 and speech delay phenotypes observed in the child, as both parents also had several psychiatric
530 features, while the unique changes in developmental genes could be related to hypotonia, growth
531 delay, and craniofacial features uniquely observed in the child. Overall, these results suggest that
532 expression changes of neurodevelopmental-related genes could account for phenotypic
533 differences among carriers of the 16p12.1 deletion.

534

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

537 We next investigated whether changes in gene expression could be attributed to "second-hits", or
538 rare genetic modifiers elsewhere in the genome. Rare variants disrupting protein-coding regions
539 and nearby regulatory elements have been previously linked to gene expression changes in both
540 control populations (23,83–85) and disease cohorts, where causal genes may be missed by DNA
541 sequencing methods (18,20–22). We hypothesized that "second-hits" by themselves or in

542 combination with the deletion could contribute to the observed gene expression changes in
543 affected children. We therefore identified 25 classes of rare gene-disruptive "second-hit" variants
544 from WGS data for each individual (**Figure S2; Tables S9, S10**), including SNVs and indels in
545 coding and non-coding regulatory regions (UTRs, introns, and putative promoter, enhancer, and
546 silencer elements within 50 kbp of a gene) with Phred-like CADD scores>10 (34), and CNVs
547 and short tandem repeats (STRs) that spanned gene-coding regions. We then calculated the
548 likelihood that these "second-hit" variants are associated with changes in expression of a
549 proximal gene, as determined by either differential expression analysis between carrier offspring
550 and their parents in the 13 trios, or outlier expression analysis among all individuals in the cohort
551 (18,83) (see Methods). While family-based differential expression analysis detects all expression
552 changes between affected children and their parents (86), including those due to the downstream
553 effects of the deletion, outlier analysis more robustly identifies specific effects of "second-hits"
554 towards larger changes in expression, including synergistic effects in combination with the
555 deletion. Overall, we observed an average of 285 outlier genes (|z-score| >2) per individual,
556 including candidate neurodevelopmental genes (49) such as *CTNNB1*, *FOXG1*, *DISC1*, and
557 *ZNF804A* (**Figure S6; Tables S8, S11**). We found that 10.8% of outlier genes (avg. 31/286 per
558 person) and 11.2% of differentially expressed genes (avg. 310/2,774 per carrier offspring) were
559 potentially disrupted by a rare coding or non-coding variant (**Table S10**). Altered expression of
560 genes without such variants could be due to several factors, such as common variants, DNA
561 methylation events, downstream effects of other dysregulated genes, or environmental factors.

562 In agreement with previous studies (23,83), we found that genes with outlier expression
563 were significantly enriched after Benjamini-Hochberg correction (FDR<0.05, Fisher's exact test)
564 for 5/25 classes of rare variants that directly affected gene-coding regions, including loss-of-
565 function (LOF), missense, and splice-site SNVs, and 5' UTR overhanging and gene-
566 encapsulating duplications (**Figure 4A**). Similarly, we found that 10/25 variant classes were
567 significantly associated with differentially expressed genes in carrier offspring for the 13 trios
568 (FDR<0.05, Fisher's exact test), including coding missense SNVs, duplications overhanging the
569 5' UTR, and encapsulated deletions (**Figure S10A**). We further found that outlier genes had
570 higher burdens of rare variants in aggregate ($p=1.01\times10^{-3}$, one-tailed t-test) and for 7/25
571 individual classes compared with non-outlier genes ($p<0.05$, one-tailed t-test), in particular loss-
572 of-function variants ($FDR=1.73\times10^{-3}$) and encapsulated duplications ($FDR=5.43\times10^{-3}$), which

passed multiple-testing correction (**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 enrichments of LOF variants ($p=4.96\times10^{-6}$; $FDR=6.20\times10^{-5}$), 5' UTR overhanging duplications (p=0.017; FDR=0.108), and 5' UTR-disrupting SNVs ($p=7.37\times10^{-3}$; FDR=0.058) towards outlier expression in carrier children but not in carrier parents (**Figure 4B**). We observed similar findings among differentially expressed genes, where missense SNVs ($FDR= 4.23\times10^{-5}$), upstream SNVs ($FDR=7.00\times10^{-4}$), encapsulated (FDR=0.039) and interstitial (FDR=0.031) deletions, and 5' UTR overhanging duplications (FDR=0.045) were only likely to alter gene expression in carrier children (**Figure S10B**). Notably, LOF variants ($p=1.67\times10^{-5}$, $FDR=2.09\times10^{-4}$) and 5' UTR overhanging duplications (p=0.018; FDR=0.090) 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 several classes of "second-hit" variants in carrier children, including 5' UTR overhanging duplications ($p=1.54\times10^{-3}$, FDR=0.025), gene-encapsulating (p=0.025; FDR=0.125) and 3' UTR overhanging deletions (p=0.025; FDR=0.125), and missense (p=0.023; FDR=0.125) and upstream SNVs (p=0.032; FDR=0.133), that were enriched for outlier expression when inherited from the noncarrier parent but not from the carrier parent (**Figure 4C**). Similarly, we found that LOF variants ($FDR=5.05\times10^{-5}$) and SNVs in upstream ($FDR=6.25\times10^{-3}$) and silencer regions ($FDR=7.70\times10^{-4}$) correlated with differential gene expression in carrier children only when inherited from the noncarrier parent (**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*, associated with motor dysfunction (87), 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

604 genes, such as *ANK3, DOCK10,* and *SLC26A1*, that were differentially expressed in all deletion
605 carriers, showed outlier expression in an individual deletion carrier, and had a nearby variant
606 (two coding and nine non-coding) inherited from the noncarrier parent (**Figure 4E; Figure S12;**
607 **Table S12**). For example, two carrier children in family GL_005 inherited an intronic variant
608 within *ZEB2* from their noncarrier parent, whose altered dosage is associated with Mowat-
609 Wilson syndrome (88,89). While both the 16p12.1 deletion and the "second-hit" variant
610 individually corresponded with increased *ZEB2* expression, the presence of both variants in the
611 carrier children resulted in even stronger overexpression of the gene compared to those with
612 either individual variant (**Figure 4E**). Overexpression of *ZEB2* could contribute to the Mowat-
613 Wilson like features observed in the carrier child P1C_05, including ID/DD, seizures, hypotonia,
614 and digestive abnormalities. Similarly, a carrier child in GL_005 inherited a rare variant in a
615 promoter region upstream of the mRNA splicing-associated (90) gene *SNRNP27* from their
616 noncarrier parent. *SNRNP27* is over-expressed in deletion carriers but under-expressed in both
617 the carrier child and the noncarrier parent, highlighting a case where a "second-hit" variant
618 reverses an expression change caused by the deletion (**Figure 4E**). Furthermore, a carrier child in
619 GL_004 shared an intronic variant with two noncarrier relatives in the gene *SYNJ1*, which is
620 associated with synaptic transmission (91) and is under-expressed in carriers of the deletion.
621 While other individuals with the same variant had normal *SYNJ1* expression, the carrier child
622 exhibited under-expression of the gene compared to both carriers and noncarriers of the deletion,
623 suggesting that the variant may alter *SYNJ1* expression only in the presence of the deletion
624 (**Figure 4E**). While it is possible that other variants elsewhere in the genome could also
625 influence expression levels of these genes, these examples highlight putative synergistic effects
626 between the 16p12.1 deletion and "second-hit" variants towards gene expression, where the
627 "second-hit" variants may amplify or reduce the effects of the CNV.
628

629 **A broad range of transcriptomic alterations contribute to phenotypic variability of the**
630 **16p12.1 deletion**

631 To identify a complete spectrum of gene expression alterations in each individual, we next
632 evaluated alternative gene splicing, allele-specific expression (ASE), and expression quantitative
633 trait loci (eQTL) among individuals in our cohort. We first identified an average of 3,267
634 alternative isoforms present in carrier offspring of the 13 trios compared to their parents (**Tables**

635 **S8, S13**), including for several neurodevelopmental-associated genes (49) such as *KANSL1*,
636 *SHANK2*, and *SYNGAP1*. After categorizing splicing events by family-specific patterns, we
637 found no differences between splicing events in offspring shared with carrier (average=1,307) or
638 noncarrier parents (average=1,392; p=0.635, two-tailed paired Mann-Whitney test), with fewer
639 unique changes in the offspring (average=568; $p=2.44\times10^{-4}$; **Figure S13A**). We next found
640 enrichments for alternative splicing in genes disrupted by "second-hit" splice-site ($p=7.47\times10^{-4}$,
641 Fisher's exact test; $FDR=2.99\times10^{-3}$), intronic ($p=9.91\times10^{-9}$; $FDR=5.95\times10^{-8}$), or missense SNVs
642 (p=0.012; FDR=0.036), interstitial (p=0.043; FDR=0.086) and 3' UTR overhanging deletions
643 (p=0.024; FDR=0.058), and intronic STRs ($p=5.75\times10^{-10}$; $FDR=6.90\times10^{-9}$) (**Figure S13B;**
644 **Table S10**). We also found that intronic SNVs were more likely to disrupt splicing in carrier
645 children if they were inherited from the noncarrier parent (p=0.034, FDR=0.204) than the carrier
646 parent (**Figure 5A**), while intronic SNVs ($FDR=6.36\times10^{-7}$) and interstitial deletions
647 (FDR=0.018) were more likely to lead to alternative splicing when present in carrier children
648 than in carrier parents (**Figure S13C**). These results suggest potential correlations between
649 classes of inherited rare variants and alternative splicing events, although changes in isoform
650 expression can only be confidently attributed to rare variants at or near the splice-site. For
651 example, a deleterious splice-site variant in the transcriptional activator *TADA2A* led to an
652 alternate isoform (*TADA2A*-003) in multiple family members of GL_007 that was not observed
653 in GTEx LCL data (**Figure 5B**). *TADA2A* is a candidate gene within the schizophrenia-
654 associated 17q12 deletion (92), and five out of six family members with the splicing variant have
655 schizophrenia-like clinical features (i.e. hallucinations or delusions), including four deletion
656 carriers and one noncarrier child.

657 Next, we identified an average of 285 genes with ASE per individual in our cohort
658 (**Tables S8, S14**), including for the neurodevelopmental-associated genes (49) *DNMT3A*,
659 *NSUN2*, and *HDAC8*. ASE events in the 13 trios were more likely to uniquely occur in the
660 offspring than be shared with a parent ($p=2.44\times10^{-4}$, two-tailed paired Mann-Whitney test), in
661 contrast to differential expression and alternative splicing events (**Figure S14A**). Genes with
662 ASE have previously been shown to have a higher burden of nearby rare deleterious variants
663 (83), and the pathogenicity of a gene with ASE increases with the presence of a deleterious
664 variant on the overexpressed allele (85). In our cohort, we found five ASE events in carrier
665 children that led to overexpression of a deleterious "second-hit" coding variant (**Figure S14B**).

666 For example, two carrier children with autism in family GL_007 showed overexpression of a
667 deleterious "second-hit" missense variant in the candidate autism gene (93) *CARD11,* which was
668 inherited from their noncarrier parent (**Figure 5C**).

669 We further performed eQTL discovery analysis to identify variants statistically correlated
670 with expression changes in our cohort, agnostic to variant pathogenicity or population frequency.
671 We identified 21 eQTLs which modulated the expression of 23 eGenes, or genes whose
672 expression is significantly correlated with an eQTL (**Figure S15A; Table S15**). Interestingly,
673 19/21 identified eQTLs were not present in GTEx LCL data, representing unique loci in our
674 cohort. Carrier children showed a trend (p=0.107, two-tailed Mann-Whitney test) towards
675 carrying a higher number of minor eQTLs alleles (average=4.3/person) than their carrier parents
676 (average=3.2/person) (**Figure S15B; Table S8**). Furthermore, several eGenes had biological
677 functions related to neuronal processes (94–96) (**Table S15**), including *SERPINF1, BEGAIN,*
678 and *ARFGEF2*. For example, we identified a relatively rare eQTL (allele frequency = 0.015) for
679 overexpression of *ARHGAP39*, a key regulator of neurogenesis and dendrite morphology
680 associated with learning and memory (97) (**Figure 5D**). The eQTL minor allele, located in a
681 transcription factor binding cluster, was only observed in a carrier child and their noncarrier
682 parent within GL_007, who both presented with neuropsychiatric phenotypes.

683 To assess the joint contributions of each type of expression change among the individuals
684 in our cohort, we assessed correlations between the numbers of gene expression changes
685 assessed in our study by family-specific pattern (**Figure 5E**). We observed three significant
686 positive correlations (FDR<0.05, Pearson correlation) between pairs of gene expression changes
687 in each person, which often shared the same family-specific patterns. Specifically, the number of
688 genes with differential expression strongly correlated with the number of genes with alternative
689 splicing when shared with either the carrier parent (r=0.93, FDR=2.91×10^{-4}) or non-carrier
690 parent (r=0.91, FDR=4.52×10^{-4}), while unique splicing events in the offspring correlated with
691 splicing events shared with the non-carrier parent (r=0.83, FDR=0.011). Together, the
692 correlations between transcriptomic alterations suggest that different types of gene expression
693 changes could co-occur in parents and offspring, potentially due to the same inherited "second-
694 hit" variants disrupting expression of similar genes and biological pathways, as is observed for
695 signals in genome-wide association studies (98).

696

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

699 Finally, to determine whether associations between transcriptomic changes and "second-hit"
700 variants in LCL samples were also relevant in the nervous system, we assessed connectivity
701 patterns of genes with "second-hit" variants and altered gene expression using a brain-specific
702 gene interaction network (72,73). We generated individual networks for carrier offspring in the
703 13 trios, and calculated shortest distances between genes with protein-coding "second-hit"
704 variants and genes with LCL-derived expression changes in each offspring (see Methods). We
705 found that the average shortest distances between genes with "second-hits" and expression
706 changes were significantly smaller in 6/13 offspring than those derived from permuted networks
707 (FDR<0.05, one-tailed z-test with n=100 permutations). In fact, networks for offspring in
708 aggregate had significantly smaller shortest distances ($p=4.88\times10^{-4}$, two-tailed paired Mann-
709 Whitney test) than the shortest distances from the sets of permuted networks, where genes were
710 randomly reassigned to different nodes in the network (**Figure 5F**). These data indicate that
711 "second-hit" variants closely interact with genes with expression changes detected from LCL
712 samples in a brain-specific context, suggesting a potential mechanism for how gene expression
713 changes that underlie developmental phenotypes can be influenced by "second-hit" variants in
714 the genome. However, these findings should be confirmed using expression data from patient-
715 derived neuronal models of the 16p12.1 deletion, as expression changes in LCL samples may not
716 be conserved in the nervous system.

717 **DISCUSSION**

718 We previously described a two-hit model for the 16p12.1 deletion, where the presence of both
719 the deletion and "second-hit" variants determine the phenotypic trajectory of affected children
720 (11,16). Here, we propose a potential mechanism for how the deletion and "second-hits" jointly
721 interact to alter clinical phenotypes by way of the transcriptome. We found that the 16p12.1
722 deletion itself disrupts the expression of genes across the genome through direct effects, such as
723 chromatin interactions, and through indirect effects, such as downstream genetic interactions
724 (**Figure 6**). For example, chromatin interactions were observed between regions within the
725 16p12.1 deletion and flanking genes such as *STX1B* and *DNAH3*, and 1,493 genes outside of
726 chromosome 16 were also dysregulated in deletion carriers. The identification of flanking genes
727 downregulated by the deletion is in line with similar findings for the 16p11.2, 1q21.1, and
728 22q11.2 deletion disorders (19,99). Each of these CNVs exhibited altered gene expression in
729 adjacent regions that is putatively mediated by chromatin interactions, highlighting the
730 importance of considering the three-dimensional structure of the genome to elucidate CNV
731 pathogenicity. Similarly, we found that "second-hits" disrupt gene expression through both direct
732 and indirect mechanisms. Genes with nearby "second-hit" variants were more likely to exhibit
733 outlier expression and alternative splicing, and genes with "second-hits" were more closely
734 connected to genes with expression changes in a brain-specific network than random sets of
735 genes in permuted networks. In fact, we observed 11 examples of combined effects of the
736 deletion and "second-hit" variants towards expression in our cohort, including the candidate
737 neurodevelopmental genes *SYNJ1* and *ZEB2*. These synergistic effects towards gene expression
738 are similar to those previously observed for eQTLs (100) and HLA alleles (101), except that
739 these effects are potentially due to the combined effects of rare deleterious variants. We note that
740 only a subset (~11%) of genes with altered expression in our cohort harbored deleterious
741 "second-hit" variants that could affect expression. It is likely that the downstream effects of both
742 the deletion and "second-hit" variants could be responsible for a larger proportion of gene
743 expression changes, along with other common variants and environmental factors. Thus, our
744 results suggest that the 16p12.1 deletion and the "second-hit" variants interact with each other in
745 a complex manner to mold the shape of the transcriptome, resulting in strong dysregulation of
746 developmental genes and contributing to neuropsychiatric features in 16p12.1 deletion carriers.

747 Results from our study align with previous studies, which found that rare variants of
748 different classes have varying effect sizes towards gene expression (23,83). Our study extends
749 this paradigm by identifying classes of rare "second-hit" variants whose contributions to gene
750 expression changes differ by inheritance pattern. We found that high-effect variants, such as
751 whole-gene duplications, cause expression changes regardless of parent-of-origin, and splice-site
752 variants lead to changes in isoform expression independently of inheritance. In contrast, lower-
753 effect variants, including missense, silencer, and upstream SNVs, were more strongly associated
754 with gene expression changes when inherited from the noncarrier parent than the carrier parent.
755 These findings indicate that noncarrier parents are more likely to pass gene expression-altering
756 "second-hit" variants down to their carrier children, potentially accounting for more severe
757 phenotypic manifestations in children with the deletion compared with their carrier parents
758 (**Figure 6**). One potential explanation for why carrier children receive a higher number of
759 deleterious variants from their non-carrier parent is assortative mating among their parents, as
760 8/8 carrier parents and 7/9 non-carrier parents in our cohort manifested at least mild
761 neuropsychiatric features. Assortative mating has been extensively observed among individuals
762 with neurodevelopmental or psychiatric disorders (102,103), in particular autism (104),
763 suggesting its relevance towards phenotypic variability among deletion carriers on our cohort.
764 Future family-based transcriptome studies with larger sample sizes may be able to pinpoint
765 specific rare variants within dysregulated genes that are associated with distinct phenotypes in
766 the carrier children.

767 We also identified putative biological and developmental pathways disrupted by both the
768 deletion and "second-hit" variants. For example, we found that genes differentially expressed by
769 the deletion were preferentially expressed in multiple brain tissues during development, and were
770 enriched for core signaling and developmental pathways. In fact, knockdown of individual
771 homologs of 16p12.1 genes in *Drosophila melanogaster* models showed neuronal phenotypes
772 and transcriptome disruptions, suggesting that the individual effects of multiple genes in the
773 deletion sensitize the genome for neuropsychiatric phenotypes (105). Interestingly, we found
774 several examples of biological functions and mechanisms that were simultaneously dysregulated
775 by both the deletion and "second-hit" variants in the carrier children. For example, most carrier
776 children shared differential expression in genes enriched for nervous system development, cell
777 adhesion, signaling, and locomotion with both their carrier and noncarrier parents. These results

778 provide insights into how the deletion and "second-hit" variants synergistically dysregulate
779 genes and pathways related to development, ultimately contributing towards a wide range of
780 developmental phenotypes observed in children with the deletion.

781 Some limitations can be noted in the context of our study. *First*, we investigated gene
782 expression changes within patient-derived LCL samples, which have reduced relevance for brain
783 expression. However, over 80% of genes expressed in GTEx brain samples, as well as over 70%
784 of neurodevelopmental disease genes, were expressed in our LCL samples (**Figures S4A-B**).
785 Nevertheless, repeating the study in tissues that are implicated in neurodevelopmental disorders,
786 potentially using patient-derived reprogrammed neuronal progenitor cells, would verify the
787 associations between variants, expression changes, biological functions, and clinical features.
788 *Second*, we have a relatively small cohort of 32 individuals within five families. It would be
789 useful to determine whether the identified associations are strengthened in a larger cohort,
790 especially those that did not pass multiple-testing corrections. Phenotypically more diverse
791 cohorts would also allow for performing additional correlations between gene expression
792 changes and specific clinical features, such as whether more outlier genes are present among
793 families with stronger histories of neuropsychiatric disease.

794

795 **CONCLUSIONS**

796 Overall, our work supports a model for complex disorders, where combinatorial effects of
797 multiple variants with different effect sizes affect expression of genes in developmental
798 pathways, which further influence the expressivity of clinical features. These results exemplify
799 that family-based transcriptome studies, similar to family-based genome studies, can help
800 explain changes in phenotypes from parents to children and between siblings, especially in
801 complex disorders with a high degree of intra- and inter-familial variability.

802

803 **LIST OF ABBREVIATIONS**

804 ASE, allele-specific expression

805 CNV, copy-number variant

806 ID/DD, intellectual disability/developmental delay

807 LCL, lymphoblastoid cell line

808 LOF, loss-of-function

809 PAGE, parametric analysis of gene set enrichment

810 RIN, RNA integrity number

811 SNV, single-nucleotide variant

812 STR, short tandem repeat

813 TPM, transcripts per million

814 TSS, transcription start site

815 UTR, untranslated region

816 WGCNA, weighted gene correlation network analysis

817 WGS, whole genome sequencing

818

DECLARATIONS

Ethics approval and consent to participate: 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).

Consent for publication: Not applicable.

Availability of data and materials: 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). Accession numbers for LCL samples are provided in **Table S3**. Whole genome sequencing, SNP microarray, and RNA-sequencing data generated in this study are available at NCBI dbGaP study accession phs002450 (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002450.v1.p1) and BioProject accession number PRJNA734670. All other data generated or analyzed during this study are included in this article and its supplementary information files. All code generated for this project, including pipelines for running bioinformatic software and custom analysis scripts, are available at https://github.com/girirajanlab/16p12_RNAseq_project and https://github.com/girirajanlab/16p12_WGS_project.

Competing interests: The authors declare that they have no competing interests.

Funding: 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 the design of the study, the collection, analysis, and interpretation of data, or in writing the manuscript.

Authors' contributions: MJ, AT, and SG conceived the project. MJ and AT performed all analyses in the manuscript, generated the plots and images, and wrote and revised the manuscript. LP recruited the families, obtained and assessed clinical phenotypes, and isolated

850 DNA and RNA for sequencing. CS assisted with collection of patient clinical information and
851 bioinformatic pipelines to identify WGS variants. MD assisted with extraction of RNA from
852 LCL samples. EH assisted with collection of patient clinical information. AK provided the brain-
853 specific network and assisted with the network and WGCNA analyses. SG supervised the
854 research, recruited the families, assisted with collection of patient clinical information, and wrote
855 and revised the manuscript. All authors read and approved the final draft of the manuscript.
856

857 **Acknowledgements:** We are grateful to all members of the five families who participated in this
858 study. We also thank Sherryann Wert, Melissa Berkowitz, and Tara Schmidlen (Coriell Institute)
859 for their assistance in the generation of stable LCL lines; Craig Praul and Yuka Imamura (Penn
860 State Genomics Core Facility) for assistance with designing the RNA-sequencing strategy;
861 Casey Brown (UPenn), Istvan Albert (Penn State), and Qunhua Li (Penn State) for assistance
862 with statistical analysis of transcriptome data; and Jesse Gillis (CSHL), Dajiang Liu (Penn State),
863 and members of the Girirajan lab for useful discussions and critical reading of the manuscript.
864 We appreciate access to data from the Genotype-Tissue Expression (GTEx) Project, which was
865 supported by the Common Fund of the Office of the Director of the National Institutes of Health,
866 and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses
867 described in this manuscript were obtained from the GTEx Portal in January 2020.
868

869 **REFERENCES**

870 1. Eichler EE. Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. N
871 Engl J Med. 2019 Jul 4;381(1):64–74.

872 2. Berkovic SF, Scheffer IE, Petrou S, Delanty N, Dixon-Salazar TJ, Dlugos DJ, et al. A
873 roadmap for precision medicine in the epilepsies. Lancet Neurol. 2015 Dec
874 1;14(12):1219–28.

875 3. Schaaf CP, Betancur C, Yuen RKC, Parr JR, Skuse DH, Gallagher L, et al. A framework
876 for an evidence-based gene list relevant to autism spectrum disorder. Nat Rev Genet. 2020
877 Jun 1;21(6):367–76.

878 4. Kavanagh DH, Tansey KE, O'Donovan MC, Owen MJ. Schizophrenia genetics:
879 Emerging themes for a complex disorder. Mol Psychiatry. 2015 Feb 5;20(1):72–6.

880 5. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of
881 common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51(3):431–
882 44.

883 6. Niemi MEK, Martin HC, Rice DL, Gallone G, Gordon S, Kelemen M, et al. Common
884 genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature.
885 2018 Oct 11;562(7726):268–71.

886 7. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding
887 the missing heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747–53.

888 8. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to
889 Omnigenic. Cell. 2017 Jun;169(7):1177–86.

890 9. Coe BP, Girirajan S, Eichler EE. A genetic model for neurodevelopmental disease. Curr
891 Opin Neurobiol. 2012 Oct;22(5):829–36.

892 10. Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, Itsara A, et al. A
893 recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay.
894 Nat Genet. 2010 Mar 14;42(3):203–9.

895 11. Girirajan S, Rosenfeld JA, Coe BP, Parikh S, Friedman N, Goldstein A, et al. Phenotypic
896 Heterogeneity of Genomic Disorders and Rare Copy-Number Variants. N Engl J Med.
897 2012 Oct 4;367(14):1321–31.

898 12. Rees E, Kendall K, Pardiñas AF, Legge SE, Pocklington A, Escott-Price V, et al. Analysis
899 of intellectual disability copy number variants for association with schizophrenia. JAMA
900 Psychiatry. 2016 Sep 1;73(9):963–9.

901 13. Mefford HC, Muhle H, Ostertag P, von Spiczak S, Buysse K, Baker C, et al. Genome-
902 wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized
903 and focal epilepsies. PLoS Genet. 2010 May 20;6(5):e1000962.

904 14. Stefansson H, Meyer-Lindenberg A, Steinberg S, Magnusdottir B, Morgen K, Arnarsdottir
905 S, et al. CNVs conferring risk of autism or schizophrenia affect cognition in controls.
906 Nature. 2014;505(7483):361–6.

907 15. Girirajan S, Eichler EE. Phenotypic variability and genetic susceptibility to genomic
908 disorders. Hum Mol Genet. 2010 Oct 15;19(R2):R176-87.

909 16. Pizzo L, Jensen M, Polyak A, Rosenfeld JA, Mannik K, Krishnan A, et al. Rare variants in
910 the genetic background modulate cognitive and developmental phenotypes in individuals
911 carrying disease-associated variants. Genet Med. 2019;21(4):816–25.

912 17. Merla G, Howald C, Henrichsen CN, Lyle R, Wyss C, Zabot MT, et al. Submicroscopic
913 deletion in patients with Williams-Beuren syndrome influences expression levels of the
914 nonhemizygous flanking genes. Am J Hum Genet. 2006;79(2):332–41.

915 18. Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, et al. Genome-wide
916 transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs
917 in autism spectrum disorders. Am J Hum Genet. 2012 Jul 13;91(1):38–55.

918 19. Blumenthal I, Ragavendran A, Erdin S, Klei L, Sugathan A, Guide JR, et al.
919 Transcriptional consequences of 16p11.2 deletion and duplication in mouse cortex and
920 multiplex autism families. Am J Hum Genet. 2014 Jun 5;94(6):870–83.

921 20. Frésard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, et al. Identification of rare-
922 disease genes using blood transcriptome sequencing and large control cohorts. Nat Med.
923 2019 Jun 1;25(6):911–9.

924 21. Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, et al.
925 Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian
926 Disease. Am J Hum Genet. 2019 Mar 7;104(3):466–83.

927 22. Mohammadi P, Castel SE, Cummings BB, Einson J, Sousa C, Hoffman P, et al. Genetic
928 regulatory variation in populations informs transcriptome analysis in rare disease. Science.
929 2019 Oct 18;366(6463):351–6.

930 23. Pala M, Zappala Z, Marongiu M, Li X, Davis JR, Cusano R, et al. Population- and
931 individual-specific regulatory variation in Sardinia. Nat Genet. 2017 May 1;49(5):700–7.

932 24. De Vries BBA, White SM, Knight SJL, Regan R, Homfray T, Young ID, et al. Clinical
933 studies on submicroscopic subtelomeric rearrangements: A checklist. J Med Genet.
934 2001;38(3):145–50.

935 25. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence
936 data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

937 26. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform.
938 Bioinformatics. 2010 Jul 15;26(5):589–95.

939 27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
940 Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

941 28. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GA Van der,
942 et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv.
943 2017 Jul 24;201178.

944 29. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants

945 from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep 1;38(16):e164.

946 30. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, et al. Excess of
947 rare, inherited truncating mutations in autism. Nat Genet. 2015 Jun 11;47(6):582–8.

948 31. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The
949 mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020
950 May 28;581(7809):434–43.

951 32. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: Mutation prediction
952 for the deep-sequencing age. Nat Methods. 2014;11(4):361–2.

953 33. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping
954 and analysis of chromatin state dynamics in nine human cell types. Nature. 2011 May
955 5;473(7345):43–9.

956 34. Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. A general framework
957 for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar
958 2;46(3):310–5.

959 35. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, et al. PennCNV: An integrated
960 hidden Markov model designed for high-resolution copy number variation detection in
961 whole-genome SNP genotyping data. Genome Res. 2007 Nov;17(11):1665–74.

962 36. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover,
963 genotype, and characterize typical and atypical CNVs from family and population genome
964 sequencing. Genome Res. 2011 Jun;21(6):974–84.

965 37. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural
966 variant discovery by integrated paired-end and split-read analysis. Bioinformatics.
967 2012;28(18):333–9.

968 38. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic framework for
969 structural variant discovery. Genome Biol. 2014 Jun 26;15(6):R84.

970 39. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta:
971 Rapid detection of structural variants and indels for germline and cancer sequencing
972 applications. Bioinformatics. 2016;32(8):1220–2.

973 40. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat
974 Rev Genet. 2011 May;12(5):363–76.

975 41. Pounraja VK, Jayakar G, Jensen M, Kelkar N, Girirajan S. A machine-learning approach
976 for accurate detection of copy number variants from exome sequencing. Genome Res.
977 2019 Jul 1;29(7):1134–43.

978 42. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide
979 landscape of tandem repeat expansions. Nucleic Acids Res. 2019;47(15):e90.

980 43. Mousavi N, Margoliash J, Pusarla N, Saini S, Yanicky R, Gymrek M. TRTools: a toolkit
981 for genome-wide analysis of tandem repeats. Bioinformatics. 2020 Aug 17;ePub.

982 44. Ardlie KG, DeLuca DS, Segrè A V., Sullivan TJ, Young TR, Gelfand ET, et al. The

983 Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in
984 humans. Science. 2015 May 8;348(6235):648–60.

985 45. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast
986 universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21.

987 46. Wang L, Nie J, Sicotte H, Li Y, Eckel-Passow JE, Dasari S, et al. Measure transcript
988 integrity using RNA-seq data. BMC Bioinformatics. 2016 Feb 3;17(1):58.

989 47. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or
990 without a reference genome. BMC Bioinformatics. 2011 Aug 4;12(1):323.

991 48. Deluca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, et al. RNA-
992 SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics.
993 2012 Jun;28(11):1530–2.

994 49. Gonzalez-Mantilla AJ, Moreno-De-Luca A, Ledbetter DH, Martin CL. A cross-disorder
995 method to identify novel candidate genes for developmental brain disorders. JAMA
996 Psychiatry. 2016 Mar 1;73(3):275–83.

997 50. Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, et al.
998 SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders
999 (ASDs). Mol Autism. 2013 Oct 3;4(1):36.

1000 51. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic
1001 burden of rare disruptive mutations in schizophrenia. Nature. 2014 Feb
1002 22;506(7487):185–90.

1003 52. Firth H V., Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER:
1004 Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl
1005 Resources. Am J Hum Genet. 2009 Apr 10;84(4):524–33.

1006 53. Wright CF, Fitzgerald TW, Jones WD, Clayton S, McRae JF, Van Kogelenberg M, et al.
1007 Genetic diagnosis of developmental disorders in the DDD study: A scalable analysis of
1008 genome-wide research data. Lancet. 2015 Apr 4;385(9975):1305–14.

1009 54. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential
1010 expression analysis of digital gene expression data. Bioinformatics. 2009 Jan 1;26(1):139–
1011 40.

1012 55. Sun S, Zhu J, Mozaffari S, Ober C, Chen M, Zhou X. Heritability estimation and
1013 differential analysis of count data with generalized linear mixed models in genomic
1014 sequencing studies. Bioinformatics. 2019 Feb 1;35(3):487–96.

1015 56. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A
1016 tool set for whole-genome association and population-based linkage analyses. Am J Hum
1017 Genet. 2007;81(3):559–75.

1018 57. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association
1019 studies. Nat Genet. 2012 Jul;44(7):821–4.

1020 58. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of

1021 expression residuals (PEER) to obtain increased power and interpretability of gene
1022 expression analyses. Nat Protoc. 2012 Mar;7(3):500–7.

1023 59. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq:
1024 accounting for selection bias. Genome Biol. 2010 Feb 4;11(2):R14.

1025 60. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing
1026 strong. Nucleic Acids Res. 2019 Jan 8;47(D1):D330–8.

1027 61. Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A, et al. Transcriptional
1028 landscape of the prenatal human brain. Nature. 2014 Apr 2;508(7495):199–206.

1029 62. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to
1030 Functional Variation and the Interpretation of Personal Genomes. PLoS Genet. 2013 Aug
1031 22;9(8):e1003709.

1032 63. Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of
1033 protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285–
1034 91.

1035 64. Kim S-Y, Volsky DJ. PAGE: parametric analysis of gene set enrichment. BMC
1036 Bioinformatics. 2005 Jun 8;6(1):144.

1037 65. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network
1038 analysis. BMC Bioinformatics. 2008 Dec 29;9(1):559.

1039 66. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: Transcript-level
1040 estimates improve gene-level inferences. F1000Res. 2016;4:1521.

1041 67. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
1042 RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):550.

1043 68. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data
1044 using empirical Bayes methods. Biostatistics. 2007 Jan;8(1):118–27.

1045 69. Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. Rare variant phasing and
1046 haplotypic expression from RNA sequencing with phASER. Nat Commun. 2016 Sep
1047 8;7:12817.

1048 70. Martin M, Patterson M, Garg S, O Fischer S, Pisanti N, Klau G, et al. WhatsHap: fast and
1049 accurate read-based phasing. bioRxiv. 2016 Nov 14;085050.

1050 71. Delaneau O, Ongen H, Brown AA, Fort A, Panousis NI, Dermitzakis ET. A complete tool
1051 set for molecular QTL discovery and analysis. Nat Commun. 2017 May 18;8:15452.

1052 72. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al.
1053 Understanding multicellular function and disease with human tissue-specific networks.
1054 Nat Genet. 2015 Jun 27;47(6):569–76.

1055 73. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, et al. Genome-wide
1056 prediction and functional characterization of the genetic basis of autism spectrum disorder.
1057 Nat Neurosci. 2016 Nov 1;19(11):1454–62.

1058 74. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function
1059 using NetworkX. In: 7th Python in Science Conference (SciPy 2008). 2008. p. 11–5.

1060 75. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al.
1061 Array programming with NumPy. Nature. 2020 Sep 17;585(7825):357–62.

1062 76. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy
1063 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020 Mar
1064 1;17(3):261–72.

1065 77. McKinney W. Data Structures for Statistical Computing in Python. In: Proceedings of the
1066 9th Python in Science Conference. 2010. p. 56–61.

1067 78. Schubert J, Siekierska A, Langlois M, May P, Huneau C, Becker F, et al. Mutations in
1068 STX1B, encoding a presynaptic protein, cause fever-associated epilepsy syndromes. Nat
1069 Genet. 2014 Dec 11;46(12):1327–32.

1070 79. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A
1071 3D map of the human genome at kilobase resolution reveals principles of chromatin
1072 looping. Cell. 2014 Dec 18;159(7):1665–80.

1073 80. Werling DM, Pochareddy S, Choi J, An JY, Sheppard B, Peng M, et al. Whole-Genome
1074 and RNA Sequencing Reveal Variation and Transcriptomic Coordination in the
1075 Developing Human Prefrontal Cortex. Cell Rep. 2020 Apr 7;31(1):107489.

1076 81. Courchesne E, Gazestani VH, Lewis NE. Prenatal Origins of ASD: The When, What, and
1077 How of ASD Development. Trends Neurosci. 2020 May 1;43(5):326–42.

1078 82. Skene NG, Roy M, Grant SG. A genomic lifespan program that reorganises the young
1079 adult brain is targeted in schizophrenia. Elife. 2017 Sep 12;6:e17915.

1080 83. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, et al. The impact of rare
1081 variation on gene expression across tissues. Nature. 2017 Oct 11;550(7675):239–43.

1082 84. Zhao J, Akinsanmi I, Arafat D, Cradick TJ, Lee CM, Banskota S, et al. A Burden of Rare
1083 Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. Am J
1084 Hum Genet. 2016 Feb 4;98(2):299–309.

1085 85. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, Wolman A, et al. Modified
1086 penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat
1087 Genet. 2018 Sep 1;50(9):1327–34.

1088 86. Ballouz S, Dörfel M, Crow M, Crain J, Faivre L, Keegan CE, et al. Not by systems alone:
1089 Replicability assessment of disease expression signals. bioRxiv. 2017 Apr 18;128439.

1090 87. Mao D, Reuter CM, Ruzhnikov MRZ, Beck AE, Farrow EG, Emrick LT, et al. De novo
1091 EIF2AK1 and EIF2AK2 Variants Are Associated with Developmental Delay,
1092 Leukoencephalopathy, and Neurologic Decompensation. Am J Hum Genet. 2020 Apr
1093 2;106(4):570–83.

1094 88. Yuan H, Zhang L, Chen M, Zhu J, Meng Z, Liang L. A de novo triplication on 2q22.3
1095 including the entire ZEB2 gene associated with global developmental delay, multiple

1096 congenital anomalies and behavioral abnormalities. Mol Cytogenet. 2015 Dec 23;8(1):99.

1097 89. Baxter AL, Vivian JL, Hagelstrom RT, Hossain W, Golden WL, Wassman ER, et al. A
1098 Novel Partial Duplication of ZEB2 and Review of ZEB2 Involvement in Mowat-Wilson
1099 Syndrome. Mol Syndromol. 2017 Jun 1;8(4):211–8.

1100 90. Zahler AM, Rogel LE, Glover ML, Yitiz S, Ragle JM, Katzman S. SNRP-27, the C.
1101 Elegans homolog of the tri-snRNP 27K protein, has a role in 5′ splice site positioning in
1102 the spliceosome. RNA. 2018 Oct 1;24(10):1314–25.

1103 91. Saito T, Guan F, Papolos DF, Lau S, Klein M, Fann CSJ, et al. Mutation analysis of
1104 SYNJ1: A possible candidate gene for chromosome 21q22-linked bipolar disorder. Mol
1105 Psychiatry. 2001;6(4):387–95.

1106 92. Mitchel MW, Moreno-De-Luca D, Myers SM, Finucane B, Ledbetter DH, Martin CL.
1107 17q12 Recurrent Deletion Syndrome. In: Adam MP, Ardinger HH, Pagon RA, Wallace
1108 SE, Bean LJ, Stephens K, et al., editors. GeneReviews®. University of Washington,
1109 Seattle; 2016.

1110 93. Lim ET, Uddin M, De Rubeis S, Chan Y, Kamumbu AS, Zhang X, et al. Rates,
1111 distribution and implications of postzygotic mosaic mutations in autism spectrum
1112 disorder. Nat Neurosci. 2017 Sep 1;20(9):1217–24.

1113 94. Ramírez-Castillejo C, Sánchez-Sánchez F, Andreu-Agulló C, Ferrón SR, Aroca-Aguilar
1114 JD, Sánchez P, et al. Pigment epithelium-derived factor is a niche signal for neural stem
1115 cell renewal. Nat Neurosci. 2006 Mar 19;9(3):331–9.

1116 95. Yao I, Iida J, Nishimura W, Hata Y. Synaptic and nuclear localization of brain-enriched
1117 guanylate kinase-associated protein. J Neurosci. 2002 Jul 1;22(13):5354–64.

1118 96. De Wit MCY, De Coo IFM, Halley DJJ, Lequin MH, Mancini GMS. Movement disorder
1119 and neuronal migration disorder due to ARFGEF2 mutation. Neurogenetics.
1120 2009;10(4):333–6.

1121 97. Nowak F V. Porf-2 = arhgap39 = vilse: A pivotal role in neurodevelopment, learning and
1122 memory. eNeuro. 2018 Sep 1;5(5):e0082-18.2018.

1123 98. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare Variants Create
1124 Synthetic Genome-Wide Associations. PLoS Biol. 2010 Jan 26;8(1):e1000294.

1125 99. Zhang X, Zhang Y, Zhu X, Purmann C, Haney MS, Ward T, et al. Local and global
1126 chromatin interactions are altered by large genomic deletions associated with human brain
1127 development. Nat Commun. 2018 Dec 1;9(1):5356.

1128 100. Powell JE, Henders AK, McRae AF, Kim J, Hemani G, Martin NG, et al. Congruence of
1129 Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and
1130 SNP Data. PLoS Genet. 2013 May 16;9(5):e1003502.

1131 101. Lenz TL, Deutsch AJ, Han B, Hu X, Okada Y, Eyre S, et al. Widespread non-additive and
1132 interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat Genet.
1133 2015 Aug 27;47(9):1085–90.

1134 102. Nordsletten AE, Larsson H, Crowley JJ, Almqvist C, Lichtenstein P, Mataix-Cols D.
1135 Patterns of nonrandom mating within and across 11 major psychiatric disorders. JAMA
1136 Psychiatry. 2016 Apr 1;73(4):354–61.

1137 103. Owen MJ. Intellectual disability and major psychiatric disorders: A continuum of
1138 neurodevelopmental causality. Br J Psychiatry. 2012 Apr;200(4):268–9.

1139 104. Connolly S, Anney R, Gallagher L, Heron EA. Evidence of Assortative Mating in Autism
1140 Spectrum Disorder. Biol Psychiatry. 2019 Aug 15;86(4):286–93.

1141 105. Pizzo L, Lasser M, Yusuff T, Jensen M, Ingraham P, Huber E, et al. Functional
1142 assessment of the "two-hit" model for neurodevelofpmental defects in Drosophila and X.
1143 laevis. PLoS Genet. 2021 Apr 1;17(4):e1009112.

1144 106. Wang Y, Song F, Zhang B, Zhang L, Xu J, Kuang D, et al. The 3D Genome Browser: A
1145 web-based browser for visualizing 3D genome organization and long-range chromatin
1146 interactions. Genome Biol. 2018 Oct 4;19(1):151.

1147

1148 **FIGURE TITLES AND LEGENDS**

1149 **Figure 1. Overview of experimental design.** We performed whole genome sequencing, RNA
1150 sequencing, and clinical phenotyping on five large families (32 total individuals) with the
1151 16p12.1 deletion, indicated with red asterisks in the pedigrees. Children (green) and adults (blue)
1152 in the pedigrees are shaded by phenotypic severity score, with white indicating no clinical
1153 features, lighter shades indicating mild features (child de Vries score of 1-4; adult score of 1-2
1154 features), medium shades indicating moderate features (child de Vries score of 5-8; adult score
1155 of 3-4 features), darker shades indicating severe features (child de Vries score of 9-13; adult
1156 score of 5-6 features), and grey indicating no phenotypic data available. Phenotypic severity
1157 scores are described in the Methods and are listed for each person in **Table S1**. We then
1158 performed multiple analyses to assess the role of the deletion and rare "second-hit" variants
1159 towards the observed transcriptomic changes and developmental phenotypes, including
1160 differential expression between carriers and noncarriers of the deletion, differential expression
1161 between parents and carrier offspring in 13 trios from the five families, outlier gene expression
1162 among all individuals, identification of additional transcriptomic alterations such as alternative
1163 splicing and allele-specific expression, and gene interaction patterns in the context of a brain-
1164 specific network.

1165

1166 **Figure 2. Differential expression of genes between carriers and noncarriers of the 16p12.1**
1167 **deletion. (A)** Volcano plot showing downstream (non-16p12.1 deletion) genes differentially
1168 expressed (FDR<0.05) between carriers (n=19) and noncarriers (n=13) of the deletion. Red
1169 circles indicate genes preferentially expressed in GTEx cerebral cortex tissues. **(B)** Scatter plot
1170 showing all genes differentially expressed between carriers and noncarriers of the deletion by
1171 chromosome, excluding genes on sex chromosomes. Genes are colored by FDR of differential
1172 expression. Labeled genes indicate candidate autism genes with differential expression. **(C)**
1173 Expression changes and chromatin connectivity of genes within the 16p12.1 region. The top plot
1174 shows pairwise chromatin interactions within the 3.5 Mbp 16p12.1 region, with red lines
1175 representing stronger Hi-C intensity, while the bottom plot shows $\log_2$-fold change of expression
1176 in deletion carriers of genes within and adjacent to the 16p12.1 deletion. The Hi-C data is from
1177 previously reported Hi-C experiments of LCL samples (79), and the heatmap was generated
1178 using the 3D Genome Browser (106). **(D)** Line plot shows enrichment ($\log_{10}$ FDR) of

1179 differentially expressed genes in deletion carriers for genes preferentially expressed in six select
1180 BrainSpan tissues across 11 developmental timepoints.
1181
1182 **Figure 3. Differential expression of genes between offspring and carrier and noncarrier**
1183 **parents. (A)** Boxplot shows the proportion of differentially expressed genes in carrier offspring
1184 of 13 trios (**Table S1**) that were either unique to the offspring or shared with their carrier or
1185 noncarrier parents (*p<0.05, two-tailed paired Mann-Whitney test). Boxplot indicates median
1186 (center line), 25th and 75th percentiles (bounds of box), and minimum and maximum (whiskers).
1187 **(B)** Table shows observed clinical features in eight carrier children with overt developmental
1188 phenotypes, as well as enrichments (FDR<0.05) of differentially expressed genes in each carrier
1189 child for biological functions related to each clinical feature. Cells are colored according to the
1190 family-specific patterns (uniquely observed or shared with a parent) of differentially expressed
1191 genes for each enriched biological process.
1192
1193 **Figure 4. Enrichment of "second-hit" variants near genes with outlier expression. (A)**
1194 Forest plot shows enrichment (Fisher's exact test, **=FDR<0.05, *=uncorrected p<0.05) of
1195 genes with outlier expression in all individuals in the cohort (n=32) for rare proximal coding and
1196 non-coding variants, including single-nucleotide variants (SNVs) and insertions/deletions
1197 (indels) with CADD scores >10 (34), structural variants (SVs), and short tandem repeats (STRs).
1198 **(B)** Forest plot shows classes of "second-hit" variants with significant enrichment (Fisher's exact
1199 test, **=FDR<0.05, *=uncorrected p<0.05) towards genes with outlier expression in carrier
1200 children (n=10), carrier parents (n=6), or noncarrier parents (n=6). **(C)** Forest plot shows classes
1201 of "second-hit" variants with significant enrichment (Fisher's exact test, **=FDR<0.05,
1202 *=uncorrected p<0.05) towards genes with outlier expression in carrier children (n=9) that are
1203 shared with either carrier or noncarrier parents. All forest plots show log-odds ratios (dots) and
1204 95% confidence intervals (whiskers). Odds ratios, confidence intervals, p-values, and Benjamini-
1205 Hochberg corrected FDR values for comparisons with all classes of "second-hit" variants are
1206 listed in **Table S5**. **(D)** Scatter plot shows expression values (transcripts per million, or TPM) for
1207 *EIF2AK1* in LCL replicates for all individuals (n=32). Samples in blue have outlier expression of
1208 *EIF2AK1* (z-score <-2) and carry a deleterious "second-hit" variant in the 5' UTR of the gene.
1209 **(E)** Scatter plots show expression values (TPM) of genes with synergistic effects due to the

1210 16p12.1 deletion and inherited "second-hit" variants. Blue circles indicate expression values for
1211 samples from carrier children and family members with rare "second-hit" variants, orange circles
1212 indicate expression values for samples from family members without the "second-hit" variant,
1213 and green circles indicate expression values of samples from other deletion carriers and
1214 noncarriers in the cohort. Black lines denote median gene expression for LCL replicates of each
1215 individual used to identify genes with outlier expression in individual deletion carriers.
1216
1217 **Figure 5. Alternative splicing, allele specific expression, eQTL, and network analysis. (A)**
1218 Forest plot shows classes of rare variants with significant enrichment (Fisher's exact test,
1219 **=FDR<0.05, *=uncorrected p<0.05) towards genes with alternative splicing in carrier children
1220 (n=9) that are shared with either carrier or noncarrier parents. Forest plot shows log-odds ratios
1221 (dots) and 95% confidence intervals (whiskers). Odds ratios, confidence intervals, p-values, and
1222 Benjamini-Hochberg corrected FDR values for comparisons with all classes of "second-hit"
1223 variants are listed in **Table S5**. (**B)** Scatter plot shows isoform usage percentage for *TADA2A*-
1224 003 in replicates for individuals in family GL_007. Samples in blue carry a "second-hit" splice-
1225 site variant in *TADA2A* and exhibit a higher frequency of the alternative isoform. **(C)** Scatter plot
1226 shows allele frequencies for the autism-associated gene *CARD11* in carrier child P2C_07,
1227 noncarrier parent F2NC_07, and carrier parent M2C_07 in GL_007. Blue circles indicate allele
1228 frequency for haplotypes carrying a "second-hit" coding variant disrupting *CARD11*. **(D)** Scatter
1229 plot shows z-scores for expression values of *ARHGAP39* for all individuals with available
1230 genotypes for the gene. Individuals who carry the minor allele for the *ARHGAP39* eQTL (blue
1231 dots) have higher expression of the gene than the rest of the cohort (orange dots). **(E)** Plot shows
1232 correlations among the numbers of gene expression alterations in carrier offspring for the 13 trios
1233 assessed in our study. Colors and sizes of the circles are proportional to the correlation
1234 coefficients between gene expression changes, where blue indicates a positive correlation and
1235 red indicates a negative correlation. Asterisks denote significant correlations (FDR<0.05,
1236 Pearson correlation with Benjamini-Hochberg correction). **(F)** Boxplot shows the average
1237 shortest distances for carrier offspring (n=13) between pairs of genes with "second-hit" coding
1238 variants and genes with identified expression changes in a brain-specific network. Genes with
1239 expression changes were more strongly connected to genes with "second-hit" variants in the
1240 brain-specific network than the average distances for genes within 100 permuted brain-specific

1241 networks per sample ($p=4.88\times10^{-4}$, two-tailed paired Mann-Whitney test). Boxplot indicates
1242 median (center line), 25th and 75th percentiles (bounds of box), and minimum and maximum
1243 (whiskers).
1244
1245 **Figure 6. Genetic and transcriptomic mechanisms for phenotypic variability in 16p12.1**
1246 **deletion families.** Affected children inherit the 16p12.1 deletion (red) and a smaller number of
1247 rare "second-hit" variants (green) from a carrier parent, and a larger number of "second-hit"
1248 variants from the noncarrier parent (blue). Altered expression of genes due to these "second-hit"
1249 variants affects nearby downstream connected genes in an interaction network (grey), causing
1250 additional transcriptomic perturbation. Because of this, carrier children have numerous gene
1251 expression changes compared with their carrier parents, including genes showing synergistic
1252 effects of the deletion and "second-hit" variants (orange), potentially accounting for more severe
1253 developmental phenotypes observed in the children.

1254 **ADDITIONAL FILES**

1255

1256 **Additional file 1:** Fifteen supporting Figures S1-S15. A figure caption for each is given within
1257 the file (Format: PDF).

1258

1259 **Additional files 2-16:** Individual files for supporting Tables S1-S15 (Format: Excel). Table
1260 captions are as follows:

1261

1262 **Table S1.** This file lists 32 individuals in the five 16p12.1 deletion families by family
1263 relationship, sex, deletion carrier status, and observed developmental or neuropsychiatric
1264 phenotypes, including modified de Vries scores for children and adult phenotypic severity
1265 scores. The file also lists membership of 13 trios with carrier offspring assessed for family-based
1266 comparisons in this study.

1267

1268 **Table S2.** This file summarizes the number of genomic variants (SNVs, CNVs, and STRs)
1269 present in each individual in the 16p12.1 deletion cohort.

1270

1271 **Table S3**. This file lists Coriell Institute accession numbers for the LCL samples used in this
1272 study.

1273

1274 **Table S4.** This file lists differentially expressed transcripts between carriers and noncarriers of
1275 the 16p12.1 deletion, using both the main analysis and relatedness correction methods. It also
1276 includes enrichment of differentially expressed genes for Gene Ontology terms, candidate
1277 neurodevelopmental-associated genes, and genes preferentially expressed in GTEx and
1278 BrainSpan datasets.

1279

1280 **Table S5**. This file contains all information on the statistic tests performed in the manuscript,
1281 including sample sizes, test statistics, log-odds ratios, confidence intervals, p-values, and
1282 Benjamini-Hochberg corrected FDR. * indicates $p<0.05$ without multiple testing correction, and
1283 ** indicated FDR<0.05 after correction.

1284

1285 **Table S6**. This file lists significantly up- or down-regulated Gene Ontology biological process
1286 terms in 16p12.1 deletion carriers, as identified using Parametric Analysis of Gene Set
1287 Enrichment (PAGE).
1288
1289 **Table S7.** This file lists module assignments for genes derived from weighted gene co-
1290 expression network analysis, and the enrichment of genes in six modules that correspond to
1291 deletion carrier status for Gene Ontology terms.
1292
1293 **Table S8**. This file summarizes the numbers of gene expression changes, by family-specific
1294 pattern where applicable, identified in each individual in the 16p12.1 deletion cohort. Boxes
1295 shaded grey and labeled N/A indicate samples without available family-specific patterns for
1296 expression changes.
1297
1298 **Table S9.** This file lists differentially expressed genes identified in each of the offspring in all
1299 trios (n=13 with carrier offspring and n=4 with noncarrier offspring) by family-specific pattern
1300 (unique occurrence or shared with a parent), and the enrichment of each gene set for Gene
1301 Ontology terms.
1302
1303 **Table S10**. This file lists all genes in each individual that showed any gene expression change
1304 (differential expression, outlier expression, alternative splicing, ASE, or eQTL minor allele),
1305 with family-specific patterns when applicable, alongside the number of identified rare variants
1306 disrupting each gene.
1307
1308 **Table S11.** This file lists all outlier genes identified in each individual in the 16p12.1 deletion
1309 cohort, along with their expression z-scores, preferential expression in the human brain, and pLI
1310 and RVIS intolerance to variation scores.
1311
1312 **Table S12.** This file lists rare second-hit variants that may contribute to synergistic gene
1313 expression changes along with the 16p12.1 deletion in carrier children.
1314

1315 **Table S13.** This file lists isoforms and genes with alternative splicing identified in offspring of
1316 all trios (n=13 with carrier offspring and n=4 with noncarrier offspring) by family-specific
1317 pattern (unique occurrence or shared with a parent).
1318
1319 **Table S14.** This file lists genes with allele-specific expression identified in all individuals in the
1320 cohort, including the presence of rare deleterious coding variants on the overexpressed haplotype
1321 of each gene.
1322
1323 **Table S15.** This file lists identified eQTL variants in the 16p12.1 deletion cohort, including beta
1324 and FDR values, population frequency, associated eGene, and presence in GTEx LCL datasets.
1325 The file also lists all individuals in the cohort who carry a minor allele for the identified eQTLs.
1326

Figure 1

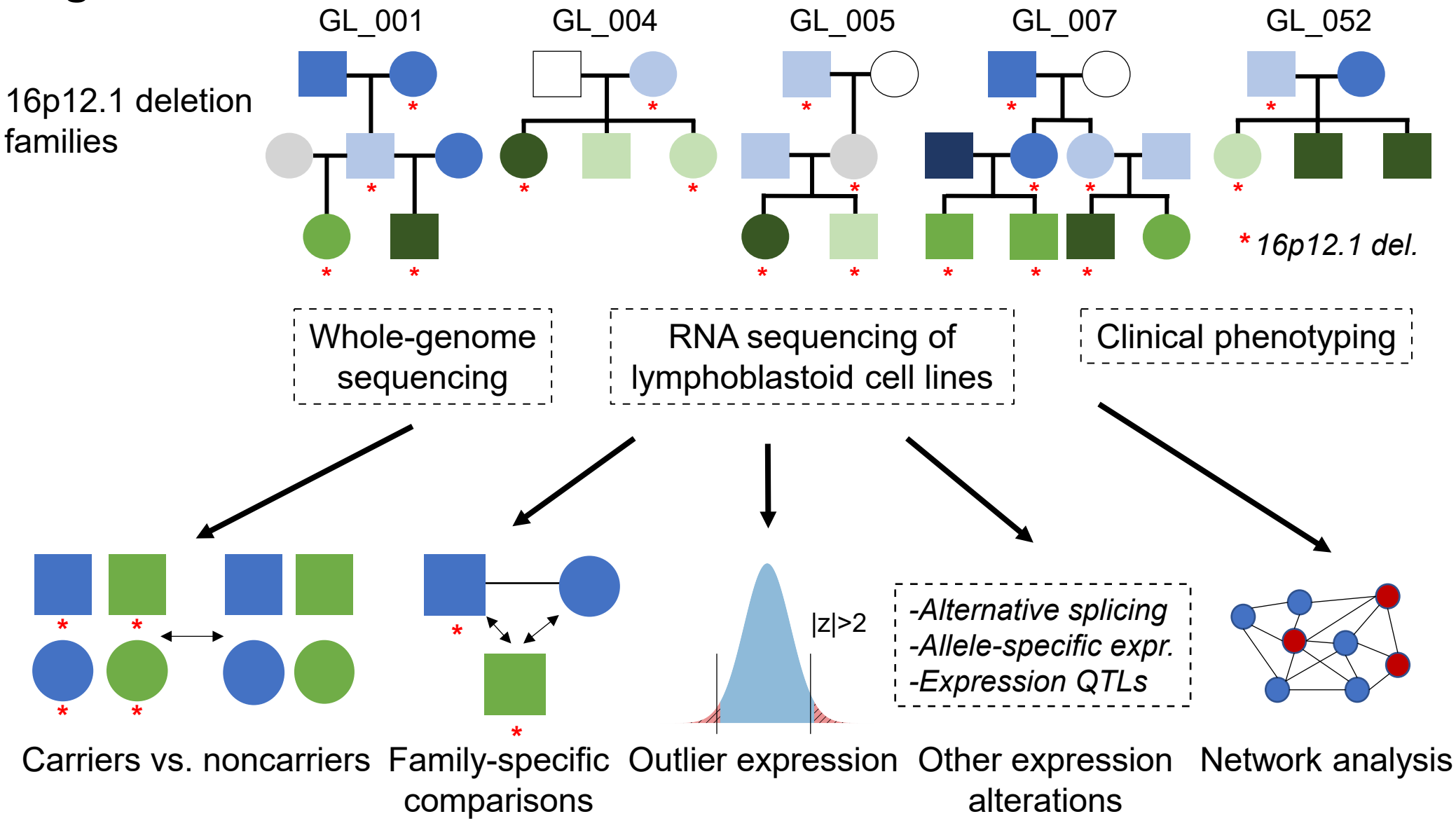

16p12.1 deletion families
GL_001
GL_004
GL_005
GL_007
GL_052
* 16p12.1 del.
Whole-genome sequencing
RNA sequencing of lymphoblastoid cell lines
Clinical phenotyping
|z|>2
-Alternative splicing
-Allele-specific expr.
-Expression QTLs
Carriers vs. noncarriers
Family-specific comparisons
Outlier expression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

Figure 2

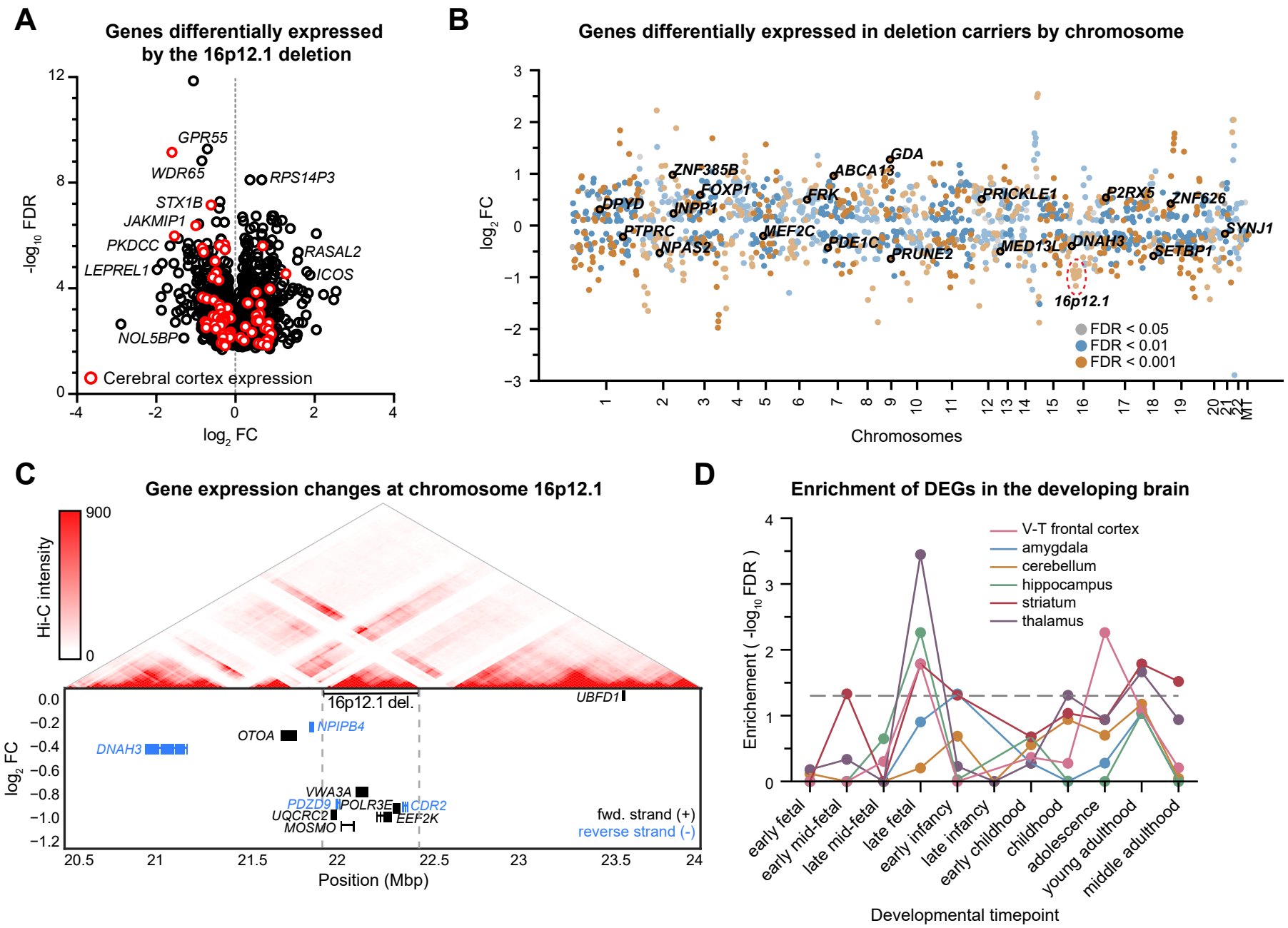
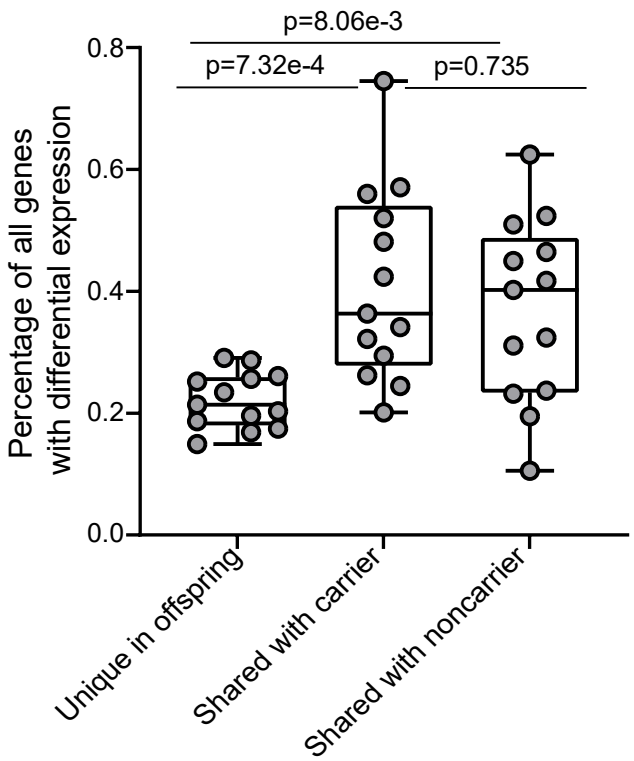

A
Genes differentially expressed by the 16p12.1 deletion
-log10 FDR
log2 FC
GPR55
WDR65
RPS14P3
STX1B
JAKMIP1
PKDCC
LEPREL1
RASAL2
ICOS
NOL5BP
Cerebral cortex expression
B
Genes differentially expressed in deletion carriers by chromosome
log2 FC
Chromosomes
GDA
ZNF385B
ABCA13
FOXP1
FRK
PRICKLE1
P2RX5
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.		Unique and shared	Unique and shared		Unique and shared						Unique	Unique and shared	Unique and shared	Unique and shared			Unique and shared		Unique and shared
Neurotransmitter metab.			Shared w/ noncarrier									Shared w/ carrier	Shared w/ carrier	Shared w/ carrier					Unique
Neuron projection					Shared w/ carrier							Unique	Unique	Unique					Unique
Synapse organization											Unique						Unique and shared		
Synaptic transmission			Shared w/ carrier					Shared w/ noncarrier				Unique	Unique				Shared w/ carrier		Unique
Axon formation												Shared w/ noncarrier	Shared w/ noncarrier	Unique					
Axon outgrowth			Shared w/ carrier		Shared w/ carrier				Unique	Unique									
Skeletal development	Unique						Unique									Unique and shared		Shared w/ carrier	
Kidney development						Unique													
Muscle development				Unique			Unique								Unique and shared	Unique and shared			

Unique | Shared w/ carrier | Shared w/ noncarrier | Unique and shared

Figure 4

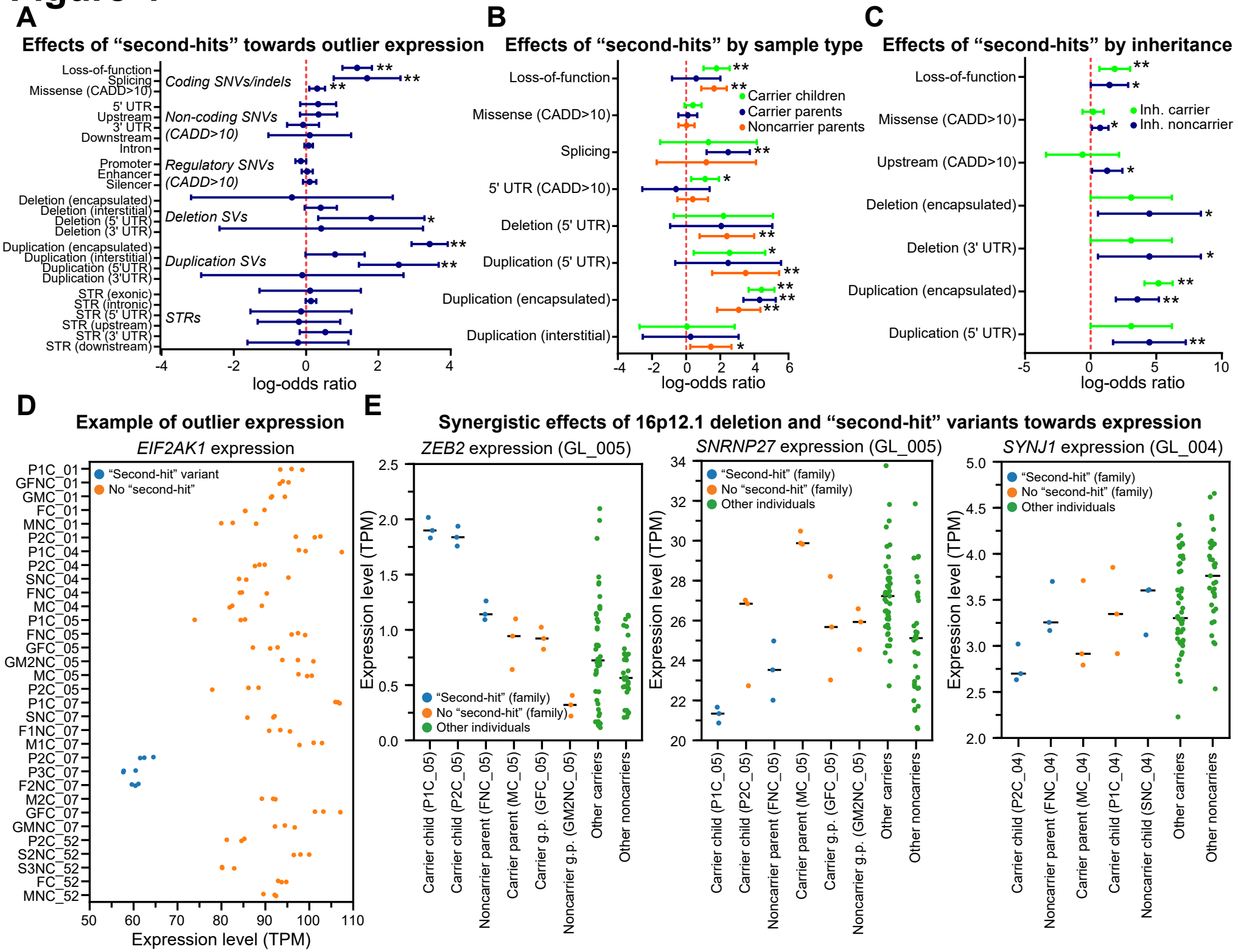

A
Effects of "second-hits" towards outlier expression
Coding SNVs/indels
Loss-of-function
Splicing
Missense (CADD>10)
Non-coding SNVs (CADD>10)
5' UTR
Upstream
3' UTR
Downstream
Intron
Regulatory SNVs (CADD>10)
Promoter
Enhancer
Silencer
Deletion SVs
Deletion (encapsulated)
Deletion (interstitial)
Deletion (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

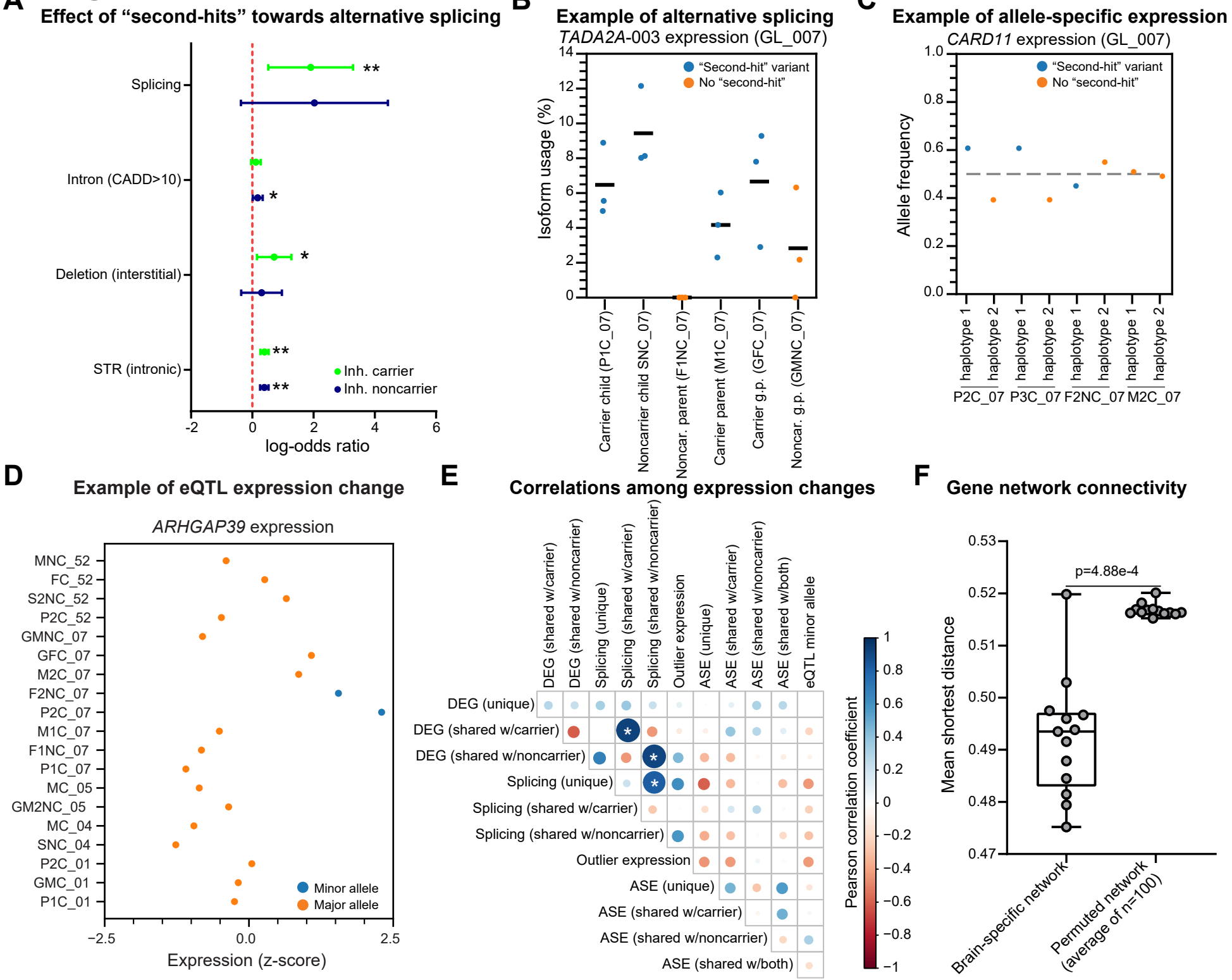
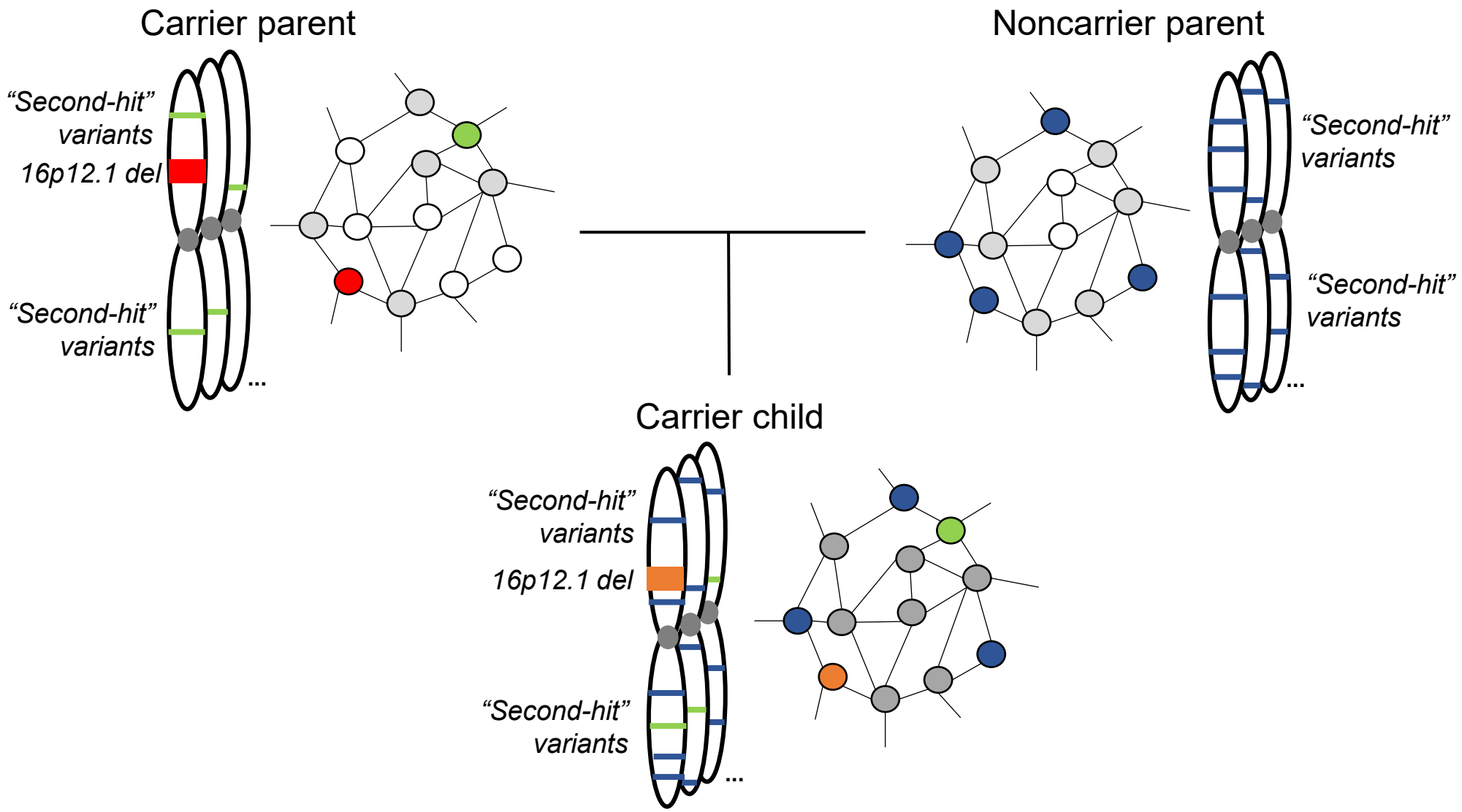

A
Effect of "second-hits" towards alternative splicing
Splicing
Intron (CADD>10)
Deletion (interstitial)
STR (intronic)
Inh. carrier
Inh. noncarrier
log-odds ratio
B
Example of alternative splicing
TADA2A-003 expression (GL_007)
"Second-hit" variant
No "second-hit"
Isoform usage (%)
Carrier child (P1C_07)
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 (shared w/carrier)
DEG (shared w/noncarrier)
Splicing (unique)
Splicing (shared w/carrier)
Splicing (shared w/noncarrier)
Outlier expression
ASE (unique)
ASE (shared w/carrier)
ASE (shared w/noncarrier)
ASE (shared w/both)
eQTL minor allele
DEG (unique)
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
...