Genetic modifiers and ascertainment drive variable expressivity of complex disorders

Matthew Jensen[1,2]*, Corrine Smolen[1,2]*, Anastasia Tyryshkina[1]*, Lucilla Pizzo[1]*, Deepro Banerjee[1], Matthew Oetjens[3], Hermela Shimelis[3], Cora M. Taylor[3], Vijay Kumar Pounraja[1,2], Hyebin Song[4], Laura Rohan[1], Emily Huber[1], Laila El Khattabi[5], Ingrid van de Laar[6], Rafik Tadros[6], Connie Bezzina[6], Marjon van Slegtenhorst[6], Janneke Kammeraad[6], Paolo Prontera[7], Jean-Hubert Caberg[8], Harry Fraser[9], Siddhartha Banka[9,10], Anke Van Dijck[11], Charles Schwartz[12], Els Voorhoeve[13], Patrick Callier[14], Anne-Laure Mosca-Boidron[14], Nathalie Marle[14], Mathilde Lefebvre[15], Kate Pope[16], Penny Snell[16], Amber Boys[16], Paul J. Lockhart[16,17], Myla Ashfaq[18], Elizabeth McCready[19], Margaret Nowacyzk[19], Lucia Castiglia[20], Ornella Galesi[20], Emanuela Avola[20], Teresa Mattina[20], Marco Fichera[20,21], Maria Grazia Bruccheri[20], Giuseppa Maria Luana Mandarà[22], Francesca Mari[23], Flavia Privitera[23], Ilaria Longo[23], Aurora Curró[23], Alessandra Renieri[23], Boris Keren[24], Perrine Charles[24], Silvestre Cuinat[25], Mathilde Nizon[25], Olivier Pichon[25], Claire Bénéteau[25], Radka Stoeva[25], Dominique Martin-Coignard[26], Sophia Blesson[27], Cedric Le Caignec[28,29], Sandra Mercier[27], Marie Vincent[27], Christa Martin[3], Katrin Mannik[30,31], Alexandre Reymond[32], Laurence Faivre[14,15], Erik Sistermans[13], R. Frank Kooy[11], David J. Amor[13], Corrado Romano[20,21], Joris Andrieux[33], and Santhosh Girirajan[1,2,34]

1. Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA.
2. Bioinformatics and Genomics Graduate program, Pennsylvania State University, University Park, PA 16802, USA.
3. Autism & Developmental Medicine Institute, Geisinger, Lewisburg, PA 17837, USA.
4. Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA.
5. Institut Cochin, Inserm U1016, CNRS UMR8104, Université Paris Cité, CARPEM, Paris, France.
6. Department of Clinical Genetics, Erasmus MC, Univ. Medical Center Rotterdam, Rotterdam, The Netherlands.
7. Medical Genetics Unit, Hospital Santa Maria della Misericordia, Perugia, Italy.
8. Centre Hospitalier Universitaire de Liège. Domaine Universitaire du Sart Tilman, Liège, Belgium.
9. Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK.
10. Manchester Centre for Genomic Medicine, St. Mary's Hospital, Central Manchester University Hospitals, NHS Foundation Trust Manchester Academic Health Sciences Centre, Manchester, UK.
11. Department of Medical Genetics, University and University Hospital Antwerp, Antwerp, Belgium.
12. Greenwood Genetic Center, Greenwood, SC 29646, USA.

13. Department of Clinical Genetics, Amsterdam UMC, Amsterdam, The Netherlands.
14. Center for Rare Diseases and Reference Developmental Anomalies and Malformation Syndromes, CHU Dijon, Dijon, France.
15. Laboratoire de Genetique Chromosomique et Moleculaire, CHU Dijon, France.
16. Bruce Lefroy Centre, Murdoch Children's Research Institute, Melbourne, Australia.
17. Department of Paediatrics, University of Melbourne, Melbourne, Australia.
18. Department of Pediatrics, McGovern Medical School, University of Texas Health Science Center, Houston, TX 77030, USA.
19. Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada.
20. Research Unit of Rare Diseases and Neurodevelopmental Disorders, Oasi Research Institute-IRCCS, Troina, Italy.
21. Section of Clinical Biochemistry and Medical Genetics, Department of Biomedical and Biotechnological Sciences, University of Catania School of Medicine, Catania, Italy.
22. Medical Genetics, ASP Ragusa, Ragusa, Italy.
23. Laboratory of Clinical Molecular Genetics and Cytogenetics, IRCCS San Raffaele Scientific Institute, Milan, Italy.
24. Département de Génétique, Hôpital Pitié-Salpêtrière, Assistance Publique-Hôpitaux de Paris, Sorbonne Université, 75019 Paris, France.
25. CHU Nantes, Medical Genetics Department, Nantes, France.
26. Service de Cytogenetique, CHU de Le Mans, Le Mans, France.
27. Department of Genetics, Bretonneau University Hospital, Tours, France.
28. CHU Toulouse, Department of Medical Genetics, Toulouse, France.
29. Toulouse Neuro Imaging, Center, Inserm, UPS, Université de Toulouse, Toulouse, France.
30. Institute of Genomics, University of Tartu, Estonia.
31. Health2030 Genome Center, Fondation Campus Biotech, Geneva, Switzerland.
32. Center for Integrative Genomics, Faculty of Biology and Medicine, University of Lausanne, Switzerland.
33. Institut de Genetique Medicale, Hopital Jeanne de Flandre, CHRU de Lille, Lille, France.
34. Department of Anthropology, Pennsylvania State University, University Park, PA 16802, USA.

Correspondence:

Santhosh Girirajan

205A Life Sciences Building

Pennsylvania State University

University Park, PA 16803

Email: sxg47@psu.edu

SUMMARY

Variable expressivity of disease-associated variants implies a role for secondary variants that modify clinical features. We assessed the effects of modifier variants towards clinical outcomes of 2,252 individuals with primary variants. Among 132 families with the 16p12.1 deletion, distinct rare and common variant classes conferred risk for specific developmental features, including short tandem repeats for neurological defects and SNVs for microcephaly, while additional disease-associated variants conferred multiple genetic diagnoses. Within disease and population cohorts of 773 individuals with the 16p12.1 deletion, we found opposing effects of secondary variants towards clinical features across ascertainments. Additional analysis of 1,479 probands with other primary variants, such as 16p11.2 deletion and *CHD8* variants, and 1,084 without primary variants, showed that phenotypic associations differed by primary variant context and were influenced by synergistic interactions between primary and secondary variants. Our study provides a paradigm to dissect the genomic architecture of complex disorders towards personalized treatment.

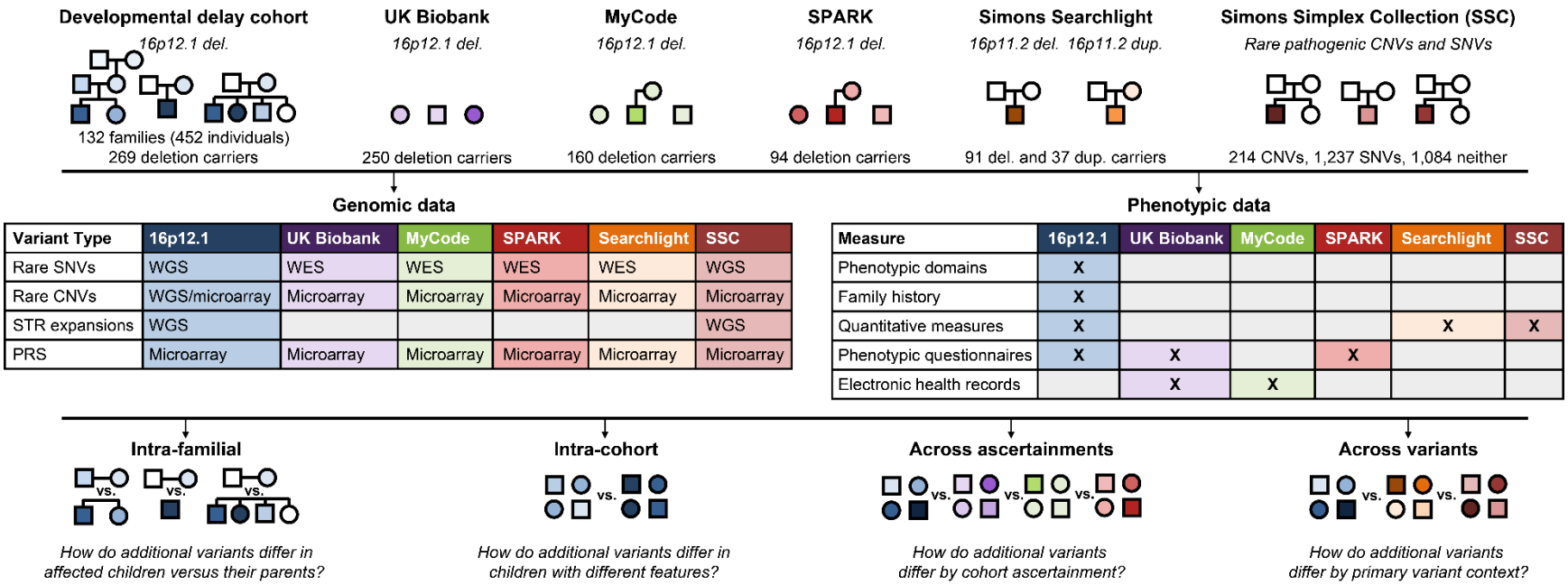
INTRODUCTION

As large-scale sequencing studies uncover increasingly complex links between genomic variants and heritable disorders, identifying the genetic etiology in an affected individual has become more challenging[1]. In contrast to Mendelian disorders caused by single genes, many disorders are characterized by extensive phenotypic heterogeneity, where individuals with the same variant exhibit a range of phenotypes with variable penetrance and expressivity[2,3]. Some instances of phenotypic heterogeneity can be explained by multiple genetic diagnoses, where more than one pathogenic variant contributes to largely independent disorders in the same individual[4]. These variants can even synergistically contribute to new phenotypes not associated with the individual variants, such as seizures observed among individuals with variants in both *MKS1* (associated with Meckel-Gruber syndrome) and *BBS1* (associated with Bardet-Biedl syndrome)[5]. Other cases of phenotypic heterogeneity could occur when the clinical features of causal variants are modified by secondary variants that do not cause disease themselves[6]. For example, rare variants in histone modifier genes were enriched among individuals with the 22q11.2 deletion who exhibited variably expressive congenital heart defects[7]. The complexity increases exponentially for neurodevelopmental disorders, where the combined effects of primary and secondary variants with differing frequency and effect sizes explain their broad heterogeneity[8,9]. For example, recent studies have found significant contributions of polygenic risk from common variants towards phenotypes of individuals with pathogenic copy-number variants (CNVs)[10,11], such as schizophrenia risk in individuals with 22q11.2 deletion[12–17]. Further, variable expressivity could also be explained by cohort ascertainment, as many pathogenic variants are enriched among individuals across disease ascertainments and lead to medical consequences in the general population or healthy-biased cohorts[18–22]. For example, the autism-associated 16p11.2 deletion[23] is also associated with obesity, musculoskeletal, pulmonary, hematologic, and renal features in the general population[24,25]. This complex interplay necessitates a systematic assessment to fully understand which modifier variants contribute to specific phenotypes when ascertained for the same primary variant.

Rare recurrent CNVs represent excellent models to study variable expressivity, as the large number of duplicated or deleted genes increases the likelihood of interactions with modifiers elsewhere in the genome[3,26]. For example, the rare heterozygous 520-kbp 16p12.1 deletion (hg18/NCBI36 when originally reported; currently maps to 16p12.2 based on

hg19/GRCh37) is enriched among children with severe neurodevelopmental features and is inherited in >90% of cases from a parent who manifests milder psychiatric and cognitive features[27–29]. The phenotypic manifestation among individuals with this deletion differs based on cohort ascertainment. For instance, the deletion was originally described in children with developmental delay[27], but studies from the general population also identified associations with multiple psychiatric and cognitive features[22,30–32]. Thus, the 16p12.1 deletion is an ideal paradigm for studying the effects of modifier variants on the clinical trajectory of a primary variant. We previously found that severely affected children with the deletion have a global increase in rare variant burden compared to parents with the deletion, and these trends are consistent for other primary variants[27,28,33]. Our findings suggested a "multi-hit" model for complex disease etiology, where a primary variant sensitizes an individual for disease and the clinical outcome is determined by other "hits" elsewhere in the genome[3]. However, it is not completely understood how specific variant classes of differing effect size and frequency modify clinical features across different ascertainment and primary variant contexts.

Here, we performed deep clinical and quantitative phenotyping and comprehensive analysis of genomic data for 2,252 individuals with primary variants from diverse disease and population-based cohorts (**Fig. 1**). We systematically dissected the roles of multiple secondary variant classes towards developmental features in 132 families with the 16p12.1 deletion and expanded our analysis to uncover phenotypic associations in 773 16p12.1 deletion individuals from disease cohorts and healthy populations as well as 1,479 autism probands who carry a range of other primary variants and 1,084 autism probands without primary variants. Our results show that variant-phenotype associations are dependent on both the primary and secondary variant context as well as cohort ascertainment (**Fig. 1**), allowing for more accurate dissection of the genetic etiology of variably expressive traits associated with pathogenic variants.



Variant Type	16p12.1	UK Biobank	MyCode	SPARK	Searchlight	SSC
Rare SNVs	WGS	WES	WES	WES	WES	WGS
Rare CNVs	WGS/microarray	Microarray	Microarray	Microarray	Microarray	Microarray
STR expansions	WGS					WGS
PRS	Microarray	Microarray	Microarray	Microarray	Microarray	Microarray

Measure	16p12.1	UK Biobank	MyCode	SPARK	Searchlight	SSC
Phenotypic domains	X					
Family history	X					
Quantitative measures	X				X	X
Phenotypic questionnaires	X	X		X		
Electronic health records		X	X			

Figure 1. Overview of variant-phenotype analyses in 2,252 individuals with primary pathogenic variants. We assessed associations between variant classes and clinical phenotypes in six cohorts of individuals and families with primary variants. We directly recruited and assessed 132 families with the 16p12.1 deletion primarily ascertained for children with developmental delay (DD) (also including ten individuals from eight families from the Estonian Biobank not ascertained for DD). We further assessed 16p12.1 deletion carriers from cohorts with different ascertainments, including healthy volunteer-biased (UK Biobank), clinically-derived (MyCode), and single-disorder (SPARK, for autism) ascertainments. We finally assessed probands ascertained for autism with various primary pathogenic variants, including the 16p11.2 deletion or duplication (Simons Searchlight) and other large CNVs or rare SNVs in neurodevelopmental genes (SSC). We note that 100 probands in SSC have both pathogenic SNVs and CNVs and are included in both categories. Within and across these cohorts, we identified associations between up to 17 classes of rare and common variants (identified from WGS, WES, and microarrays) with phenotypic features from deep clinical datasets and electronic health records.

RESULTS

Variability of clinical features in 16p12.1 deletion

We recruited a cohort of 442 individuals from 124 families with the 16p12.1 deletion ("DD cohort"), including multi-generational families, primarily ascertained for having children with developmental delay (DD) (**Fig. 1**). We analyzed phenotypes from medical records, family interviews, and online assessments for quantitative traits, such as non-verbal IQ[34] and social responsiveness scores for autism-related social traits (SRS[35]) (**Table S1A**). In total, 93% of probands (84/90) inherited the deletion from a parent, with a slight bias towards maternal inheritance (48/84, 57%), and 70% (87/124) of probands were male. Probands with the deletion exhibited clinical features grouped across six broadly defined phenotypic domains, including intellectual disability/developmental delay (ID/DD), behavioral, psychiatric, nervous system, congenital, and growth/skeletal abnormalities (**Fig. 2A, Table S1A-B**). Probands also showed a higher number of childhood developmental and behavioral features (i.e., increased complexity scores, see *Methods*) compared to their siblings and cousins, while carrier siblings and cousins manifested more features than non-carriers (**Fig. 2A**). Parents who transmitted the deletion ("carrier parents") often manifested milder cognitive or psychiatric phenotypes (**Fig. 2B)**. Probands had a 1.98 SD decrease in non-verbal IQ ($p=2.13\times10^{-5}$) and a 1.91 SD increase in SRS ($p=2.59\times10^{-7}$) compared with their carrier parents (**Fig. 2C, Table S2A**). The average IQ score among 16p12.1 deletion probands was 1.06 SD lower than all probands ascertained for autism from the Simons Simplex Collection[34] (SSC) ($p=0.004$). The average SRS of 16p12.1 deletion probands was 0.96 SD higher than probands with 16p11.2 deletions or duplications from the Simons Searchlight cohort ($p=8.31\times10^{-6}$) and 0.38 SD higher than SSC probands ($p=6.39\times10^{-3}$) (**Fig. 2C, Table S2A**). Beyond psychiatric traits, 16p12.1 deletion probands also showed decreased head size ($p=0.001$) and increased body mass index (BMI, $p=0.009$) (**Fig. 2C, Table S2A**). Finally, consistent with their ascertainment, probands exhibited significant delays in several developmental milestones[36] compared to their siblings ($p<0.05$) (**Fig. 2D, Table S2B**). Thus, our cohort represents families ascertained for probands who exhibit a range of developmental features, including more severe IQ and social responsiveness deficits than probands ascertained for autism or the 16p11.2 deletion.

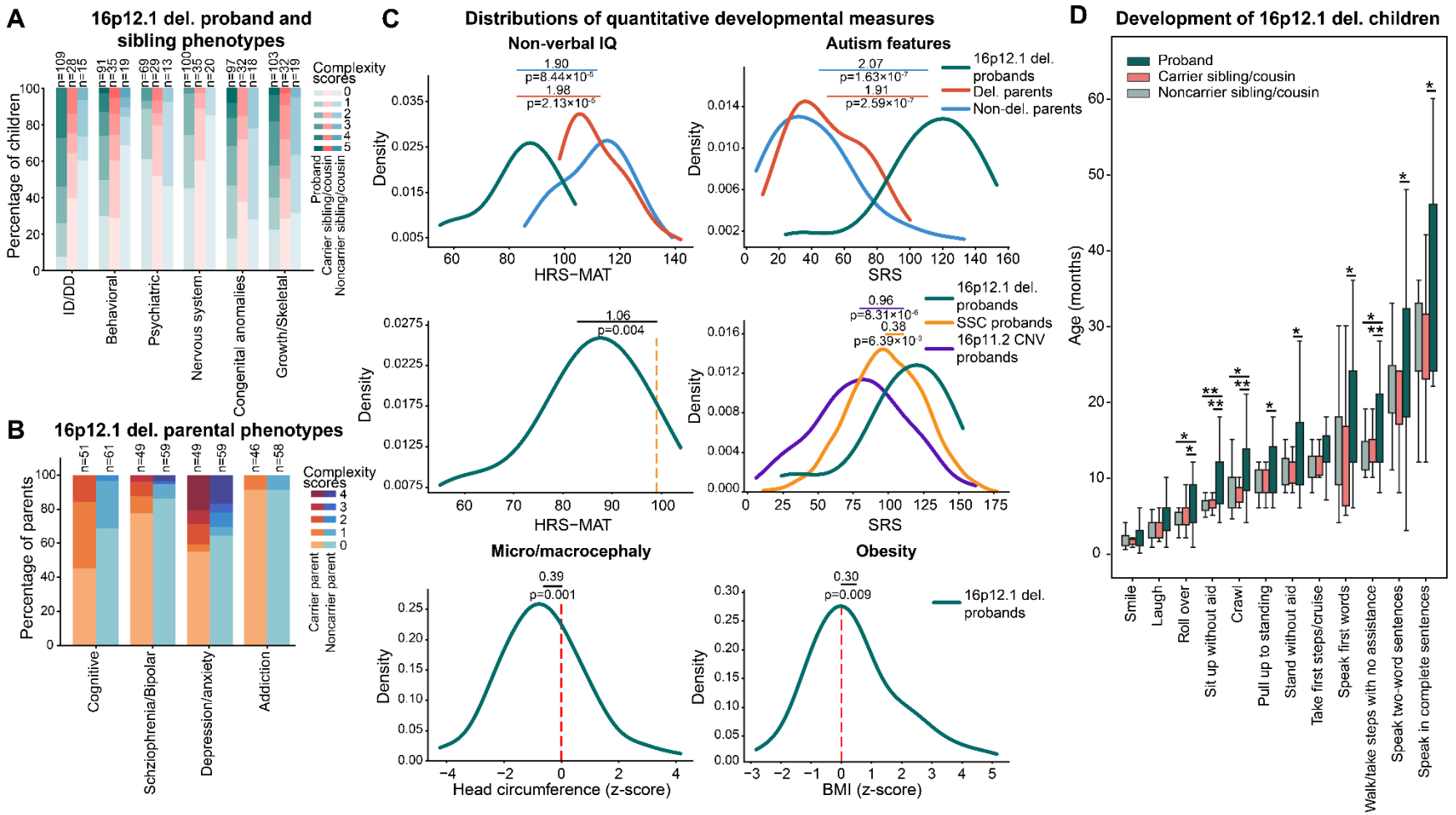


Figure 2. Variably expressive phenotypes of family members with the 16p12.1 deletion. (A) Distribution of complexity scores for six phenotypic domains in probands (n=69-109), carrier siblings and cousins (n=28-35), and noncarrier siblings and cousins (n=13-20) in 16p12.1 deletion families (numbers vary due to clinical data availability). Complexity scores were determined by identifying the number of clinical features manifested within each phenotypic domain (see Methods). **(B)** Distribution of complexity scores for four phenotypic domains in carrier parents (orange, n=46-51, orange) and non-carrier parents (blue, n=58-61) of 16p12.1 deletion probands. **(C)** Distributions of quantitative phenotypes observed in 16p12.1 deletion probands. Top plots

176

show the distribution of non-verbal IQ (HRS-MAT) and social responsiveness scores (SRS) in probands (green, n=10-27) compared to carrier (red, n=17-21) and non-carrier parents (blue, n=20-26). Middle plots compare the same scores in probands to the score for probands in the SSC cohort (SRS n=2,844, yellow; HRS-MAT mean derived from [34]) and probands with the 16p11.2 deletion or duplication from Simons Searchlight (n=139, purple). Bottom plots show the distribution of head circumference (n=64) and BMI z-scores (n=67) in deletion probands; red vertical lines represent the general population mean (i.e. z-score=0). P-values from Mann Whitney tests or one-sample t-tests. Individual scores for 16p12.1 deletion probands and parents are listed in **Table S1A**. **(D)** Distribution of the age of attainment for developmental milestones in probands (n=13-33), carrier siblings and cousins (n=16-18), and noncarrier siblings and cousins (n=11-15). One-tailed t-test, *p≤0.05, **Benjamini-Hochberg FDR≤0.05.

Patterns of secondary variants within and across families

Using WGS and microarray data, we evaluated 17 classes of secondary variants, including rare coding SNVs (missense and splice variants with CADD Phred scores ≥ 25 and LOF variants), non-coding SNVs (5' untranslated region [UTR], promoter [1kb upstream of transcription start site], and enhancer [variants in regions with enhancer chromatin signatures in fetal brain] variants), rare CNVs (deletions and duplications), and short tandem repeat expansions (STRs, defined as repeat length ≥2SD than the cohort mean), a subset of which disrupted genes under evolutionary constraint[37] (defined as LOEUF<0.35 and referred to as "(LF)" variants) (**Table S1A**). We also calculated polygenic risk scores (PRS) for four psychiatric features, including schizophrenia, intelligence, educational attainment, and autism[38–41]. These secondary variants could contribute to independent diagnoses from the 16p12.1 deletion, additively contribute to the same phenotypes as the deletion, or synergistically modify the phenotypes of the deletion. We first assessed whether probands carried additional pathogenic CNVs[33] or secondary variants that were also present in ClinVar[42] or in genes present in clinically relevant databases, such as Online Mendelian Inheritance in Man[43] (OMIM), Developmental Brain Disorder database[44] (DBD), and SFARI Gene[45]. Overall, 58% of probands (57/99) had at least one such variant, including 19% (19/99) of probands who had ClinVar-defined pathogenic variants (**Fig. S1A, Table S1D**). A subset of these cases represented probands with multiple genetic diagnoses[4]. For instance, one proband had a loss-of-function (LOF) variant in *KMT2A* and manifested Wiedemann-Steiner syndrome features, including ID/DD, dysmorphic features, and hypertrichosis[46]. Another proband with an LOF variant in *DMD* showed expected hypotonia and muscular abnormalities as well as ID/DD and craniofacial defects[28]. Additionally, we found 17 probands with STR expansions in spinocerebellar ataxia genes[47] such as *ATXN7* and *CACNA1A*[45]. Although these probands had fewer repeats than the clinical threshold for ataxia, 64.7% (11/17) of them manifested nervous system phenotypes. We also identified a deleterious missense variant in *POLR3E* on the non-deleted allele in a proband with global developmental delay and multiple congenital defects (such as bilateral club feet and natal teeth).

We next sought to identify patterns of secondary variants in probands compared to their parents (**Fig. S1B**). Probands carried more coding (LF) SNVs (union of missense, LOF, and splice variants) (p=0.041) and missense (LF) variants (p=0.017), as well as increases in non-coding SNVs in 5' UTRs (p=0.045), compared to their carrier parents (**Fig. 3A, Fig. S1C, Table**

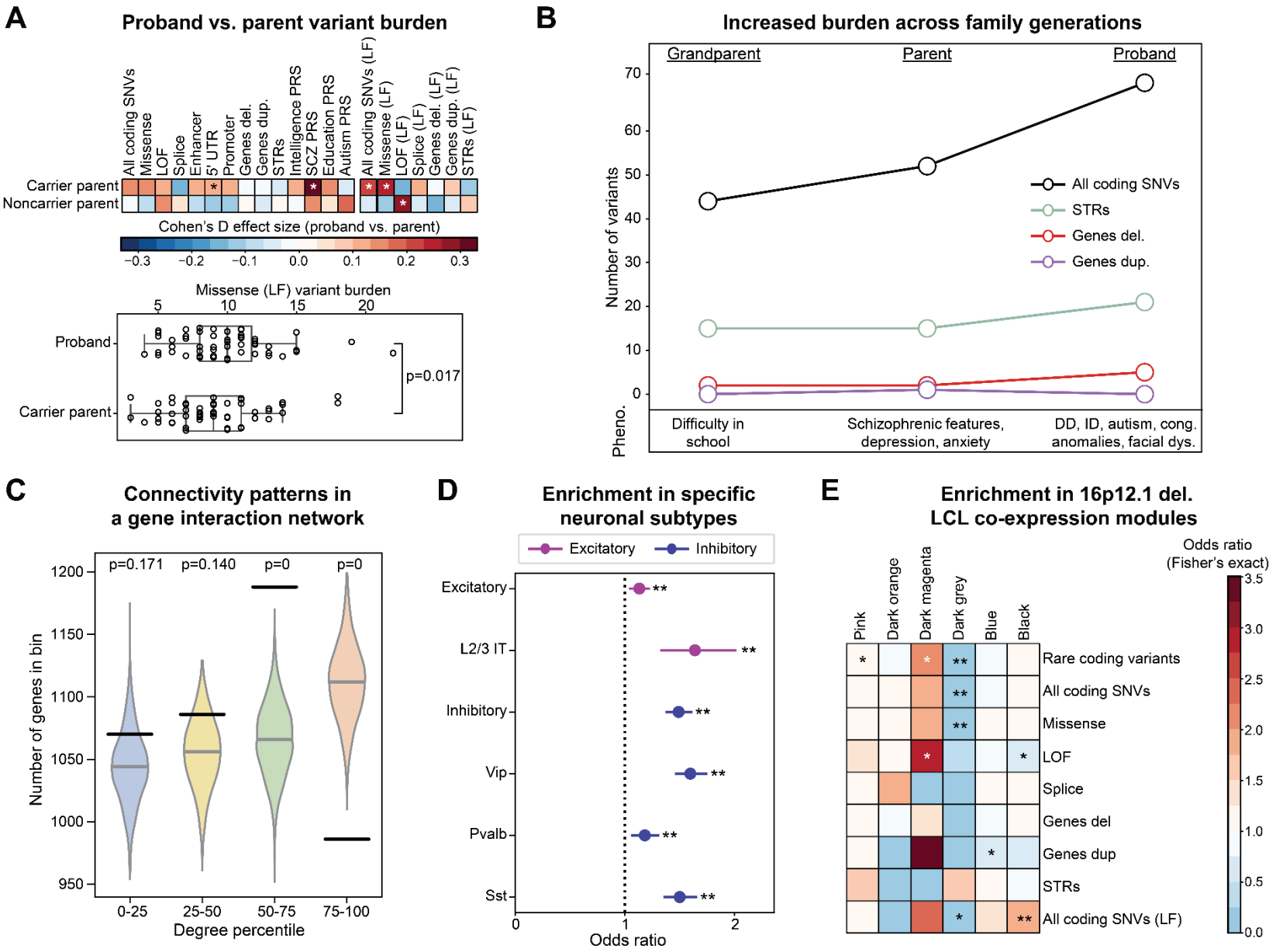


Figure 3. Secondary variants contribute to phenotypic variability within 16p12.1 deletion families. (A) Cohen's D effect sizes (top) for changes in secondary variant burden (i.e. rare variant burden or PRS) between probands and their carrier or noncarrier parents (n=49-54 pairs). *p≤0.05, paired one-tailed (rare variant classes) or two-tailed (PRS) t-test. Red indicates increased burden in probands relative to their parents. Boxplot (bottom) highlights increased burden of missense (LF) variants between probands and carrier parents. **(B)** Increased burden of rare variants corresponds with more severe clinical features across successive generations of 16p12.1 deletion carriers in a multi-generational family. **(C)** Distribution of genes by average connectivity (degree) within a brain-specific interaction network, binned into quartiles from 1000 simulations of randomly selected gene sets. Black lines represent the observed number of genes with secondary variants in 16p12.1 deletion probands in each degree quartile. Empirical p-values derived from simulation distributions. **(D)** Enrichment of genes with secondary SNVs in 16p12.1 deletion probands for genes preferentially expressed in neuronal classes (excitatory and inhibitory) and sub-classes (colored by main class) in the adult motor cortex. Fisher's exact test, **Benjamini-Hochberg FDR≤0.05. Full results are listed in **Table S2E**. **(E)** Enrichment of genes with secondary variants in probands for six gene co-expression modules identified from WGCNA analysis of lymphoblastoid cell lines (LCL) from individuals with the 16p12.1 deletion[54]. Fisher's exact test, *p≤0.05, **Benjamini-Hochberg FDR≤0.05.

209

S2C). Probands also carried higher schizophrenia polygenic risk than their carrier parents (p=0.009), showing that polygenic risk may also contribute to the features observed among 16p12.1 deletion probands (**Fig. 3A, Fig. S1C, Table S2C**). Except for an increase in LOF (LF) variants (p=0.039), no differences across other variant classes were observed in probands compared to non-carrier parents (**Fig. 3A, Table S2C**). This is consistent with a model in which the deletion and secondary variants are transmitted in specific patterns that lead to different clinical outcomes in probands. In fact, assessment of multi-generational families showed that variant burden tends to compound over generations towards increased phenotypic severity. For example, in one multi-generational family, the carrier grandparent had mild cognitive features, while the carrier parent manifested psychiatric features and the proband had neurodevelopmental features (**Fig. 3B**). This increase in phenotypic severity corresponded with an increase in the burden of multiple rare variant classes across generations, akin to the genetic anticipation observed for certain Mendelian disorders.

We further profiled the putative functions of secondary variants and found that missense variants were enriched for brain-expressed, constrained, and post-synaptic density genes (FDR≤7.41×10^{-9})[37,48], while genes with LOF variants were depleted for these gene sets (FDR≤0.007) (**Fig. S2A, Table S2G**). This suggests that LOF variants in essential genes may not be tolerated, particularly in the presence of the deletion, while less severe variants in these genes may contribute to neurodevelopmental disorders seen in probands[26]. As further evidence of this, secondary variants were enriched (empirical p=0.000) for genes with intermediate connectivity within a brain-specific gene interaction network[49,50] but depleted for genes with high network connectivity, which typically represent essential genes across species[51] (empirical p=0.000; **Fig. 3C, Fig. S2B, Table S2D**). Secondary variant genes were also preferentially expressed in several brain regions during early and mid-fetal development[52], including the frontal cortex (FDR≤0.05) and hippocampus (FDR=1.73×10^{-5}) (**Fig. S2C, Table S2H**). SNVs in particular were enriched for genes preferentially expressed across multiple neuronal classes in the adult motor cortex[53], including excitatory (FDR=0.013) and inhibitory (FDR=3.82×10^{-23}) neurons (**Fig. 3D, Table S2E**). All coding SNVs (LF) were also enriched for genes co-expressed with 16p12.1 deletion genes (black module, FDR=0.016) in lymphoblastoid cell lines (LCL) derived from a subset of 19 individuals with the deletion[54] (**Fig. 3E, Table S2F**). Overall, our results indicate that a

diverse range of biologically relevant modifiers contribute to variable phenotypes in probands with 16p12.1 deletion.

Distinct secondary variant classes contribute to specific phenotypic outcomes

We next used a series of logistic regression models to assess contributions of rare variant classes and PRS towards individual phenotypic domains. Overall, rare coding variants contributed to nervous system (logOR=0.640, p=0.032) and growth/skeletal features (logOR=0.941, p=0.004) (**Fig. 4A, Table S3A**). Specifically, STRs were associated with nervous system features (logOR=0.596, p=0.036) while SNVs were associated with growth/skeletal features (logOR=0.962, p=0.004) (**Fig. 4A, Fig. S3A, Table S3A**). In contrast, schizophrenia PRS was negatively associated with behavioral phenotypes (logOR=-0.563, p=0.046) (**Fig. 4A, Table S3A**). Combined variant models explained an average of 8% variance (McFadden's pseudo-R^2; range 2% to 14%) for each phenotypic domain, and showed better performance than models built using individual variant classes (average of 2% explained variance) (**Fig. S3B, Table S3A**). These estimates further highlighted the specificity of variant-phenotype associations; for example, STRs (LF) explained 12% of variance in nervous system defects but less than 4% of variance for other features (**Fig. S3B, Table S3A**). Orthogonal burden tests also identified fewer rare variants in enhancers, promoters, and 5' UTR elements ($p \leq 0.012$) as well as increased autism PRS (p=0.028) among probands with psychiatric features (**Fig. S3C, Table S3D**). These results suggest that the modifying roles of different secondary variant classes vary across specific phenotypes, with PRS in particular modulating behavioral features.

Linear regression models testing specific variant classes towards quantitative traits revealed negative associations of head circumference z-scores with SNVs (LF) (β=-0.357, p=0.039) (**Fig. 4B, Fig. S3A, Table S3A**). Secondary CNVs were associated with increased de Vries scores, a quantitative assessment for global developmental features[55] (deletions: β=0.288, p=0.013; duplications: β=0.246, p=0.030) (**Fig. 4B, Fig. S3A, Table S3A**). Correlation analyses revealed that intelligence and educational attainment PRS were positively correlated with head circumference (education r=0.318, p=0.026; intelligence r=0.287, p=0.045), while duplications (LF) were negatively correlated with social responsiveness deficits (r=-0.605, FDR=0.030) (**Fig. S3D**, **Table S3B**).

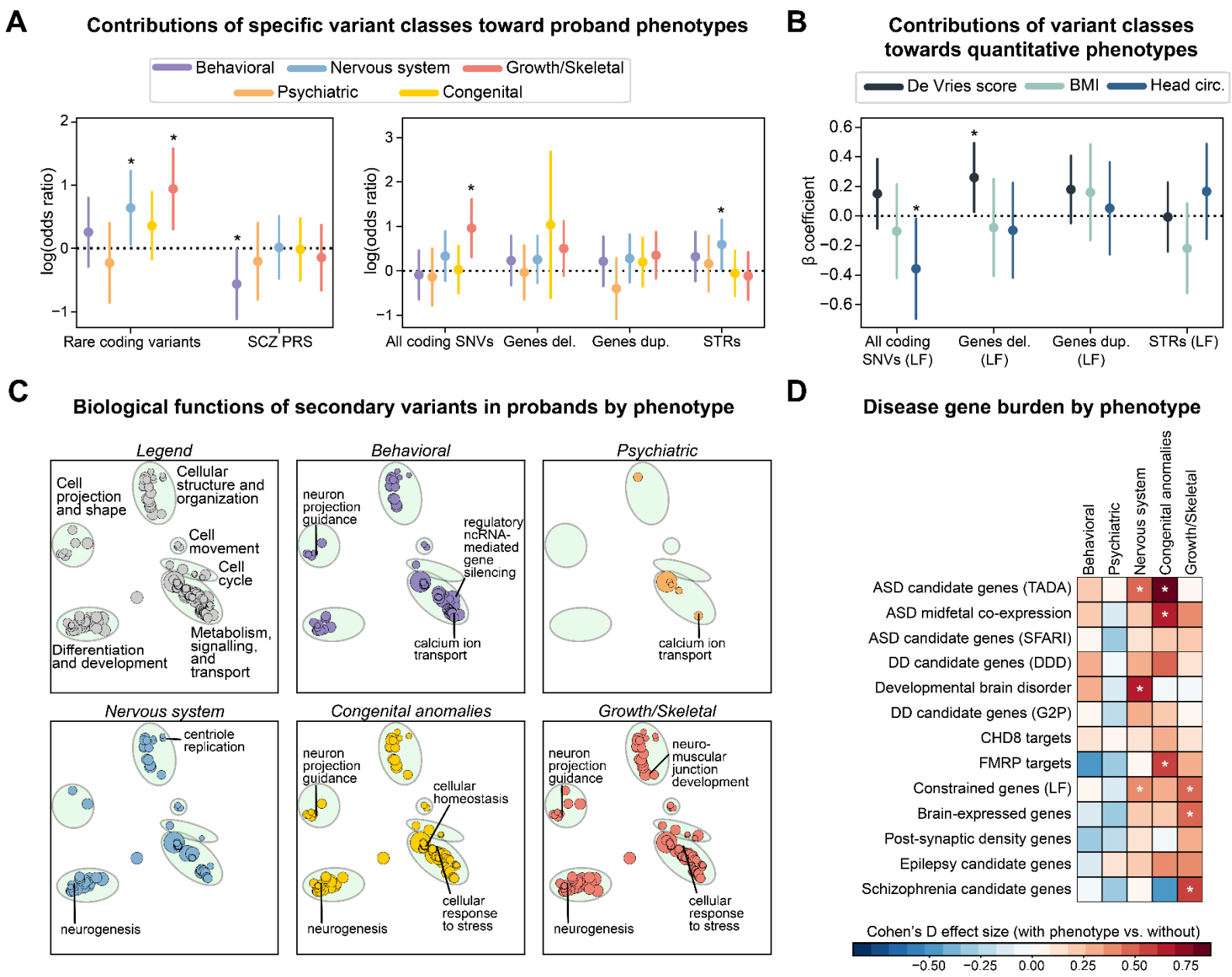


Figure 4. Secondary variant associations for phenotype domains of the 16p12.1 deletion. (A) Forest plots show log-scaled odds ratios from logistic regression models for secondary variant burden in 16p12.1 deletion probands with higher complexity scores for five phenotypic domains, compared with probands with lower complexity scores (n=47-71). *p≤0.05. Model results for variants (LF) are shown in **Fig. S3A**. **(B)** Forest plots show β coefficients from linear regression models for secondary variant burden in genes under evolutionary constrain (LF genes) towards quantitative phenotypes in deletion probands (n=43-76). *p≤0.05. Model results for variants without LF filter are shown in **Fig. S3A**. **(C)** Gene Ontology (GO) biological process terms enriched among secondary variants in probands with each phenotypic domain. Circles represent individual GO terms, clustered based on semantic similarity into broad categories (green ovals, as defined in the "legend" plot). Size of each circle represents the number of genes in each term, such that broader terms are larger. Colored circles in each plot indicate significant enrichment of the GO terms for the given phenotype. **(D)** Changes in burden of secondary variants disrupting sets of genes involved with neurodevelopmental disease and related functions (see Methods) in probands with phenotypic domains (n=23-67) compared to probands without each domain (n=12-36). *p≤0.05, one-tailed t-test.

Secondary variants in probands ascertained for the same phenotypes showed specific enrichments for biological function, including neuromuscular junction development genes in probands with growth/skeletal defects and axonogenesis-related genes in probands with behavioral and nervous system features (**Fig. 4C, Table S3B**). Additionally, probands with specific phenotypes showed increased burden of rare variants in key neuronal genes, such as *FMRP*-binding targets[48] (p=0.050) in probands with congenital anomalies and candidate autism[45] (p=0.014) and developmental brain disorder genes[44] (p=0.001) in probands with nervous system defects (**Fig. 4D, Table S3C**). Overall, our findings indicate that the disruption of distinct biological functions and molecular pathways by secondary variants may underlie specific phenotypic features of individuals with the deletion.

Differing ascertainments confer distinct genotype-phenotype patterns

Clinical outcomes associated with the same genetic variant may vary across cohorts with different ascertainments, especially for cohorts composed of affected individuals compared to those drawn from the general population[56,57]. We sought to compare the phenotypic effects of the 16p12.1 deletion in 757 individuals with complete phenotypic data, including 253 pediatric and adult carriers from the DD cohort and three independent cohorts with distinct ascertainments: SPARK (n=94), where families were ascertained for probands with autism features[58], and two population-based cohorts[59,60], the healthy-biased UK Biobank (UKB; n=250) and the hospital-derived Geisinger MyCode (MyCode; n=160) (**Fig. 1**). Assessment of UKB individuals with the deletion showed enrichment for a variety of clinical phenotypes within electronic health record (EHR) data, including circulatory, endocrine, and genitourinary features (n=3,488, FDR≤0.004) (**Fig. S4A, Table S5H**). PheWAS analysis further revealed enrichment of obesity- and kidney-related features, including type 2 diabetes and hypertension (n=99,363-255,262, $p \leq 3.48 \times 10^{-7}$) (**Fig. S4B, Table S5I**). This pattern of obesity-related features is in line with the pattern of increased BMI we observed in probands in the DD cohort (**Fig. 2C**). To more directly compare phenotype prevalence across cohorts, we next harmonized EHR and clinical questionnaire responses (**Table S4A**). As expected, the prevalence of neuropsychiatric phenotypes in both pediatric and adult deletion carriers varied across the cohorts (**Fig. 5A**, **Fig. S4C-D**, **Table S4B**). For example, we found increased anxiety symptoms in adults from the DD cohort compared to

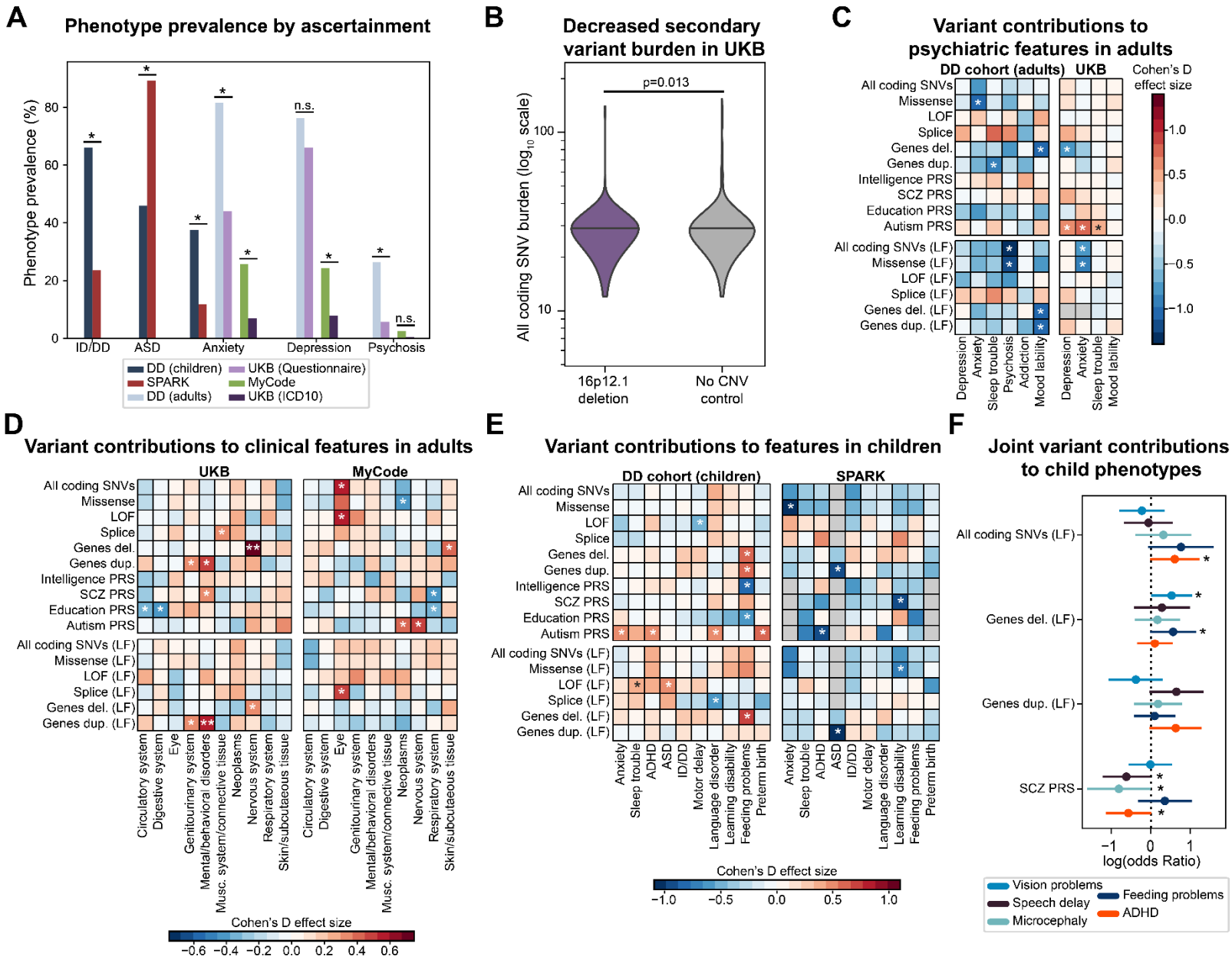


Figure 5. Effects of ascertainment on associations of 16p12.1 deletion. (A) Prevalence of phenotypes among adults and children with 16p12.1 deletion from four ascertainments: DD cohort (adults n=38, children n=93-151), SPARK (n=51-56), UK Biobank (UKB; questionnaire n=50-53, ICD10 n=217), and MyCode (n=160). Fisher's exact test, *p≤0.05. **(B)** Distribution of rare secondary SNVs in UKB individuals with 16p12.1 deletion (n=240, left) and age and sex-matched controls without large rare (>500kb) CNVs (n=2,640, right). P-value from two-tailed t-test. **(C)** Associations of secondary variant burden with select psychiatric phenotypes derived from clinical questionnaires in 16p12.1 deletion adults from DD cohort (n=24-31) and UKB (n=46-249). Two-tailed t-test, *p<0.05. **(D)** Associations of secondary variant burden with select clinical phenotypes derived from EHR data (ICD10 codes) in 16p12.1 deletion individuals from UKB (n=187-218) and MyCode (n=143-159). Two-tailed t-test, *p≤0.05. **Benjamini-Hochberg FDR≤0.05. **(E)** Associations of secondary variant burden and select developmental phenotypes in children with 16p12.1 deletion from the DD cohort (n=67-125) and SPARK (n=27-56). Two-tailed t-test, *p≤0.05. **(F)** Associations of secondary variant burden and developmental phenotypes from joint logistic models of 16p12.1 deletion children from the DD and SPARK cohorts (n=98-125). Joint models for non-LF are shown in **Fig. S4J**, and joint models for adults are shown in **Fig. S4H-I**. *p≤0.05.

304

UKB ($p=4.18\times10^{-4}$) (**Fig. 5A**, **Table S4B**), likely reflecting biases due to ascertainment for severely affected family members in the DD cohort compared to healthy volunteers in UKB[61].

We next investigated how patterns of secondary variants differed across cohorts. Adult deletion carriers in UKB showed decreased burden of additional rare coding SNVs compared to individuals without large CNVs (p=0.013) (**Fig. 5B, Table S4C**). Reduced secondary variant burden in 16p12.1 deletion carriers may explain the less severe features observed in the UKB compared to those typically observed among deletion carriers in the clinic. This trend was reversed in SPARK, where individuals with the deletion had an increased burden of SNVs (LF) compared to individuals without large CNVs (p=0.048) (**Fig. S4E, Table S4C**). Thus, we observed a higher rare variant burden in deletion carriers compared to controls in cohorts with more severe disease ascertainment (SPARK) and reduced burden in cohorts with less severe ascertainment (UKB). We also directly compared the variant burden between deletion carriers in the DD cohort to identically processed data from eight deletion carriers in the Estonian Biobank[62]. As expected, Estonian Biobank carriers showed a depletion of missense (FDR≤0.012) and non-coding SNVs (FDR≤0.019) compared with probands and carrier parents in the DD cohort (**Fig. S4F, Table S4J**).

We next assessed how the relationship of secondary variants and phenotypes varies by ascertainment by assessing the burden of variant classes in individuals with and without specific phenotypes across cohorts. We found similar trends for psychiatric features based on self-reported questionnaires for adults in both the DD cohort and UKB (**Fig. 5C, Table S4D**). For example, autism PRS was associated with depression, anxiety, and sleep disturbance (p≤0.037) in UKB (**Fig. 5C, Table S4D**). Conversely, rare variants were negatively associated with psychiatric features in both UKB and DD cohorts, including SNVs (LF) with anxiety in UKB (p=0.010) and psychosis in DD (p=0.003), and deletions with depression in UKB (p=0.016) and mood lability in DD (p=0.027). These data suggest potential opposing roles of PRS and rare variants towards specific psychiatric features. We further compared secondary variant profiles for broader groups of clinical features represented by ICD10 chapters in UKB and MyCode, in contrast to assessment of specific psychiatric features from questionnaires. In UKB, nervous system features were associated with deletions (FDR=0.007), while mental health features were associated with duplications (LF) (FDR=0.035) and schizophrenia PRS (p=0.045) (**Fig. 5D, Table S4E**). In MyCode, eye phenotypes were associated with SNVs (p=0.016), while autism

PRS was associated with both nervous system defects and cancer (p≤0.039) (**Fig. 5D, Table S4E**). These differences in genotype-phenotype patterns reflect ascertainment differences between the cohorts, potentially due to healthcare system differences, phenotyping modalities, or biases stemming from healthy volunteers versus clinical patients.

We further observed differences in children with the deletion from the DD cohort and those in SPARK. In general, both rare variants and PRS were associated with increased risk for neurodevelopmental features in DD probands (for example, duplications and deletions for feeding difficulty; p≤0.028) but decreased risk in SPARK probands (for example, decreased missense variants in individuals with anxiety; p=0.025) (**Fig. 5E, Table S4F**). In fact, individuals with ADHD had increased autism PRS in the DD cohort (p=0.017) but decreased autism PRS risk in SPARK (p=0.035) (**Fig. 5E, Table S4F**). This trend reflects the influence of ascertainment towards variant-phenotype associations, where secondary variants may not show the expected associations in highly ascertained cohorts due to saturated genetic risk for the ascertained phenotype, such as ADHD and autism PRS in SPARK (**Fig. S4G**). Overall, we found marked differences in secondary variant-phenotype associations between cohorts (**Fig. 6C-E**), which potentially explains the variable phenotypic trajectories of the deletion observed across ascertainments.

We finally combined individuals with the 16p12.1 deletion across cohorts and developed logistic regression models to identify variant-phenotype associations independent of ascertainment bias. We found 12 associations among children in the combined SPARK and DD cohorts (n=84-125), including SNVs with preterm birth (logOR=1.24, p=0.039), deletions with vision problems (logOR=0.878, p=0.018), and SNVs (LF) with ADHD (logOR=0.602, p=0.049) (**Fig. 5F**, **Fig. S4J**, **Table S4G**). Across adults in the DD, UKB, and MyCode cohorts (n=331), duplications were associated with anxiety (logOR=0.092, p=0.004) (**Fig. S4H, Table S4G**). When examining EHR-derived features in UKB and MyCode (n=321), duplications were associated with circulatory system features (logOR=0.278, p=0.037) and deletions were associated with nervous system (logOR=0.352, p=0.016) and skin/subcutaneous tissue (logOR=0.336, p=0.020) phenotypes (**Fig. S4I**, **Table S4G**). Thus, leveraging combined data from multiple cohorts allowed for the increased statistical power necessary for deriving robust variant-phenotype associations across ascertainments.

Differing contributions of secondary variants by primary variant ascertainment

To extend our findings beyond the 16p12.1 deletion, we assessed contributions of secondary variants towards developmental, cognitive, and behavioral features among 1,479 probands with different rare pathogenic CNVs or SNVs in known neurodevelopmental genes who were ascertained for the same disorder, i.e. autism (**Fig. 1**). We first assessed 128 probands with reciprocal 16p11.2 deletions (n=91) and duplications (n=37) in the Simon Searchlight cohort[63] and found eight variant-phenotype associations using linear regression models (**Fig. 7A**, **Table S6A**). Among 16p11.2 deletion probands, schizophrenia PRS contributed to higher full-scale IQ (ß=0.343, p=0.020), while secondary deletions contributed to decreased IQ (ß=-0.283, p=0.040), similar to previous findings[11] (**Fig. 6A, Fig. S5A, Table S5A**). Different trends were observed in 16p11.2 duplication probands; deletions and duplications were negatively associated with autism behavioral features (BSI; duplications: ß=-0.497, p=0.022; deletions: ß=-0.432, p=0.037) and duplications (LF) were negatively associated with SRS (ß=-0.701, p=0.002) (**Fig. 6A, Fig. S5A, Table S5A)**. Orthogonal correlation analyses revealed several other trends, including opposing effects of secondary duplications towards BSI (16p11.2 deletion individuals: r=0.241, p=0.023; 16p11.2 duplication individuals: r=-0.391, p=0.019) (**Fig. S5B, Table S5E**).

We next assessed 214 probands with a more heterogeneous set of large CNVs, including pathogenic deletions and duplications[33] at 15q13.3, 3q29, and 16p13.11, from SSC[64]. Among probands with large deletions, linear regression models uncovered negative associations between secondary duplications (LF) with BMI (ß =-0.275, p=0.049), while secondary deletions (LF) were associated with decreased IQ in probands with large duplications (ß =-0.255, p=0.021) (**Fig. 6A, Fig. S6A, Table S5A)**. Correlation analyses revealed additional associations, including SNVs with coordination impairment in probands with large duplications (DCDQ, r=0.178, p=0.042) and decreased SRS in those with large deletions (r=-0.246, p=0.035) (**Fig. S6B, Table S5F**). We further assessed 1,237 SSC probands with SNVs disrupting canonical neurodevelopmental genes[44], such as *CHD8*, *DYRK1A*, *SCN1A*, and *PTEN*. We again identified a negative association for deletions (LF) with IQ (ß =-0.154, $p=5.23\times10^{-5}$), while STRs were associated with increased IQ (ß=0.093, p=0.015) (**Fig. 6A, Fig. S6A, Table S5A**). Correlation analyses uncovered additional effects, such as externalizing behavior (r=-0.111, p=0.001) and repetitive behavior (r=-0.083, p=0.017) negatively correlating with educational attainment PRS (**Fig. S6B, Table S5F**). Thus, we found some consistent patterns across primary variant

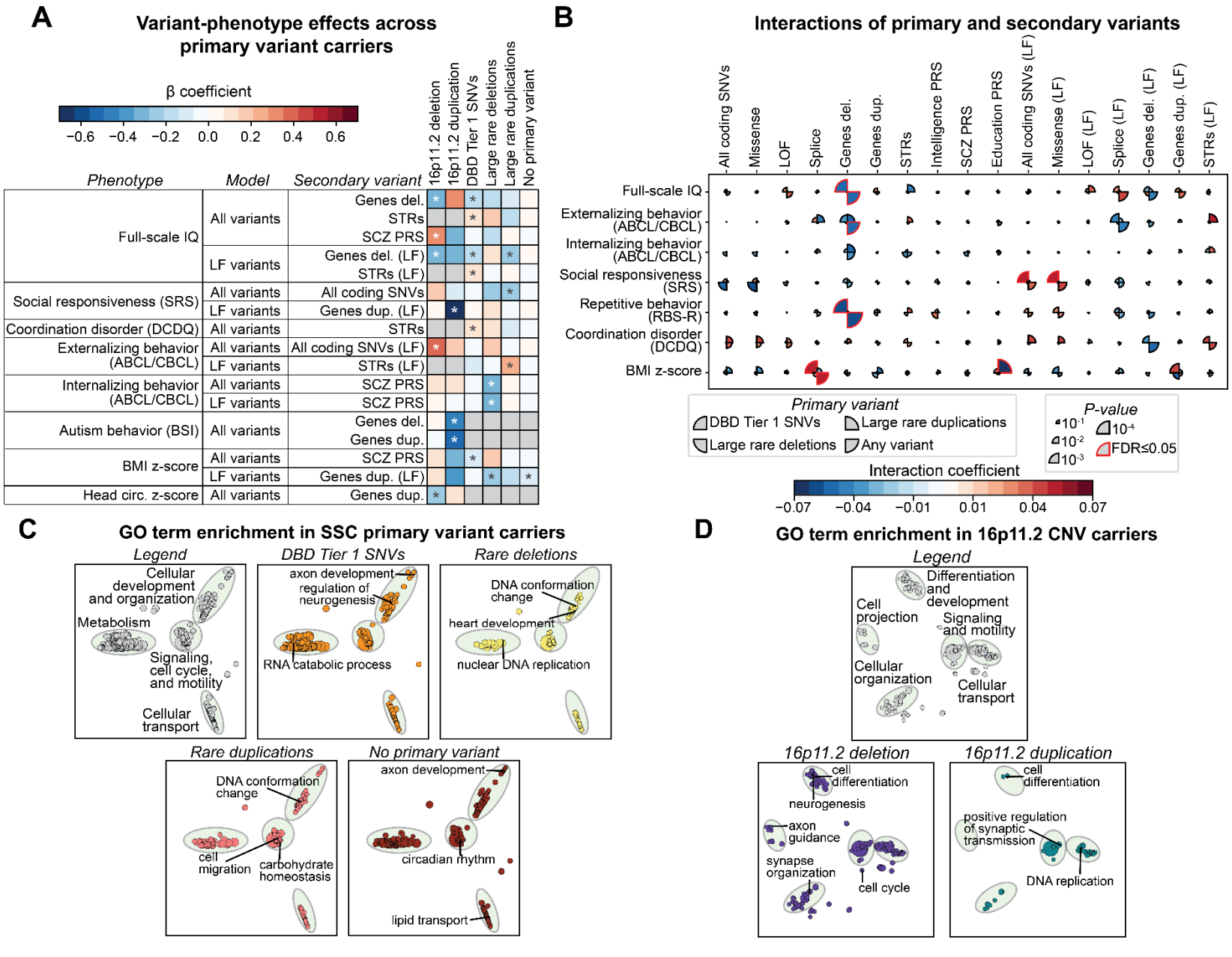


Figure 6. Secondary variant associations in probands with primary variants. (A) Heatmap shows β coefficients from select linear regression models for secondary variant burden (y-axis, third column) towards quantitative developmental phenotypes (y-axis, first column) in probands from SSC and Simons Searchlight cohorts with different classes of primary variants (x-axis) (n=21-660). *p≤0.05. **(B)** β coefficients from linear regression models examining interactions between primary variants (pie chart slices) and specific secondary variant classes (x-axis) towards quantitative phenotypes (y-axis) in SSC probands (n=1,597-2,591). Color of pie chart slices indicate interaction coefficients, and size of pie chart slices indicate p-value for strength of interaction coefficient. Red highlights indicate Benjamini-Hochberg FDR ≤0.05 **(C-D)** Gene Ontology (GO) biological process terms enriched among secondary variants observed in **(C)** probands with different classes of primary variants from the SSC cohort and **(D)** probands with 16p11.2 deletions and duplications from the Searchlight cohort. Circles represent individual GO terms, clustered based on semantic similarity into broad categories (green ovals, as defined in the two "legend" plots). Size of each circle represents the number of genes in each term, such that broader terms are larger. Colored circles in each plot indicate significant enrichment of the GO terms for the given primary variant.

399

ascertainments, such as negative effects of rare deletions on IQ, while secondary variant effects on other features were more dependent on the primary variant context.

We additionally examined secondary variants in 1,084 SSC probands without primary variants in the above categories to assess the role of the genetic background outside of a primary variant context. The only observed association from regression analysis was for duplications (LF) and lower BMI (ß=-0.087, p=0.031) (**Fig. 6A, Table S5A**). The paucity of associations in the absence of primary variants suggests that secondary variant classes mostly exert their effects through interactions with primary variants instead of contributing directly towards disease phenotypes. To assess this, we used linear models to identify interactions between primary and secondary variants towards clinical features. We found ten instances of multiplicative interactions, including primary SNVs and secondary SNVs (LF) towards SRS (ß=0.054, FDR=0.034) as well as primary SNVs and secondary deletions towards full-scale IQ (ß=-0.058, FDR=0.020) and repetitive behavior (ß=-0.062, FDR=0.011) (**Fig. 6B**, **Table S5B**). Notably, these interactions tended to be primary variant-specific, further supporting the hypothesis that secondary variant effects are influenced by primary variant context.

The relevance of primary variant context was further evident when we assessed the biological functions of secondary variants (**Fig. 6C-D)**. For example, secondary variants in probands with primary SNVs showed specific enrichments for neuronal development and cell-to-cell signaling, while probands with primary deletions showed enrichments for DNA repair and replication (**Table S5C**). Secondary variants in 16p11.2 duplication probands were enriched for DNA replication and synaptic transmission regulation, while variants in probands with the reciprocal deletion were enriched for cell cycle regulation and synapse organization (**Table S5D**). In fact, multiple genes within the 16p11.2 deletion have similar molecular functions (i.e. *MAPK3* and cell cycle regulation[65]), and many of the same GO enrichments, including neuronal differentiation and projection, were identified among differentially expressed genes in animal models of genes within the 16p11.2 region[66] (**Table S5D**). We therefore posit that modifier variants influence developmental features by acting additively or synergistically in molecular pathways disrupted by the primary variant, further underscoring the importance of primary variant context[67].

DISCUSSION

Our comprehensive analysis of 2,252 individuals with primary variants from several diverse cohorts allowed us to find strong evidence that modifier variants confer distinct risks towards different developmental and clinical phenotypes. These effects are contingent upon the context of the primary variant, secondary variant class and function, phenotype of interest, and cohort ascertainment. Our results emphasize the importance of assessing a full spectrum of genomic variants in a variety of contexts to unravel the etiology of heterogeneous clinical features observed in individuals with the same primary variants.

Several of our results expand on previous work to identify mechanisms for variable expressivity of pathogenic variants and may help refine the broader roles of modifier variants in complex disorders. First, we identified roles for a wide set of rare variants towards developmental features of the 16p12.1 deletion, including noncoding variants and STRs. These findings expand previous definitions of secondary variants beyond additional CNVs or rare coding SNVs[28,33]. Second, we expanded the role for PRS towards various developmental and psychiatric phenotypes in individuals with pathogenic CNVs. These findings are in line with recent studies that have identified roles for PRS as modifiers of specific phenotypes of pathogenic CNVs, such as BMI z-scores for 16p11.2 CNVs[10]. Third, we observed cases of compounding variant burden across generations of 16p12.1 deletion carriers, which could explain our previous findings correlating rare variant burden with family history of psychiatric disorders[28,54]. This phenomenon could be attributed to assortative mating on psychiatric features between deletion and non-deletion parents; in fact, we recently reported that 16p12.1 deletion spouse pairs show strong correlations for psychiatric disorders, which may lead to increasing genetic risk over generations[68]. Fourth, while disease-relevant secondary variants contributed to multiple genetic diagnoses, we did not find any instance where a single variant solely accounted for all phenotypes observed in a proband, suggesting that secondary variants modify effects of the deletion.

Ascertainment bias can preclude a more thorough assessment of a full spectrum of phenotypes due to primary variants, including subtler or progressive effects, as deeper evaluations are typically restricted to individuals with specific disorders[69,70]. We therefore examined the effects of cohort ascertainment within a single primary variant-specific context. While the 16p12.1 deletion contributed to clinical outcomes across disease-ascertained and

general population cohorts, the specific phenotypic trajectories and the influence of secondary variants both differed substantially across ascertainments. Our findings have several implications, as management of primary variant-related symptoms may differ in individuals evaluated for severe developmental features versus those with other medical features, where the variant may first present as an incidental finding[19]. Thus, a shift in treatment focus from just the primary variant to all variants in an affected individual could allow for more effective management of individuals who carry primary variants.

The observed variability among secondary variant-phenotype patterns in each cohort makes it difficult to identify consistent patterns across ascertainments, limiting the generalizability of variant association studies conducted in a single cohort. In fact, the only consistent trends we observed across primary variants, such as 16p11.2 deletion and rare disease-associated SNVs, were for reduced IQ correlating with increased rare secondary variants, mirroring previous studies[11]. Joint models that integrate data across cohorts with appropriate covariates can be used to overcome this ascertainment bias; for example, we found several significant associations using joint models of 16p12.1 deletion carriers, including rare SNV burden towards ADHD and speech delay in affected children. More broadly across primary variant and ascertainments, we observed general trends for more PRS effects towards psychiatric features, such as autism PRS and ADHD in DD children and SCZ PRS for mental/behavioral disorders in UKB, and more rare variant effects towards cognitive features, such as rare deletions (LF) with full-scale IQ in SSC probands with disease-associated SNVs and 16p11.2 deletion probands. Both trends mirror previous variant-phenotype associations in individuals with autism outside of a primary variant context[71]. However, some exceptions to these patterns exist: SSC probands with rare duplications and 16p11.2 duplication probands show higher effects of rare secondary variants towards psychiatric phenotypes, including negative associations of splice variants with externalizing behavior in both groups. Further, 16p12.1 deletion carriers from SPARK and 16p11.2 deletion probands show greater effects of PRS towards cognitive features, such as the positive association of SCZ PRS and full-scale IQ in 16p11.2 deletion probands and a negative association of SCZ PRS with learning disability in SPARK, potentially a facet of autism-specific genetic etiology for cognitive features. Therefore, when describing genotype-phenotype associations in a primary variant context, future studies should strive to assess

multiple independent cohorts with different ascertainments to determine the extent that ascertainment could bias their results.

Despite assessing the contributions of multiple secondary variant classes towards specific clinical features of pathogenic variants, much of the genetic etiology for these features is still not accounted for in our study. Several factors could account for the unexplained variance, including environmental factors or additional variant classes such as inversions, as well as those that could explain ascertainment variability, including population-specific effects in the genetic background. Another under-studied source of the unexplained variance could be non-additive interactions between variants. Only a small number of synergistic variant interactions have been identified to date in complex genomic disorders[72], and very large cohorts will be required to quantify the effects of these interactions towards clinical features[73]. Here, we identified non-additive effects of primary and secondary variants among children ascertained for autism. Molecular studies could help unravel the mechanisms by which modifier variants interact with primary variants to influence their phenotypes. For instance, we previously found 11 cases where secondary variants synergistically altered the expression of genes dysregulated by the deletion in patient-derived LCL models[54]. While the overall effects of such interactions will likely explain only a portion of the unexplained variance, they may play an outsized role in CNV disorders due to the potential for interactions among multiple genes within the primary variant[26].

Overall, we identified family-, phenotype-, ascertainment-, and primary variant-specific patterns of secondary variants that influence the variable expressivity of the 16p12.1 deletion and other primary variants. Our study emphasizes the complexity of neurodevelopmental disorders even after a causal variant is identified, suggestive of an oligogenic model for disease pathogenicity[74]. For researchers and clinicians alike, our study highlights the importance of understanding the influence of cohort ascertainment and thoroughly investigating genomic variants with smaller effect sizes. The complexity of the 16p12.1 deletion and other genomic disorders calls for personalized medicine approaches that fully account for individual-level phenotypic presentation, family history, and genome-wide variant profile towards counseling, management, and potential treatment.

Limitations of this study

One limitation of our study is the relatively low sample size of families with the 16p12.1 deletion and other primary variants. While this study represents one of the largest cohorts of individuals with the same pathogenic variant to date and is well-powered for assessing changes in rare coding SNV burden among deletion carriers, the study is under-powered for identifying enrichments of individual variants or genes towards specific phenotypes. Additionally, while our study captures major themes regarding variable expressivity of pathogenic variants, some specific associations have only nominal significance. Larger cohorts will allow for further study of these trends and could uncover roles for specific genes or molecular pathways towards clinical features. Finally, while we were able to leverage data from 773 16p12.1 deletion individuals from multiple ascertainments, differences in genotyping and phenotyping methods precluded direct comparisons between cohorts.

AUTHOR CONTRIBUTIONS

M.J., C.S., A.T., L.P., and S.G. designed the study and analyses. C.S., L.P., E.H., and L.R. recruited patients and conducted interviews, harmonized phenotypic data from interviews and medical records, and extracted DNA from blood for WGS sequencing. C.T. and C.L.M. assisted with collection and analysis of quantitative phenotypic data. Other authors provided de-identified DNA, blood samples, or genomic and phenotypic data of 16p12.1 deletion families to the study. M.J., C.S., A.T., D.B., and V.K.P. designed bioinformatics pipelines to identify variants, uniformly processed sequencing data from 16p12.1 deletion and external cohorts, and performed all statistical, enrichment, and modeling analysis. V.R. and H.S. assisted with design of the PRS calculations and modeling approaches. M.J., C.S., A.T., L.P, and S.G. wrote the manuscript with approval from all authors.

ACKNOWLEDGEMENTS

This work was supported by NIH R01-GM121907 and resources from the Huck Institutes of the Life Sciences to S.G. M.J. and C.S. were supported by NIH T32-GM102057. A.T. was supported by NIH T32-LM012415. L.P. was supported by Fulbright Commission Uruguay-ANII. A.R. was supported by grants from the Swiss National Science Foundation 31003A_182632. S.B. was supported by the NIHR Manchester Biomedical Research Centre (NIHR203308). We thank

Craig Praul and the Penn State Genomics Core Facility for assistance with designing the WGS sequencing strategy, Abby Hare-Harris (Geisinger ADMI) for assistance with RedCap analysis of quantitative phenotypic data, Veera Rajagopal (Regenron Pharmaceuticals) and Bertrand Isidor (CHU Nantes) for useful comments on the manuscript, and Jianyu Yang, Sarah Dwiekat, and Edmundo Torres-Rodriguez (Penn State) for assistance with curating annotation data for variant analysis. We are grateful to all of the families in each cohort (DD, SPARK, Simons Searchlight, SSC, MyCode, and UK Biobank) as well as clinical sites and staff who participated in the study. We thank the SSC principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman) as well as the Simons Searchlight Consortium. We appreciate obtaining access to genomic and phenotypic data on SFARI Base. Approved researchers can obtain the SPARK, SSC, and Simons Searchlight population data sets described in this study by applying at https://www.base.sfari.org. This research has been conducted using data from UK Biobank. More information about UK Biobank is available at https://www.ukbiobank.ac.uk/.

DECLARATION OF INTERESTS

The authors declare no competing interests.

METHODS

16p12.1 deletion developmental delay cohort description

We analyzed genomic and phenotypic data from a cohort of 452 individuals belonging to 132 families with the 16p12.1 deletion, which we refer to as the Developmental Delay cohort ("DD cohort"). (**Fig. 1, Table S1A**). These families comprised single probands (n=14), parent-child pairs (n=13), complete trios (n=97), and extended families, including eight families with three generations and 22 with multiple affected children. The deletion was identified through prior clinical diagnostic tests for ID/DD or other developmental disorders in probands. We note that 10 individuals from eight families with the 16p12.1 deletion representing an unselected population from the Estonian BioBank were included as a comparison group within the DD cohort[62]. Whole genome sequencing was performed on 287 individuals (107 probands), including the eight Estonian BioBank families, and microarray experiments were performed for 368 individuals. Informed consent was obtained from families recruited directly according to a protocol approved by the Pennsylvania State University Institutional Review Board (IRB #STUDY00000278), while de-identified information was obtained from families recruited through clinics according to another approved protocol (IRB #STUDY00017269). A list of all individuals in the DD cohort, including familial membership and 16p12.1 deletion status, is available in **Table S1A**. Note that some information that could be construed as personally identifiable is summarized in **Table S1A** (i.e., age ranges instead of age values); full datasets are available upon request from the authors for purposes of reproducibility (i.e., for the model covariates). We also note that sample and family identifiers used in **Table S1A** are specific for this study and were not known outside of the research group. Power calculations for detecting changes in rare variant burden within deletion family members (**Fig. S1B**) were based on estimated effect sizes of burden differences between 16p12.1 deletion probands and parents (coding SNVs and CNVs) or between autism probands and parents (non-coding SNVs and STRs) from previous studies[28,75,76].

Phenotypic analysis

Individual-level phenotypic data described below (phenotypic domain complexity scores, quantitate phenotypes, and family history status) are available in **Table S1A**.

(a) Collection and analysis of clinical features: We collected detailed medical history and used clinician-, guardian- (for children), or self- (for adults) reported standardized questionnaires to assess developmental phenotypes in children (average age=10.1 years) and psychiatric features in adults. Questionnaires for children assessed neuropsychiatric and developmental features, anthropomorphic measures, congenital abnormalities in multiple organ systems, and family history of medical or psychiatric disorders. Phone surveys were conducted to fill in missing information, and families were recontacted every 3-4 years to track longitudinal data and to note any later-onset clinical features. We analyzed phenotypes for each individual by calculating "complexity scores" for clinical features within specific domains, as well as measured specific quantitative features (see "Assessment of quantitative phenotypes" below).

For children, we first grouped clinical features into six broadly defined phenotypic domains: (i) ID/DD, (ii) behavioral phenotypes, (iii) psychiatric features, (iv) nervous system defects, (v) craniofacial and skeletal abnormalities, and (vi) congenital abnormalities (**Fig. 2A, Table S1A-B**). We determined complexity scores ranging from 0 to 4 or 5 for each phenotypic domain by assessing the total number of affected phenotypic sub-domains in each child. The full list of phenotypes considered for each sub-domain is available in **Table S1B**. Presence of at least one clinical feature within a sub-domain added an additional point of complexity to the total score, but additional phenotypes within the same sub-domain added no additional complexity score. For example, proband P1C_001 had three nervous system-related phenotypes (tremors, abnormal gait, and abnormal brain morphology) grouped into two sub-domains (nervous system abnormalities and nervous system morphology defects), and therefore received two points for complexity. We note that younger probands were not assessed for psychiatric domains based on the typical age of onset, such as for schizophrenia (**Table S1C**). As most probands (92%) exhibited >1 feature within the ID/DD domain, we focused on the other five phenotypic domains for downstream analysis.

For adult family members, we binned and calculated complexity scores in a similar manner for clinical features within four domains, including (i) cognitive (ID, learning difficulty), (ii) psychiatric (schizophrenia, bipolar disorder), (iii) depression/anxiety, and (iv) addiction phenotypes (**Fig. 2B**). We note that most phenotypes for early-onset behavioral and developmental features assessed in children were not examined for adult family members.

(b) Assessment of quantitative phenotypes: We performed online quantitative assessments using the Hansen Research Services Matrix Adaptive Test (HRS-MAT) for non-verbal IQ[34] and Social Responsiveness Scale (SRS) for autism-related social behavior[35] (**Fig. 2C**). HRS-MAT was self-administered to participants through an online platform, while the SRS was administered through a RedCap-based survey platform maintained by the Geisinger Autism and Developmental Medicine Institute. The SRS assessment was self-reported if the participant was over 18 years or completed by parents or guardians for children under the age of 18 years. Body Mass Index (BMI) and head circumference were obtained from medical records or self/guardian-reports or, for BMI, calculated from height and weight data obtained from medical records or self/guardian-reports. Both BMI and head circumference were converted into age- and sex-adjusted z-scores[77,78]. We further obtained SRS, BMI, and head circumference z-scores for probands in the SRS and Simons Searchlight cohorts (see below), while the mean HRS-MAT score in SSC probands was obtained from Hansen, 2019 [34]. Differences in phenotype distributions between groups of 16p12.1 deletion family members and between sets of probands with different primary variants were calculated using one- and two-tailed Mann Whitney tests, respectively (**Table S2A**). Differences of proband scores from a defined mean were calculated using one-tailed one-sample t-tests (**Table S2A**).

(c) Developmental milestones: We assessed the achievement of developmental milestones in children from the DD cohort based on CDC guidelines[36]. Parents/guardians of children reported the ages at which children achieved 12 milestones, including age first smiled, rolled over, crawled, walked, and spoke. Age of milestone attainment for all available samples is reported in **Table S1A**. Differences in milestone achievement between probands and their siblings/cousins were assessed using one-tailed t-tests (**Table S2B**).

DNA extraction and whole-genome sequencing

We performed DNA extraction and whole-genome sequencing on 287 individuals in the DD cohort (**Table S1A)**. Genomic DNA was extracted from peripheral blood samples from some participants using the QIAamp DNA Blood Maxi extraction kit (Qiagen, Hilden, Germany) while clinical collaborators sent isolated DNA from other participants. Illumina TruSeq DNA PCR-free libraries (San Diego, CA, USA) were constructed for 150bp paired-end whole-genome sequencing using Illumina HiSeq X by Macrogen Labs (Rockville, MD, USA). Samples were

sequenced at an average 35.7X coverage, or 716.2 M reads/sample, with 94.9% of reads mapping to the human genome. After processing for quality control using Trimmomatic[79] (leading:5, trailing:5, and slidingwindow:4:20 parameters), sequences were aligned to the hg19 reference genome using BWA v.0.7.13[80], and sorted and indexed using Samtools v.1.9[81]. We note that sequencing data from 163 individuals was described previously[54,68].

We used SNP microarrays for copy-number variant validation and genotyping experiments (i.e. CNV calling and polygenic risk score calculation) for 368 individuals. Samples were run on Illumina OmniExpress 24 v.1.1 microarrays at Northwest Genomics Center in the University of Washington (Seattle, WA, USA). We note that microarray data of 208 individuals was described previously[28,54,68].

Identification and annotation of single-nucleotide variants

We identified SNVs and small indels using the GATK Best Practices pipeline[82], followed by quality control and extensive variant and gene-level annotations. Duplicate removal with PicardTools was followed by base-pair quality score recalibration and variant calling for each sample using GATK HaplotypeCaller v.3.8. We then merged calls from all samples using GATK GenotypeGVCFs v.4.0.11, performed variant quality score recalibration to finalize variant calls, and used Vcfanno to annotate variants with GnomAD frequency[83,84]. All calls were filtered for QUAL ≥50, allele balance between 0.25 and 0.75 or ≥0.9, read depth ≥8, QUAL/alternative read depth ≥1.5, GnomAD frequency ≤0.1% (or not present), and intracohort frequency ≤10 to account for technical differences between our data and GnomAD. We annotated coding and noncoding variants within genes from GENCODE[85] v19 using ANNOVAR[86] and Vcfanno[83] as follows: *(a) Coding SNVs*: Rare coding variants were filtered for loss-of-function (LOF), missense, or splicing exonic variants in protein-coding genes, and annotated with CADD Phred-like scores, presence in ClinVar database, the gene-level pathogenicity metric LOEUF, and gene-level phenotype associations using OMIM[37,42,43,87]. Missense and splice variants were filtered to include only those with a CADD score ≥25. *(b) Noncoding SNVs*: All rare variants located 1kbp upstream of a gene transcription start site were classified as promoter variants, while genes within the 5' UTR were classified as 5' UTR variants. Fetal brain-active enhancer regions were identified using chromatin state data from the Roadmap Epigenomics consortium[88] (states 6, 7,

and 12 in the fetal brain), and rare variants in those regions were classified as enhancer variants. Rare SNV burden for all available individuals are listed in **Table S1A**.

Identification of copy-number variants and short tandem repeats

CNVs were called from both microarray data using PennCNV[89] and WGS data using CNVnator v.0.4.1, Lumpy-sv v.0.2.13 with Smoove v.0.2.5, Delly v.1, and Manta v.1.6.0[90–93]. For CNVs >50 kbp in length, we used a union of PennCNV and CNVnator calls. For CNVs <50 kbp, we used CNVs called by least two of CNVnator, Lumpy, Manta, or Delly, defined by 50% reciprocal overlap. All CNVs were annotated for 50% reciprocal overlap with known pathogenic CNVs[33]. WGS-based CNV calls were filtered to remove calls with >50% overlap with centromeres, segmental duplications, regions of low mappability, and V(D)J recombination regions, while microarray CNVs were filtered to remove samples with >50% overlap with centromeres, telomeres, and segmental duplications[94]. Known pathogenic CNVs were excluded from this filter[33]. All CNVs were then filtered for GnomAD-SV frequency[95] <0.1% and intracohort frequency ≤10, and >50 kb CNVs were additionally filtered for <0.1% frequency in a control cohort[96]. All CNVs were finally filtered for those intersecting at least one protein-coding gene, using gene annotations from GENCODE v19[85].

We identified STR expansions from WGS data using the GangSTR v.2.4 (reference file v.13.1) and TRTools pipelines[97,98]. We filtered calls with read depth >20 and <1000, excluding reads that were not spanning and bounding the STR locus and calls with maximum likelihood estimates not within the confidence interval using dumpSTR. After merging the calls with mergeSTR, we ran dumpSTR with population level filters, including locus call rate >0.8 and departure from Hardy Weinberg equilibrium (Fisher's exact p-value) >0.00001, and removed loci that overlapped with segmental duplications. For chromosome X, the Hardy-Weinberg equilibrium p-value was calculated from female samples only. For each family, we extracted the STR loci that passed variant filtering, and used GangSTR v2.5 to call STR variants, which were used in subsequent analyses. We defined STR expansions as STR variants with lengths >2SD higher than the average of STR lengths among all individuals in our in-house cohort of individuals with WGS data at a particular locus. STR expansions spanning protein coding regions defined by GENCODE v19[85] were selected for downstream analysis. All CNV and STR genes were further annotated for pathogenicity metrics (LOEUF score[37]). The number of genes

723 affected by rare CNVs and the number of STR expansions for all individuals with available
724 WGS data are listed in **Table S1A**.
725
726 **Polygenic risk score calculations**
727 Using microarray data, we calculated polygenic risk scores for educational attainment[38],
728 intelligence[39], schizophrenia[40], and autism[41] among the samples in the DD cohort, based on
729 standardized bioinformatics pipelines for quality control[99]. We first downloaded summary
730 statistics from four recent GWAS studies of neuropsychiatric traits, and filtered SNPs for
731 imputation INFO scores >0.8 and removed duplicate and ambiguous SNPs. We then merged
732 SNP genotype data from different microarray batches together using PLINK. Initial quality
733 control removed SNPs with minor allele frequency <0.05, Hardy-Weinberg equilibrium $<1.0\times10^{-6}$,
734 and genotype rate <0.01, along with samples missing >1% of genotypes. We used the HRC-
735 1000 Genomes Imputation toolkit (https://www.well.ox.ac.uk/~wrayner/tools) to process PLINK
736 files into individual chromosomes for imputation, and VcfCooker
737 (https://genome.sph.umich.edu/wiki/VcfCooker) to convert PLINK files to VCF files.
738 Microarray-based SNPs were imputed using the TOPMed v.r2 imputation server using Eagle
739 v2.4 for phasing[100]. After imputation, VCF files were converted back to PLINK format, and
740 SNPs were again filtered with identical QC filters. Additional QC filters included removing
741 samples with >±3SD of the mean heterozygosity rate. We also selected individuals with
742 European ancestry, based on imputed genetic ancestry (calculated using Peddy v.0.4.8 with 1000
743 Genomes-based population panel) or self-reported ancestry, for downstream analysis[101]. Finally,
744 we performed strand-flipping to match SNPs in microarray data with the GWAS summary
745 statistics. To calculate PRS, we used standardized pipelines for the LDPred2 software package,
746 which uses Bayesian approaches to optimize parameters for PRS calculation[102]. Briefly, we
747 filtered the four sets of GWAS summary statistics for SNPs present in the HapMap3 dataset[103],
748 and used 1000 Genomes datasets to calculate linkage disequilibrium matrices for the SNPs[99].
749 After regressing betas or odds ratios of GWAS SNPs according to linkage disequilibrium, we
750 used the LDPred2-auto model to calculate the four PRS values for all samples with available
751 genotype data. PRS for all available individuals are listed in **Table S1A**.
752

Genotype-phenotype statistical and modeling analysis

We performed multiple analyses to compare effects of rare variants and PRS towards different phenotypic domains among 16p12.1 deletion probands or between probands and their carrier and non-carrier parents. When assessing variant effects towards phenotypic domains, we binned probands with different complexity scores for each phenotypic domain into binary groups of roughly equal sizes (i.e., probands with higher versus lower complexity scores) for logistic regression models and burden tests. Burden analysis (paired and independent T-tests) and Pearson's correlation analyses were calculated in Python v.3.7 using the scipy v1.13.1 *ttest_rel*, *ttest_ind*, or *pearsonr* functions, respectively. We note that paired t-tests for PRS burden between probands and parents were two-tailed, due to the dual directionality of PRS for different phenotypes, while other t-tests were one-tailed. Benjamini-Hochberg multiple testing correction was performed using the scipy v.1.13.1 *false_discovery_control* function for all statistical analyses unless otherwise stated. Multiple testing was performed separately for analyses with all sets of rare variants and variants filtered for evolutionary constraint (defined by LOEUF<0.35, which are intolerant to loss-of-function variants in the general population[37]; referred to as "(LF)"). FDR values reported in the text are corrected for multiple testing, while p-values are not corrected for or did not pass multiple testing. Sample sizes, test statistics, and corrected and uncorrected p-values for all analyses are available in **Tables S2-S5**.

Logistic and linear regression models for phenotypic variation among probands were performed using the *Logit* and *OLM* functions in statsmodels v.0.14.2, respectively (**Table S3A**). For joint variant regression models, we used three different sets of genetic input variables to test for effects towards phenotypes: (a) all rare variants (sum of SNV, STR, and CNV gene burden) and schizophrenia PRS; (b) SNVs, STRs, duplications, and deletions; and (c) SNVs, STRs, duplications, and deletions restricted to genes with LOEUF scores<0.35 (referred to in the figures as "LF model"). Single variant regression models used only a single variant class as input. All models also included sex as a covariate, while models *b* and *c* also included schizophrenia PRS as a covariate. Additional covariates, such as age and genotype PCs, were not included due to concerns regarding potential overfitting of models with lower sample sizes. The variance explained by the models (R^2 for linear models and McFadden's pseudo-R^2 for logistic models) was calculated for all models. Sample sizes, odds ratios, uncorrected p-values,

783 confidence intervals, and variance statistics for all regression models used in the DD cohort are
784 available in **Tables S3A**.
785
786 **Variant enrichment and pathogenicity analysis**
787 *Gene set enrichment:* We assessed enrichment of genes with secondary variants among
788 sets of neurodevelopmental disease genes and genes with neuronal function from several
789 previously published resources[37,44,45,48,104–106]. We identified enrichment of variants in these gene
790 sets by performing Fisher's Exact tests against the whole genome for each gene list and
791 calculated odds ratios and p-values for genes with variants in the DD cohort using the
792 *contingency.odds_ratio* function from scipy v.1.13.1 (**Table S2G**).
793 *Gene ontology:* Gene ontology (GO) term enrichment was performed using the Panther
794 API and the GO-Slim Biological Process annotation dataset[107] (**Table S3B**). The GO term
795 network figures were created using GO term semantic similarity calculated from rrvgo[108] v1.10.0
796 in R v4.2.3.
797 *Spatio-temporal brain expression:* We assessed variant enrichment in genes
798 preferentially expressed in specific brain tissues using the BrainSpan Atlas[52] and in genes
799 preferentially expressed in specific cell types using single-cell RNA-seq expression data in the
800 M1 motor cortex[53]. We defined preferentially expressed genes as those with expression >2SD
801 higher than the median expression across all tissues or all cell types for that gene. We used
802 Fisher's exact tests as described above to find the odds that a gene both carries a variant in the
803 DD cohort and is expressed in a specific brain region or cell type (**Table S2E, S2H**).
804 *16p12.1 differentially and co-expressed genes:* We used Fisher's Exact tests as described
805 to calculate enrichment of secondary variants in gene co-expression modules previously
806 identified using WGCNA in lymphoblastoid cell line (LCL) models of the 16p12.1 deletion[54]
807 (**Table S2F**). We specifically assessed six co-expression modules whose genes showed
808 differential expression between deletion carriers and controls, one of which (black module) also
809 contained two genes within the 16p12.1 deletion region (*POLR3E*, *MOSMO*).
810 For all enrichments, we applied Benjamini-Hochberg multiple testing correction as described
811 above. Sample sizes, test statistics and corrected and uncorrected p-values for enrichments are
812 listed in **Tables S2** and **S3**.

Pathogenic variant analysis: We defined pathogenic SNVs as those that are "Pathogenic" or "Likely pathogenic" in ClinVar for neurodevelopmental phenotypes[42], or loss-of-function variants in genes that (*a*) have dominant or (if male) X-linked neurodevelopmental OMIM phenotype[43], (*b*) are a Tier S or Tier 1 SFARI gene, which represent strong candidate autism genes[45], or (*c*) are in the Tier 1 or Tier 2 gene list from the Developmental Brain Disorder Gene Database, which represent genes with well-documented connections to neurodevelopmental disease[44] (**Table S1D**). Pathogenic CNVs were identified from 50% reciprocal overlap with a previously published list of CNVs[33].

Network analysis

We assessed the connectivity of genes within a previously described brain-specific interaction network. In brief, the network was built using a machine-learning model trained to predict the likelihood of interactions between pairs of genes using brain-specific gene co-expression, protein-protein interaction, and regulatory sequence datasets[49,50]. We restricted the network to genetic interactions with weighted probabilities >2 and calculated the degree for each gene (or number of connections between a particular gene and other genes) as a descriptor of connectivity to other genes in the network. We then binned each gene into one of four quantiles based on the gene's degree of connectivity and counted the number of times that genes with coding variants fell into each quantile. To calculate empirical p-values, we compared the resulting values to 1000 simulations in which we randomly selected the same number of genes from the genome and counted the number of randomly selected genes that fell into each quantile (**Table S2D**).

Genotype-phenotype analysis in 16p12.1 deletion samples from other ascertainments

(a) Description of cohorts: We identified individuals carrying 16p12.1 deletions from three additional cohorts, each representing a distinct ascertainment. Individuals in the Simons Powering Autism Research for Knowledge (SPARK) cohort (n=94) were ascertained for families with autism[58], while the Geisinger MyCode Community Health Initiative (MyCode) (n=160) represents a health care-based cohort[60] and the UK Biobank[59] (UKB) (n=250) consists of individuals with a "healthy volunteer" bias[61]. Combined with samples from the DD cohort (after excluding samples with incomplete phenotypic information), we assessed a total of 757 deletion carriers. De-identified data from these cohorts were obtained and analyzed according to a

844 protocol approved by the Pennsylvania State University Institutional Review Board (IRB
845 #STUDY00011008). Data from the UK Biobank was accessed under application 45023.
846 Individuals from the MyCode cohort were recruited during primary care or specialty clinic visits
847 to Geisinger Health System locations, independent of condition, diagnosis, or demographic
848 characteristic. Written informed consent was obtained from adult patients and from the parents or
849 guardians of pediatric patients. The study was conducted with approval from the Geisinger
850 institutional review board.
851 *(b) CNV calling*: Carriers of the 16p12.1 deletion in each cohort were identified based on
852 CNV calls from microarray data. Samples from MyCode were genotyped using the Illumina
853 Global Screening Array and Illumina OmniExpressExome-8 Kit. SNP log-r ratio and b-allele
854 frequencies for SPARK samples were downloaded through the SFARI Base portal
855 (https://www.base.sfari.org), while signals for the UK Biobank were accessed through Data
856 Fields 22437 and 22431. CNVs for all cohorts were called using the PennCNV pipeline
857 described above for the DD cohort[89]. Additionally, 2,640 and 356 additional samples without any
858 large (>500kb), rare (<0.1% population frequency) CNVs were identified as controls for
859 additional genetic analysis from the UK Biobank and SPARK, respectively.
860 *(c) SNV calling from sequencing data:* SNVs for all three cohorts were identified from
861 whole exome sequencing (WES) data. NimbleGen (SeqCap VCRome) and xGEN probes from
862 Integrated DNA Technologies (IDT) were used for target sequence capture in the MyCode
863 cohort[109,110]. Sequencing was performed by paired end 75bp reads on an Illumina NovaSeq or
864 HiSeq at coverage >20x at >80% of the targeted bases. Alignments and variant calling were
865 based on GRCh38 human genome reference sequence. Variants were called with the WeCall
866 variant caller version 1.1.2 (https://github.com/Genomicsplc/wecall). Whole exome VCFs for
867 SPARK samples were downloaded through the SFARI Base portal (https://www.base.sfari.org).
868 WES VCFs from both cohorts were processed using the same pipeline described above for the
869 DD cohort. WES data from UKB individuals was available as multi-sample project VCFs[111] in
870 the UK Biobank Research Analysis Platform. After splitting multi-allelic records, we applied the
871 following set of quality control filters using Hail in the DNAnexus platform: (a) variant call rate
872 >90%, (b) Hardy Weinberg equilibrium p-value >10-15, (c) minimum read depth > 10, and (d) at
873 least one sample passing the allelic balance threshold of 0.2. Next, we filtered for variants with
874 an intracohort frequency <0.1% and present in at least two samples. The remaining variants were

then annotated using Variant Effect Predictor[112] (VEP) v.109 and dbNSFP[113] v.4 to identify their effects on gene transcripts. We specifically annotated variants based on VEP annotations as LOF (transcript ablation, stop gained, frameshift variant, stop lost, and start lost), missense, or splice (splice acceptor variant and splice donor variant). Missense variants were further filtered for those predicted to be deleterious by at least five of nine selected tools (SIFT, LRT, FATHMM, PROVEAN, MetaSVM, MetaLR, PrimateAI, DEOGEN2, and MutationAssessor) available through the dbNSFP database[113].

(c) Phenotype analysis: We assessed phenotypic data in these cohorts using ICD10 codes derived from electronic health records (MyCode and UKB) and self-reported questionnaire responses (UKB and SPARK) (**Table S4A**). Phenotypic information from SPARK was downloaded from the Simons Foundation through the SFARI Base portal (https://www.base.sfari.org). Electronic health records were available from participants in MyCode[60]. Electronic health records for UKB were identified from Data Fields 41202 and 41204 (main and secondary inpatient ICD10 codes), while questionnaire data was identified from additional Data Fields (**Table S4A**). For harmonization of phenotypic data across cohorts, ICD10 codes were matched to phenotypes from questionnaires, details of which are provided in **Table S4A**.

(d) Secondary variant associations: Comparisons of secondary variant burden between 16p12.1 deletion carriers and age and sex matched controls in the UK Biobank and SPARK were assessed using two-tailed t-tests. The relationship of secondary variant burden and phenotypes in all three cohorts, and comparison with adults and children in the DD cohort, were assessed using two-tailed t-tests. We note that phenotypes in these cohorts were only assessed if they were present in at least five individuals or 10% of the cohort, whichever was larger, and the cohort had at least five or 10% of the cohort controls. We further assessed the effects of secondary variant on phenotypes across cohorts using logistic regression using the model structures *(b)* and *(c)* described above for the DD cohort, without the inclusion of STR variants. Phenotypic logistic regression was performed on main and secondary ICD10 codes with age and sex included as covariates. PheWAS was performed using the *PheWAS* v.0.99.6-1 package in R[114] on all available samples from the UK Biobank using Phecodes derived from ICD9 and ICD10 codes, identified from Data Fields 41202, 41204, 41203, and 41205 (main and secondary inpatient ICD9/10 codes), while correcting for sex, age, and the top four genetic principal components.

Sample sizes, test statistics, uncorrected and corrected p-values, confidence intervals, and variance statistics for all analyses are available in **Table S4**.

Genotype-phenotype analysis in other neurodevelopmental disease cohorts

We assessed genomic and phenotypic data from individuals from the Simons Searchlight project[63], ascertained for probands with 16p11.2 deletions and duplications, and Simons Simplex Collection (SSC), ascertained for families with simplex cases of autism[64]. Within the Simons Searchlight cohort, we assessed 128 probands with the 16p11.2 duplication (n=37) or deletion (n=91). Within the SSC cohort, we assessed genomic data of 2,435 total probands, and classified probands with the following primary variant classes for downstream analysis: (i) 1,237 probands with rare, deleterious variants (<0.1% frequency, loss-of-function or missense variants with CADD Phred-like scores >25); (ii) 79 probands with large rare deletions (<0.1% population frequency, >500kbp); (iii) 148 probands with large rare duplications (<0.1% population frequency, >500kbp); and (iv) 1,084 probands who did not carry any of these variants or any other known pathogenic CNVs[33]. We note that groups with primary variants have overlapping samples. Additionally, we assessed phenotypic data for an additional 419 SSC probands and 32 Simons Searchlight probands to compare SRS distributions with 16p12.1 deletion probands (a total of 2,844 total SSC probands and 139 Simons Searchlight probands were used in these analyses). De-identified data from these cohorts were obtained and analyzed according to a protocol approved by the Pennsylvania State University Institutional Review Board (IRB #STUDY00011008). In sum, we assessed secondary variants and phenotypes in 2,252 individuals with primary variants from six cohorts: DD (n=269), SPARK (n=94), MyCode (n=160), UKB (n=250), Searchlight (n=128), and SSC (n=1,351). We also assessed data from 1,084 SSC probands without primary variants. Data from an additional 311,980 control individuals was also included, including non-carrier samples from DD (n=183), age and sex-matched controls without CNVs from SPARK (n=356), PheWAS and age and sex-matched controls without CNVs from UKB (n=310,990), 16p11.2 CNV samples from Searchlight and SSC probands without genetic data but with SRS data (n=32 and 419, respectively).

Exome sequencing VCFs and raw microarray data for Searchlight cohorts, whole genome sequencing VCFs and STR calls for SSC[75], and all phenotype data were downloaded through the SFARI Base portal (https://www.base.sfari.org). We processed and filtered exome and WGS-

based SNVs and indels for the same quality control filters used to process the DD cohort, and then annotated variants using our standardized pipeline. Short tandem repeat calls from SSC were previously published by Mitra and colleagues[75] and were processed and filtered with the same pipeline as our cohort. CNV calls from microarrays for SSC were previously published by Sanders and colleagues[115], while CNV calls from microarrays for the Searchlight cohort were processed using PennCNV[89] as previously described[28]. For this manuscript, genes within CNVs were reannotated using GENCODE v19[85], but otherwise used as-is without additional processing. Primary variant SNVs and CNVs were removed from secondary variant lists for downstream processing. We further processed microarray data and calculated PRS for both cohorts using the same pipelines as the DD cohort, except that autism PRS was not calculated in SSC samples, as the underlying GWAS summary statistics were calculated in part using SSC samples[41]. Finally, we curated results of quantitative phenotypic assessments for each cohort from SFARI Base, including full-scale IQ, internalizing and externalizing behavior (ABCL/CBCL), social responsiveness (SRS), autism-related behaviors (BSI, Searchlight only), repetitive behavior (RBS-R, SSC only), coordination disorder (DCDQ, SSC only), BMI z-score, and head circumference z-score (Searchlight only).

Linear regression models for assessing variation in these phenotypes were constructed using the *OLS* function from statsmodels v0.14.2, using the same model structures *(b)* and *(c)* described above for the DD cohort (note that Searchlight models did not include STRs). For models investigating the interactions of primary and secondary variants, Benjamini-Hochberg FDR correction was performed using the statsmodels v.0.14.2 *false_discovery_control* function. All other correlations, statistical analyses, GO enrichments, and multiple testing corrections for comparing variant classes and quantitative phenotypes were performed in the same manner as for the DD cohort. Sample sizes, test statistics, uncorrected and corrected p-values, confidence intervals, and variance statistics for all analyses are available in **Table S5**.

Data and code availability

Whole genome sequencing and SNP microarray data generated in this study are available at NCBI dbGaP phs002450.v2.p1. All code generated for this project, including pipelines for running bioinformatic software and custom analysis scripts, are available at https://github.com/girirajanlab/16p12_WGS_project. Statistical analyses and experimental

968 results for the data presented in **Figs. 2-7** and associated supplementary figures are available in
969 **Tables S2-S6.**

REFERENCES

1. Claussnitzer, M., Cho, J.H., Collins, R., Cox, N.J., Dermitzakis, E.T., Hurles, M.E., Kathiresan, S., Kenny, E.E., Lindgren, C.M., MacArthur, D.G., et al. (2020). A brief history of human disease genetics. Nature *577*, 179–189. https://doi.org/10.1038/s41586-019-1879-7.

2. Kingdom, R., and Wright, C.F. (2022). Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts. Front Genet *13*, 920390. https://doi.org/10.3389/fgene.2022.920390.

3. Girirajan, S., and Eichler, E.E. (2010). Phenotypic variability and genetic susceptibility to genomic disorders. Hum Mol Genet *19*, R176-87. https://doi.org/10.1093/hmg/ddq366.

4. Posey, J.E., Harel, T., Liu, P., Rosenfeld, J.A., James, R.A., Coban Akdemir, Z.H., Walkiewicz, M., Bi, W., Xiao, R., Ding, Y., et al. (2017). Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. N Engl J Med *376*, 21–31. https://doi.org/10.1056/NEJMoa1516767.

5. Leitch, C.C., Zaghloul, N.A., Davis, E.E., Stoetzel, C., Diaz-Font, A., Rix, S., Alfadhel, M., Lewis, R.A., Eyaid, W., Banin, E., et al. (2008). Hypomorphic mutations in syndromic encephalocele genes are associated with Bardet-Biedl syndrome. Nat Genet *40*, 443–448. https://doi.org/10.1038/ng.97.

6. Riordan, J.D., and Nadeau, J.H. (2017). From Peas to Disease: Modifier Genes, Network Resilience, and the Genetics of Health. Am J Hum Genet *101*, 177–191. https://doi.org/10.1016/j.ajhg.2017.06.004.

7. Guo, T., Chung, J.H., Wang, T., McDonald-McGinn, D.M., Kates, W.R., Hawuła, W., Coleman, K., Zackai, E., Emanuel, B.S., and Morrow, B.E. (2015). Histone Modifier Genes Alter Conotruncal Heart Phenotypes in 22q11.2 Deletion Syndrome. Am J Hum Genet *97*, 869–877. https://doi.org/10.1016/j.ajhg.2015.10.013.

8. Parenti, I., Rabaneda, L.G., Schoen, H., and Novarino, G. (2020). Neurodevelopmental Disorders: From Genetics to Functional Pathways. Trends Neurosci *43*, 608–621. https://doi.org/10.1016/j.tins.2020.05.004.

9. Jacquemont, S., Huguet, G., Klein, M., Chawner, S.J.R.A., Donald, K.A., van den Bree, M.B.M., Sebat, J., Ledbetter, D.H., Constantino, J.N., Earl, R.K., et al. (2022). Genes To Mental Health (G2MH): A Framework to Map the Combined Effects of Rare and Common Variants on Dimensions of Cognition and Psychopathology. Am J Psychiatry *179*, 189–203. https://doi.org/10.1176/appi.ajp.2021.21040432.

10. Oetjens, M.T., Kelly, M.A., Sturm, A.C., Martin, C.L., and Ledbetter, D.H. (2019). Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. Nat Commun *10*, 4897. https://doi.org/10.1038/s41467-019-12869-0.

11. Hudac, C.M., Bove, J., Barber, S., Duyzend, M., Wallace, A., Martin, C.L., Ledbetter, D.H., Hanson, E., Goin-Kochel, R.P., Green-Snyder, L., et al. (2020). Evaluating heterogeneity in ASD symptomatology, cognitive ability, and adaptive functioning among 16p11.2 CNV carriers. Autism Res *13*, 1300–1310. https://doi.org/10.1002/aur.2332.

12. Cleynen, I., Engchuan, W., Hestand, M.S., Heung, T., Holleman, A.M., Johnston, H.R., Monfeuga, T., McDonald-McGinn, D.M., Gur, R.E., Morrow, B.E., et al. (2021). Genetic contributors to risk of schizophrenia in the presence of a 22q11.2 deletion. Mol Psychiatry *26*, 4496–4510. https://doi.org/10.1038/s41380-020-0654-3.

13. Tansey, K.E., Rees, E., Linden, D.E., Ripke, S., Chambert, K.D., Moran, J.L., McCarroll, S.A., Holmans, P., Kirov, G., Walters, J., et al. (2016). Common alleles contribute to schizophrenia in CNV carriers. Mol Psychiatry *21*, 1085–1089. https://doi.org/10.1038/mp.2015.143.

14. Davies, R.W., Fiksinski, A.M., Breetvelt, E.J., Williams, N.M., Hooper, S.R., Monfeuga, T., Bassett, A.S., Owen, M.J., Gur, R.E., Morrow, B.E., et al. (2020). Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. Nat Med *26*, 1912–1918. https://doi.org/10.1038/s41591-020-1103-1.

15. Alver, M., Mancini, V., Läll, K., Schneider, M., Romano, L., Estonian Biobank Research Team, Mägi, R., Dermitzakis, E.T., Eliez, S., and Reymond, A. (2022). Contribution of schizophrenia polygenic burden to longitudinal phenotypic variance in 22q11.2 deletion syndrome. Mol Psychiatry *27*, 4191–4200. https://doi.org/10.1038/s41380-022-01674-9.

16. Bergen, S.E., Ploner, A., Howrigan, D., CNV Analysis Group and the Schizophrenia Working Group of the Psychiatric Genomics Consortium, O'Donovan, M.C., Smoller, J.W., Sullivan, P.F., Sebat, J., Neale, B., and Kendler, K.S. (2019). Joint Contributions of Rare Copy Number Variants and Common SNPs to Risk for Schizophrenia. Am J Psychiatry *176*, 29–35. https://doi.org/10.1176/appi.ajp.2018.17040467.

17. Alver, M., Mancini, V., Läll, K., Schneider, M., Romano, L., Estonian Biobank Research Team, Mägi, R., Dermitzakis, E.T., Eliez, S., and Reymond, A. (2022). Contribution of schizophrenia polygenic burden to longitudinal phenotypic variance in 22q11.2 deletion syndrome. Mol Psychiatry *27*, 4191–4200. https://doi.org/10.1038/s41380-022-01674-9.

18. Banerjee, D., and Girirajan, S. (2023). Pathogenic Variants and Ascertainment: Neuropsychiatric Disease Risk in a Health System Cohort. American Journal of Psychiatry *180*, 11–13. https://doi.org/10.1176/appi.ajp.20220934.

19. Shimelis, H., Oetjens, M.T., Walsh, L.K., Wain, K.E., Znidarsic, M., Myers, S.M., Finucane, B.M., Ledbetter, D.H., and Martin, C.L. (2023). Prevalence and Penetrance of Rare Pathogenic Variants in Neurodevelopmental Psychiatric Genes in a Health Care System Population. Am J Psychiatry *180*, 65–72. https://doi.org/10.1176/APPI.AJP.22010062.

20. Crawford, K., Bracher-Smith, M., Owen, D., Kendall, K.M., Rees, E., Pardiñas, A.F., Einon, M., Escott-Price, V., Walters, J.T.R., O'Donovan, M.C., et al. (2019). Medical consequences of pathogenic CNVs in adults: Analysis of the UK Biobank. J Med Genet *56*, 131–138. https://doi.org/10.1136/jmedgenet-2018-105477.

21. Auwerx, C., Lepamets, M., Sadler, M.C., Patxot, M., Stojanov, M., Baud, D., Mägi, R., Estonian Biobank Research Team, Porcu, E., Reymond, A., et al. (2022). The individual and global impact of copy-number variants on complex human traits. Am J Hum Genet *109*, 647–668. https://doi.org/10.1016/j.ajhg.2022.02.010.

22. Auwerx, C., Jõeloo, M., Sadler, M.C., Tesio, N., Ojavee, S., Clark, C.J., Mägi, R., Estonian Biobank Research Team, Reymond, A., and Kutalik, Z. (2024). Rare copy-number variants as modulators of common disease susceptibility. Genome Med *16*, 5. https://doi.org/10.1186/s13073-023-01265-5.

23. Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A.R., Green, T., et al. (2008). Association between Microdeletion and Microduplication at 16p11.2 and Autism. New England Journal of Medicine *358*, 667–675. https://doi.org/10.1056/NEJMoa075974.

24. Auwerx, C., Moix, S., Kutalik, Z., and Reymond, A. (2024). Disentangling mechanisms behind the pleiotropic effects of proximal 16p11.2 BP4-5 CNVs. medRxiv, 2024.03.20.24304613. https://doi.org/10.1101/2024.03.20.24304613.

25. Walters, R.G., Jacquemont, S., Valsesia, A., de Smith, A.J., Martinet, D., Andersson, J., Falchi, M., Chen, F., Andrieux, J., Lobbens, S., et al. (2010). A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature *463*, 671–675. https://doi.org/10.1038/nature08727.

26. Jensen, M., and Girirajan, S. (2019). An interaction-based model for neuropsychiatric features of copy-number variants. PLoS Genet *15*, e1007879. https://doi.org/10.1371/journal.pgen.1007879.

27. Girirajan, S., Rosenfeld, J.A., Cooper, G.M., Antonacci, F., Siswara, P., Itsara, A., Vives, L., Walsh, T., McCarthy, S.E., Baker, C., et al. (2010). A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat Genet *42*, 203–209. https://doi.org/10.1038/ng.534.

28. Pizzo, L., Jensen, M., Polyak, A., Rosenfeld, J.A., Mannik, K., Krishnan, A., McCready, E., Pichon, O., Le Caignec, C., Van Dijck, A., et al. (2019). Rare variants in the genetic background modulate cognitive and developmental phenotypes in individuals carrying disease-associated variants. Genet Med *21*, 816–825. https://doi.org/10.1038/s41436-018-0266-3.

29. Girirajan, S., Pizzo, L., Moeschler, J., and Rosenfeld, J. (1993). 16p12.2 Recurrent Deletion. In GeneReviews® (University of Washington, Seattle).

30. Rees, E., Walters, J.T.R., Chambert, K.D., O'Dushlaine, C., Szatkiewicz, J., Richards, A.L., Georgieva, L., Mahoney-Davies, G., Legge, S.E., Moran, J.L., et al. (2014). CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1. Hum Mol Genet *23*, 1669–1676. https://doi.org/10.1093/hmg/ddt540.

31. Stefansson, H., Meyer-Lindenberg, A., Steinberg, S., Magnusdottir, B., Morgen, K., Arnarsdottir, S., Bjornsdottir, G., Walters, G.B., Jonsdottir, G.A., Doyle, O.M., et al. (2014). CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature *505*, 361–366. https://doi.org/10.1038/nature12818.

32. Rees, E., Kendall, K., Pardiñas, A.F., Legge, S.E., Pocklington, A., Escott-Price, V., MacCabe, J.H., Collier, D.A., Holmans, P., O'Donovan, M.C., et al. (2016). Analysis of

intellectual disability copy number variants for association with schizophrenia. JAMA Psychiatry *73*, 963–969. https://doi.org/10.1001/jamapsychiatry.2016.1831.

33. Girirajan, S., Rosenfeld, J.A., Coe, B.P., Parikh, S., Friedman, N., Goldstein, A., Filipink, R.A., McConnell, J.S., Angle, B., Meschino, W.S., et al. (2012). Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N Engl J Med *367*, 1321–1331. https://doi.org/10.1056/NEJMoa1200395.

34. Hansen, J.A. (2019). Development and Psychometric Evaluation of the Hansen Research Services Matrix Adaptive Test: A Measure of Nonverbal IQ. J Autism Dev Disord *49*, 2721–2732. https://doi.org/10.1007/s10803-016-2932-0.

35. Constantino, J.N., Davis, S.A., Todd, R.D., Schindler, M.K., Gross, M.M., Brophy, S.L., Metzger, L.M., Shoushtari, C.S., Splinter, R., and Reich, W. (2003). Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. J Autism Dev Disord *33*, 427–433. https://doi.org/10.1023/a:1025014929212.

36. Zubler, J., and Whitaker, T. (2022). CDC's Revised Developmental Milestone Checklists. Am Fam Physician *106*, 370–371.

37. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

38. Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet *50*, 1112–1121. https://doi.org/10.1038/s41588-018-0147-3.

39. Savage, J.E., Jansen, P.R., Stringer, S., Watanabe, K., Bryois, J., De Leeuw, C.A., Nagel, M., Awasthi, S., Barr, P.B., Coleman, J.R.I., et al. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat Genet *50*, 912–919. https://doi.org/10.1038/s41588-018-0152-6.

40. Ripke, S., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K.H., Holmans, P.A., Lee, P., Bulik-Sullivan, B., Collier, D.A., Huang, H., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427. https://doi.org/10.1038/nature13595.

41. Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. Nat Genet *51*, 431–444. https://doi.org/10.1038/s41588-019-0344-8.

42. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Res *46*, D1062–D1067. https://doi.org/10.1093/nar/gkx1153.

43. Amberger, J.S., Bocchini, C.A., Scott, A.F., and Hamosh, A. (2019). OMIM.org: Leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res *47*, D1038–D1043. https://doi.org/10.1093/nar/gky1151.

44. Gonzalez-Mantilla, A.J., Moreno-De-Luca, A., Ledbetter, D.H., and Martin, C.L. (2016). A cross-disorder method to identify novel candidate genes for developmental brain disorders. JAMA Psychiatry *73*, 275–283. https://doi.org/10.1001/jamapsychiatry.2015.2692.

45. Abrahams, B.S., Arking, D.E., Campbell, D.B., Mefford, H.C., Morrow, E.M., Weiss, L.A., Menashe, I., Wadkins, T., Banerjee-Basu, S., and Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol Autism *4*, 36. https://doi.org/10.1186/2040-2392-4-36.

46. Jones, W.D., Dafou, D., McEntagart, M., Woollard, W.J., Elmslie, F. V, Holder-Espinasse, M., Irving, M., Saggar, A.K., Smithson, S., Trembath, R.C., et al. (2012). De novo mutations in MLL cause Wiedemann-Steiner syndrome. Am J Hum Genet *91*, 358–364. https://doi.org/10.1016/j.ajhg.2012.06.008.

47. Tang, H., Kirkness, E.F., Lippert, C., Biggs, W.H., Fabani, M., Guzman, E., Ramakrishnan, S., Lavrenko, V., Kakaradov, B., Hou, C., et al. (2017). Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. Am J Hum Genet *101*, 700–715. https://doi.org/10.1016/j.ajhg.2017.09.013.

48. Werling, D.M., Brand, H., An, J.-Y., Stone, M.R., Zhu, L., Glessner, J.T., Collins, R.L., Dong, S., Layer, R.M., Markenscoff-Papadimitriou, E., et al. (2018). An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat Genet *50*, 727–736. https://doi.org/10.1038/s41588-018-0107-y.

49. Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat Genet *47*, 569–576. https://doi.org/10.1038/ng.3259.

50. Krishnan, A., Zhang, R., Yao, V., Theesfeld, C.L., Wong, A.K., Tadych, A., Volfovsky, N., Packer, A., Lash, A., and Troyanskaya, O.G. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat Neurosci *19*, 1454–1462. https://doi.org/10.1038/nn.4353.

51. Hou, J., van Leeuwen, J., Andrews, B.J., and Boone, C. (2018). Genetic Network Complexity Shapes Background-Dependent Phenotypic Expression. Trends Genet *34*, 578–586. https://doi.org/10.1016/j.tig.2018.05.006.

52. Miller, J.A., Ding, S.-L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., Aiona, K., et al. (2014). Transcriptional landscape of the prenatal human brain. Nature *508*, 199–206. https://doi.org/10.1038/nature13185.

53. Bakken, T.E., van Velthoven, C.T., Menon, V., Hodge, R.D., Yao, Z., Nguyen, T.N., Graybuck, L.T., Horwitz, G.D., Bertagnolli, D., Goldy, J., et al. (2021). Single-cell and single-nucleus RNA-seq uncovers shared and distinct axes of variation in dorsal LGN neurons in mice, non-human primates, and humans. Elife *10*. https://doi.org/10.7554/eLife.64875.

54. Jensen, M., Tyryshkina, A., Pizzo, L., Smolen, C., Das, M., Huber, E., Krishnan, A., and Girirajan, S. (2021). Combinatorial patterns of gene expression changes contribute to variable expressivity of the developmental delay-associated 16p12.1 deletion. Genome Med *13*, 163. https://doi.org/10.1186/s13073-021-00982-z.

55. De Vries, B.B.A., White, S.M., Knight, S.J.L., Regan, R., Homfray, T., Young, I.D., Super, M., McKeown, C., Splitt, M., Quarrell, O.W.J., et al. (2001). Clinical studies on submicroscopic subtelomeric rearrangements: A checklist. J Med Genet *38*, 145–150. https://doi.org/10.1136/jmg.38.3.145.

56. Flannick, J., Beer, N.L., Bick, A.G., Agarwala, V., Molnes, J., Gupta, N., Burtt, N.P., Florez, J.C., Meigs, J.B., Taylor, H., et al. (2013). Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. Nat Genet *45*, 1380–1385. https://doi.org/10.1038/ng.2794.

57. Gunther, D.F., Eugster, E., Zagar, A.J., Bryant, C.G., Davenport, M.L., and Quigley, C.A. (2004). Ascertainment bias in Turner syndrome: new insights from girls who were

diagnosed incidentally in prenatal life. Pediatrics *114*, 640–644. https://doi.org/10.1542/peds.2003-1122-L.

58. Feliciano, P., Daniels, A.M., Green Snyder, L.A., Beaumont, A., Camba, A., Esler, A., Gulsrud, A.G., Mason, A., Gutierrez, A., Nicholson, A., et al. (2018). SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. Neuron *97*, 488–493. https://doi.org/10.1016/j.neuron.2018.01.015.

59. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/S41586-018-0579-Z.

60. Carey, D.J., Fetterolf, S.N., Davis, F.D., Faucett, W.A., Kirchner, H.L., Mirshahi, U., Murray, M.F., Smelser, D.T., Gerhard, G.S., and Ledbetter, D.H. (2016). The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet Med *18*, 906–913. https://doi.org/10.1038/gim.2015.187.

61. Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N.E. (2017). Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol *186*, 1026–1034. https://doi.org/10.1093/aje/kwx246.

62. Männik, K., Mägi, R., Macé, A., Cole, B., Guyatt, A.L., Shihab, H.A., Maillard, A.M., Alavere, H., Kolk, A., Reigo, A., et al. (2015). Copy number variations and cognitive phenotypes in unselected populations. JAMA *313*, 2044–2054. https://doi.org/10.1001/jama.2015.4845.

63. Spiro, J.E., and Chung, W.K. (2012). Simons Variation in Individuals Project (Simons VIP): A Genetics-First Approach to Studying Autism Spectrum and Related

Neurodevelopmental Disorders. Neuron *73*, 1063–1067. https://doi.org/10.1016/j.neuron.2012.02.014.

64. Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. Neuron *68*, 192–195. https://doi.org/10.1016/j.neuron.2010.10.006.

65. Pucilowska, J., Vithayathil, J., Tavares, E.J., Kelly, C., Karlo, J.C., and Landreth, G.E. (2015). The 16p11.2 deletion mouse model of autism exhibits altered cortical progenitor proliferation and brain cytoarchitecture linked to the ERK MAPK pathway. J Neurosci *35*, 3190–3200. https://doi.org/10.1523/JNEUROSCI.4864-13.2015.

66. Iyer, J., Singh, M.D., Jensen, M., Patel, P., Pizzo, L., Huber, E., Koerselman, H., Weiner, A.T., Lepanto, P., Vadodaria, K., et al. (2018). Pervasive genetic interactions modulate neurodevelopmental defects of the autism-associated 16p11.2 deletion in Drosophila melanogaster. Nat Commun *9*, 2548. https://doi.org/10.1038/s41467-018-04882-6.

67. Veltman, J.A., and Brunner, H.G. (2010). Understanding variable expressivity in microdeletion syndromes. Nat Genet *42*, 192–193. https://doi.org/10.1038/ng0310-192.

68. Smolen, C., Jensen, M., Dyer, L., Pizzo, L., Tyryshkina, A., Banerjee, D., Rohan, L., Huber, E., El Khattabi, L., Prontera, P., et al. (2023). Assortative mating and parental genetic relatedness contribute to the pathogenicity of variably expressive variants. Am J Hum Genet *110*, 2015–2028. https://doi.org/10.1016/J.AJHG.2023.10.015.

69. Martin, C.L., Wain, K.E., Oetjens, M.T., Tolwinski, K., Palen, E., Hare-Harris, A., Habegger, L., Maxwell, E.K., Reid, J.G., Walsh, L.K., et al. (2020). Identification of Neuropsychiatric Copy Number Variants in a Health Care System Population. JAMA Psychiatry *77*, 1276–1285. https://doi.org/10.1001/jamapsychiatry.2020.2159.

70. Männik, K., Mägi, R., Macé, A., Cole, B., Guyatt, A.L., Shihab, H.A., Maillard, A.M., Alavere, H., Kolk, A., Reigo, A., et al. (2015). Copy number variations and cognitive phenotypes in unselected populations. JAMA *313*, 2044–2054. https://doi.org/10.1001/jama.2015.4845.

71. Antaki, D., Guevara, J., Maihofer, A.X., Klein, M., Gujral, M., Grove, J., Carey, C.E., Hong, O., Arranz, M.J., Hervas, A., et al. (2022). A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex. Nat Genet *54*, 1284–1292. https://doi.org/10.1038/s41588-022-01064-5.

72. Mitra, I., Lavillaureix, A., Yeh, E., Traglia, M., Tsang, K., Bearden, C.E., Rauen, K.A., and Weiss, L.A. (2017). Reverse Pathway Genetic Approach Identifies Epistasis in Autism Spectrum Disorders. PLoS Genet *13*, e1006516. https://doi.org/10.1371/journal.pgen.1006516.

73. Hivert, V., Sidorenko, J., Rohart, F., Goddard, M.E., Yang, J., Wray, N.R., Yengo, L., and Visscher, P.M. (2021). Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. Am J Hum Genet *108*, 786–798. https://doi.org/10.1016/j.ajhg.2021.02.014.

74. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038.

75. Mitra, I., Huang, B., Mousavi, N., Ma, N., Lamkin, M., Yanicky, R., Shleizer-Burko, S., Lohmueller, K.E., and Gymrek, M. (2021). Patterns of de novo tandem repeat mutations and their role in autism. Nature *589*, 246–250. https://doi.org/10.1038/s41586-020-03078-7.

76. Turner, T.N., Coe, B.P., Dickel, D.E., Hoekzema, K., Nelson, B.J., Zody, M.C., Kronenberg, Z.N., Hormozdiari, F., Raja, A., Pennacchio, L.A., et al. (2017). Genomic

Patterns of De Novo Mutation in Simplex Autism. Cell *171*, 710–722. https://doi.org/10.1016/j.cell.2017.08.047.

77. Kuczmarski, R.J., Ogden, C.L., Guo, S.S., Grummer-Strawn, L.M., Flegal, K.M., Mei, Z., Wei, R., Curtin, L.R., Roche, A.F., and Johnson, C.L. (2002). 2000 CDC Growth Charts for the United States: methods and development. Vital Health Stat 11, 1–190.

78. Rollins, J.D., Collins, J.S., and Holden, K.R. (2010). United States head circumference growth reference charts: birth to 21 years. J Pediatr *156*, 907-913.e2. https://doi.org/10.1016/j.jpeds.2010.01.009.

79. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

80. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595. https://doi.org/10.1093/bioinformatics/btp698.

81. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

82. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Auwera, G.A. Van der, Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv, 201178. https://doi.org/10.1101/201178.

83. Pedersen, B.S., Layer, R.M., and Quinlan, A.R. (2016). Vcfanno: fast, flexible annotation of genetic variants. Genome Biol *17*, 118. https://doi.org/10.1186/s13059-016-0973-5.

84. Lek, M., Karczewski, K.J., Minikel, E. V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291. https://doi.org/10.1038/nature19057.

85. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., et al. (2021). GENCODE 2021. Nucleic Acids Res *49*, D916–D923. https://doi.org/10.1093/nar/gkaa1087.

86. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res *38*, e164. https://doi.org/10.1093/nar/gkq603.

87. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res *47*, D886–D894. https://doi.org/10.1093/nar/gky1016.

88. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–329. https://doi.org/10.1038/NATURE14248.

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

90. Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res *21*, 974–984. https://doi.org/10.1101/gr.114876.110.

91. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710.

92. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: A probabilistic framework for structural variant discovery. Genome Biol *15*, R84. https://doi.org/10.1186/gb-2014-15-6-r84.

93. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, 333–339. https://doi.org/10.1093/bioinformatics/bts378.

94. Brandler, W.M., Antaki, D., Gujral, M., Kleiber, M.L., Whitney, J., Maile, M.S., Hong, O., Chapman, T.R., Tan, S., Tandon, P., et al. (2018). Paternally inherited cis-regulatory structural variants are associated with autism. Science *360*, 327–331. https://doi.org/10.1126/science.aan2261.

95. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A. V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451. https://doi.org/10.1038/s41586-020-2287-8.

96. Coe, B.P., Witherspoon, K., Rosenfeld, J.A., Van Bon, B.W.M., Vulto-Van Silfhout, A.T., Bosco, P., Friend, K.L., Baker, C., Buono, S., Vissers, L.E.L.M., et al. (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet *46*, 1063–1071. https://doi.org/10.1038/ng.3092.

97. Mousavi, N., Margoliash, J., Pusarla, N., Saini, S., Yanicky, R., and Gymrek, M. (2021). TRTools: A toolkit for genome-wide analysis of tandem repeats. Bioinformatics *37*, 731–733. https://doi.org/10.1093/bioinformatics/btaa736.

98. Mousavi, N., Shleizer-Burko, S., Yanicky, R., and Gymrek, M. (2019). Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res *47*, e90. https://doi.org/10.1093/nar/gkz501.

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

100. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat Genet *48*, 1284–1287. https://doi.org/10.1038/ng.3656.

101. Pedersen, B.S., and Quinlan, A.R. (2017). Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. Am J Hum Genet *100*, 406–413. https://doi.org/10.1016/j.ajhg.2017.01.017.

102. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2021). LDpred2: better, faster, stronger. Bioinformatics *36*, 5424–5431. https://doi.org/10.1093/bioinformatics/btaa1029.

103. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Schaffner, S.F., Yu, F., Dermitzakis, E., Bonnen, P.E., De Bakker, P.I.W., Deloukas, P., Gabriel, S.B., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58. https://doi.org/10.1038/nature09298.

104. Wang, J., Lin, Z.J., Liu, L., Xu, H.Q., Shi, Y.W., Yi, Y.H., He, N., and Liao, W.P. (2017). Epilepsy-associated genes. Seizure *44*, 11–20. https://doi.org/10.1016/j.seizure.2016.11.030.

105. Thormann, A., Halachev, M., McLaren, W., Moore, D.J., Svinti, V., Campbell, A., Kerr, S.M., Tischkowitz, M., Hunt, S.E., Dunlop, M.G., et al. (2019). Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. Nat Commun *10*, 2373. https://doi.org/10.1038/s41467-019-10016-3.

106. Wu, Y., Li, X., Liu, J., Luo, X.-J., and Yao, Y.-G. (2020). SZDB2.0: an updated comprehensive resource for schizophrenia research. Hum Genet *139*, 1285–1297. https://doi.org/10.1007/s00439-020-02171-1.

107. Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P.D. (2019). PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res *47*, D419–D426. https://doi.org/10.1093/nar/gky1038.

108. Sayols, S. (2023). rrvgo: a Bioconductor package for interpreting lists of Gene Ontology terms. MicroPubl Biol *2023*. https://doi.org/10.17912/micropub.biology.000811.

109. Beck, D.B., Bodian, D.L., Shah, V., Mirshahi, U.L., Kim, J., Ding, Y., Magaziner, S.J., Strande, N.T., Cantor, A., Haley, J.S., et al. (2023). Estimated Prevalence and Clinical Manifestations of UBA1 Variants Associated With VEXAS Syndrome in a Clinical Population. JAMA *329*, 318–324. https://doi.org/10.1001/jama.2022.24836.

110. Staples, J., Maxwell, E.K., Gosalia, N., Gonzaga-Jauregui, C., Snyder, C., Hawes, A., Penn, J., Ulloa, R., Bai, X., Lopez, A.E., et al. (2018). Profiling and Leveraging Relatedness in a Precision Medicine Cohort of 92,455 Exomes. Am J Hum Genet *102*, 874–889. https://doi.org/10.1016/j.ajhg.2018.03.012.

111. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature *599*, 628–634. https://doi.org/10.1038/s41586-021-04103-z.

112. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol *17*, 122. https://doi.org/10.1186/s13059-016-0974-4.

113. Liu, X., Li, C., Mou, C., Dong, Y., and Tu, Y. (2020). dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med *12*, 103. https://doi.org/10.1186/s13073-020-00803-9.

114. Denny, J.C., Ritchie, M.D., Basford, M.A., Pulley, J.M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D.R., Roden, D.M., and Crawford, D.C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics *26*, 1205–1210. https://doi.org/10.1093/bioinformatics/btq126.

115. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron *87*, 1215–1233. https://doi.org/10.1016/j.neuron.2015.09.016.

1511 **SUPPLEMENTAL FIGURES**

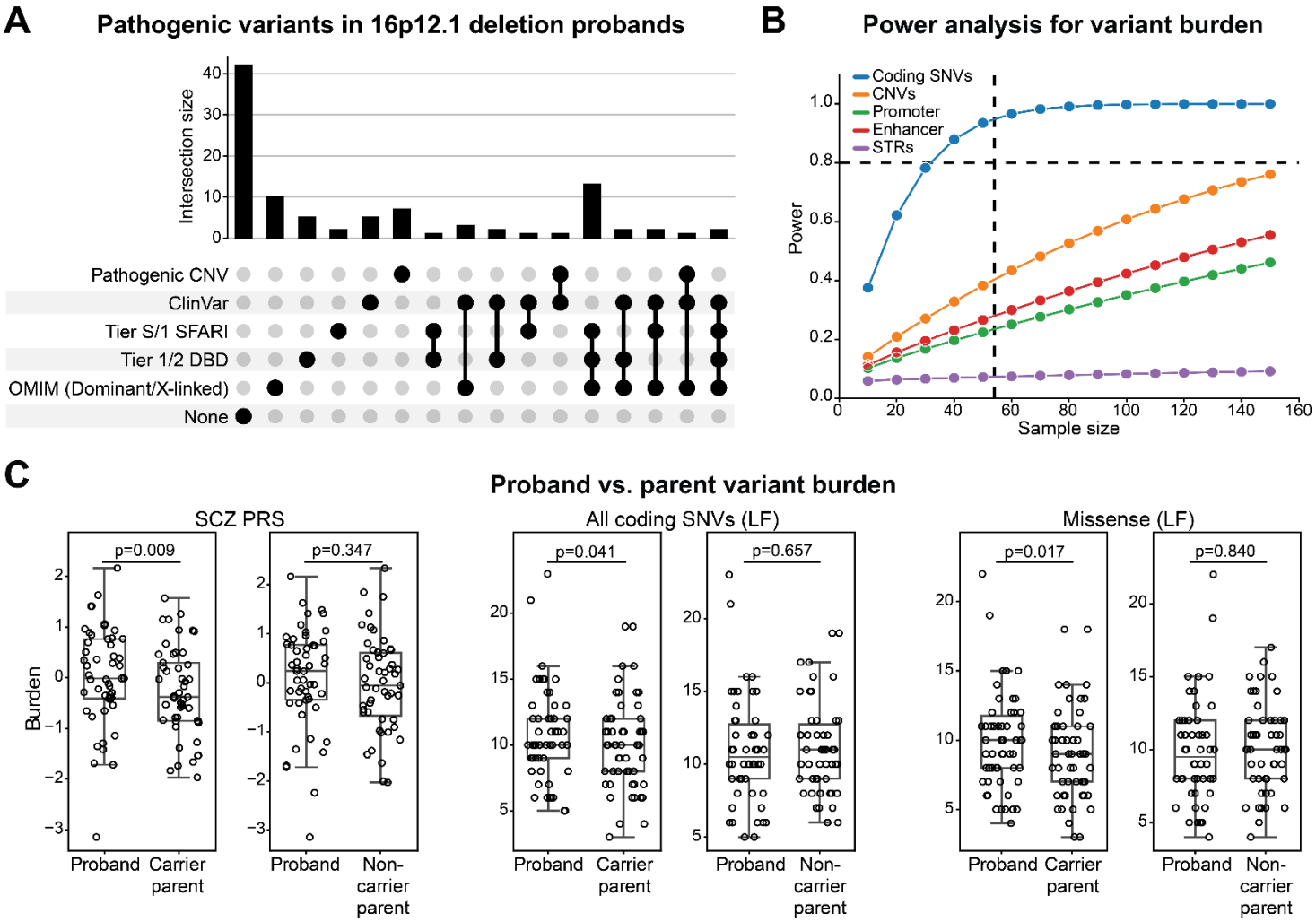


Figure S1. Secondary variant burden comparisons among 16p12.1 deletion family members (related to Figure 3). (A) UpSet plot shows the number of 16p12.1 deletion probands with secondary variants in one or more disease-associated categories, potentially indicative of multiple genetic diagnoses. **(B)** Power analysis for detecting changes in burden of rare variant classes

among individuals with the 16p12.1 deletion (see Methods). Dashed horizontal line indicates 80% power and dashed vertical line represents the sample size of proband-carrier parent pairs (n=54). **(C)** Changes in burden of SCZ PRS (left), missense (LF) variants (center), and all coding SNVs (LF) (right) between 16p12.1 deletion probands and their carrier (left, n=49-54) and noncarrier parents (right, n=50-51). P-values from one-tailed (rare variants) or two-tailed (SCZ PRS) t-tests.

1512

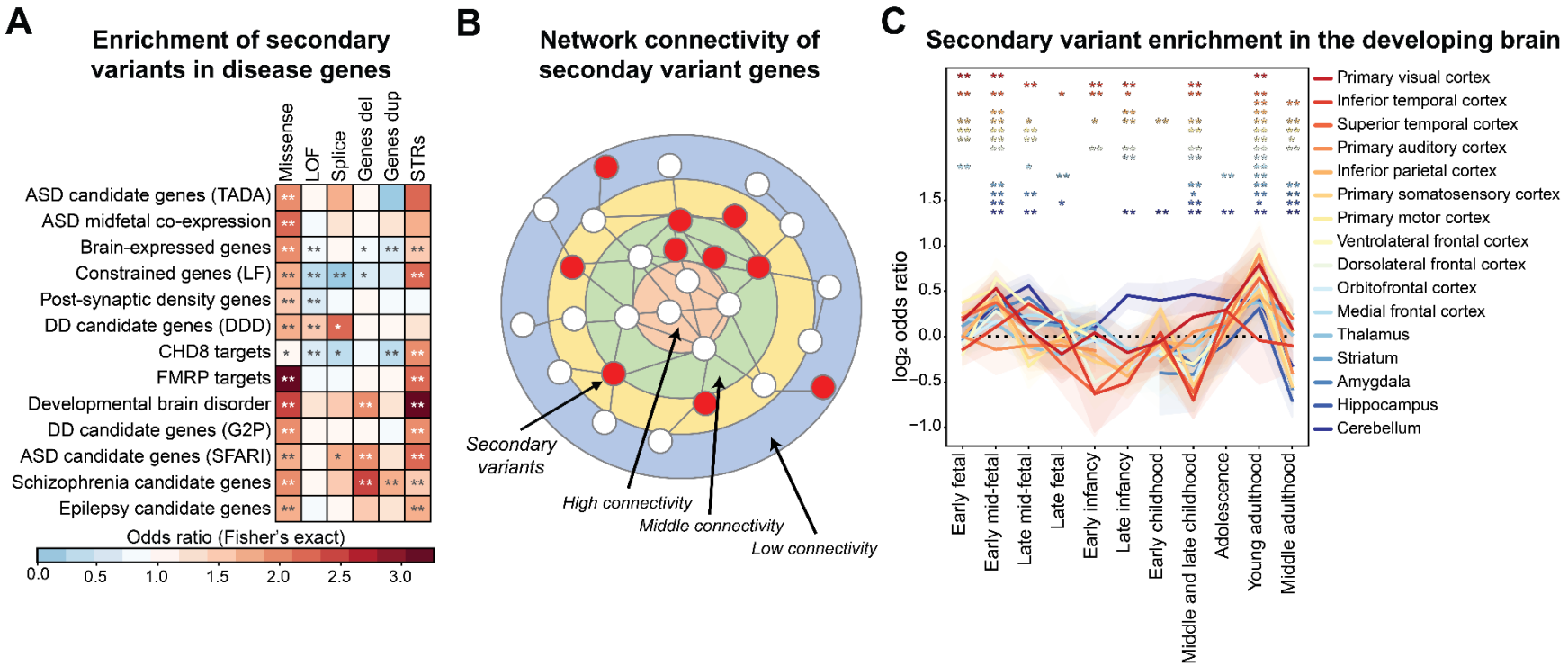


Figure S2. Functional effects of secondary variants observed in 16p12.1 deletion probands (related to Figure 3). (A) Enrichment of secondary variant classes in 16p12.1 deletion probands for sets of genes involved with neurodevelopmental disease and related functions. Fisher's exact test, *p≤0.05, **Benjamini-Hochberg FDR≤0.05. **(B)** Diagram illustrating the distribution of secondary variants (red nodes) in genes with varying connectivity (colored rings) in a brain-specific interaction network. Highly connected genes (light red ring) are depleted for secondary variants, while genes with intermediate connectivity (light green ring) are enriched for variants. **(C)** Line plot shows enrichment (log-odds ratios with 95% confidence intervals; y-axis) of secondary variants in 16p12.1 deletion probands among genes preferentially expressed in 16 brain tissues (colored lines) over 11 developmental timepoints (x-axis). Fisher's exact test, *p≤0.05, **Benjamini-Hochberg FDR≤0.05.

1513

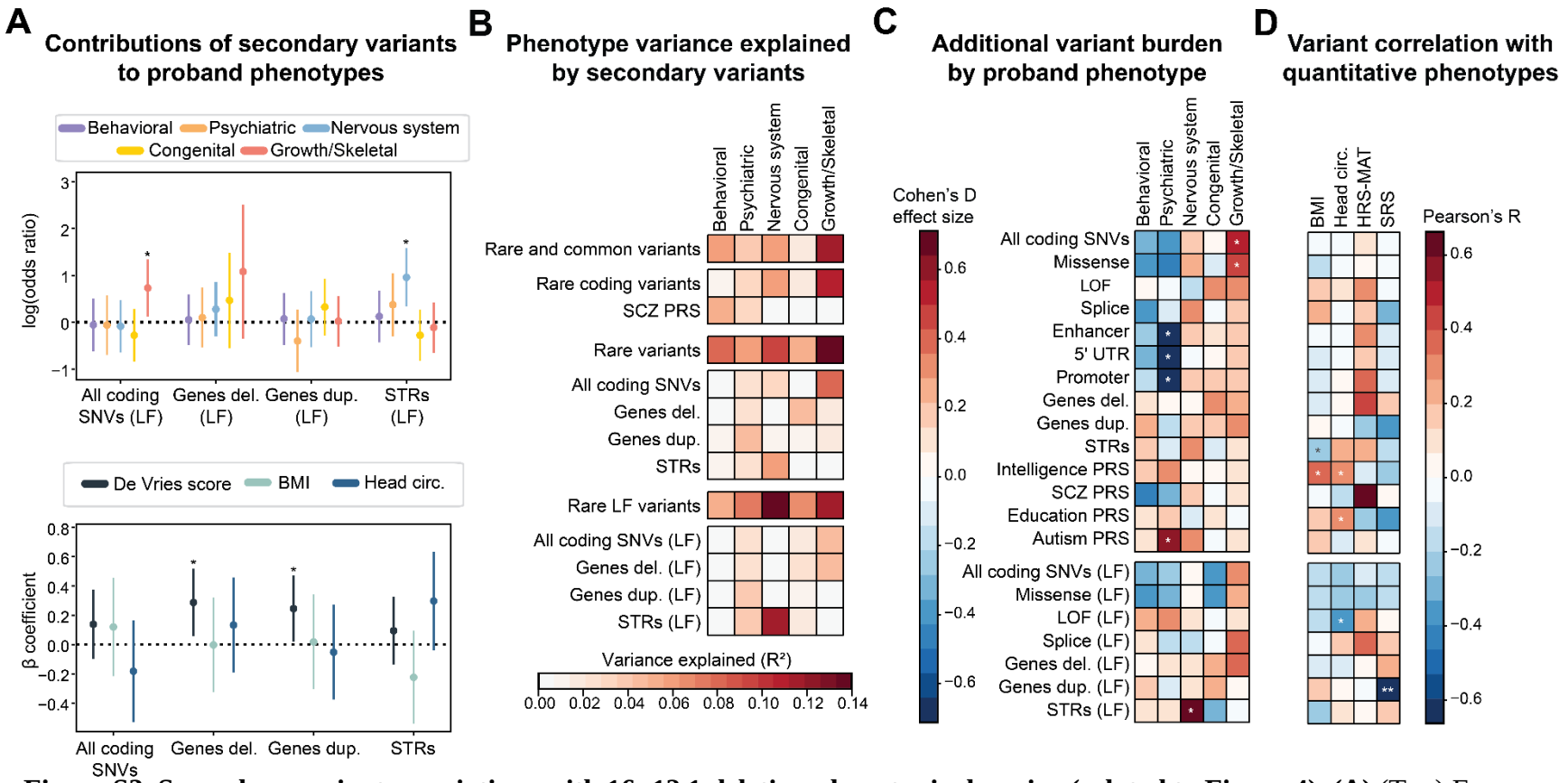


Figure S3. Secondary variant associations with 16p12.1 deletion phenotypic domains (related to Figure 4). (A) (Top) Forest plots show log-scaled odds ratios from logistic regression models for secondary variant burden in constrained genes for probands (n=47-71) with higher complexity scores in five phenotypic domains, compared with probands with lower complexity scores for each domain. *p≤0.05. (Bottom) β coefficients from linear regression models for quantitative phenotypes in probands (n=43-76). *p≤0.05. **(B)** Variance explained by secondary variant burden from logistic regression models (McFadden's pseudo-R^2), both for individual variant classes and joint contributions from combinations of classes. **(C)** Comparisons of secondary variant burden between 16p12.1 deletion probands (n=53-84) with higher and lower complexity scores for each phenotypic domain. Two-tailed t-test, *p≤0.05, **Benjamini-Hochberg FDR≤0.05. **(D)** Pearson correlations between quantitative phenotypes and secondary variant burden in deletion probands (n=9-59). *p≤0.05, **Benjamini-Hochberg FDR≤0.05.

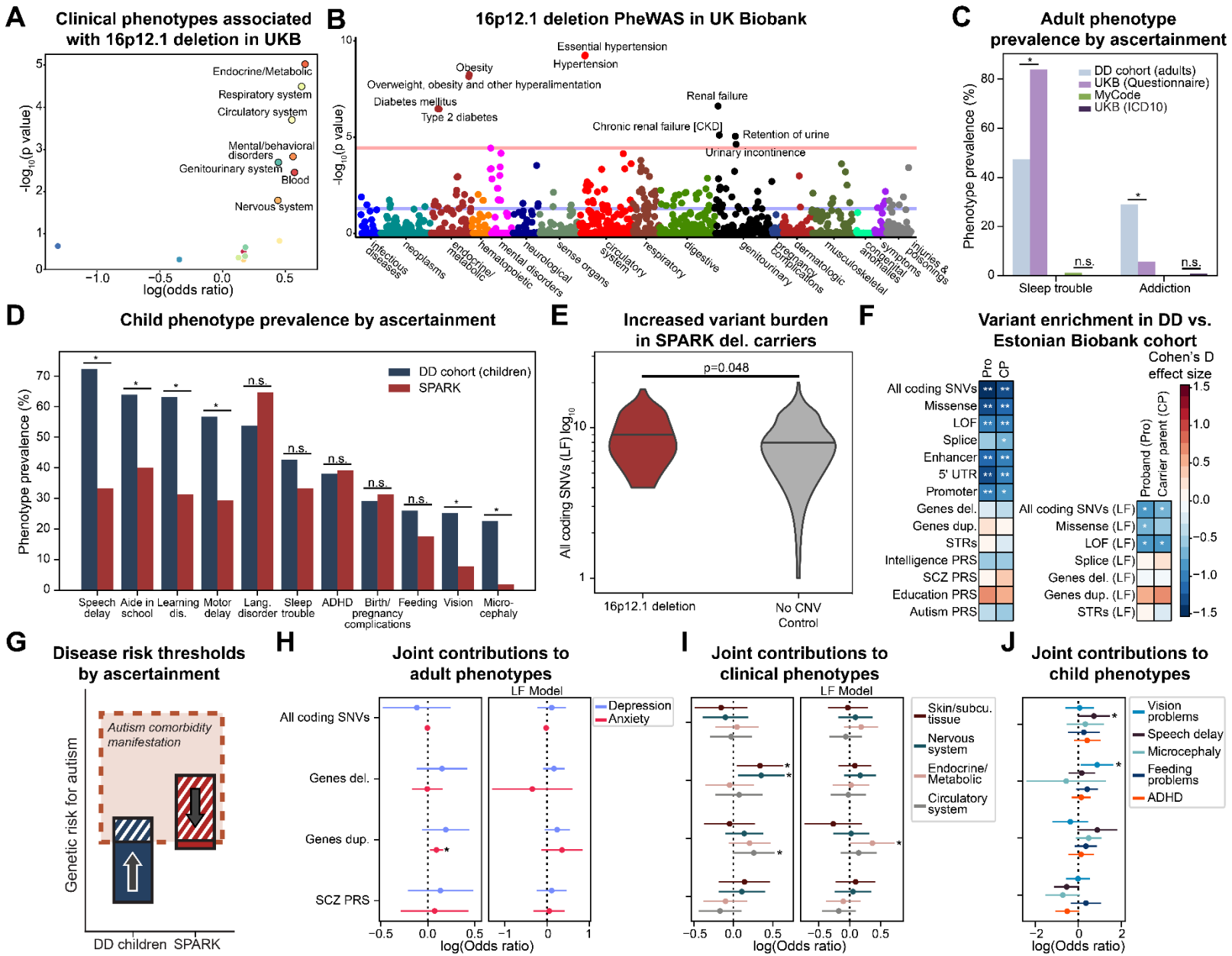


Figure S4. Effects of ascertainment on phenotypes and secondary variant associations in = 16p12.1 deletion carriers (related to Figure 5). (A) Enrichment of select ICD10 chapters from logistic regression models in UK Biobank (UKB) 16p12.1 deletion carriers compared to controls without large rare CNVs (n=3,488). P-values from logistic regression. Labeled points indicate phenotypes with Benjamini-Hochberg FDR≤0.05. **(B)** PheWAS analysis for 16p12.1 deletion carriers in UKB (n=99,363-255,262). Colored circles indicate individual phenotype membership in respective ICD10 chapters. Red line indicates phenome-wide significance (Bonferroni p=0.05) and blue line indicates nominal significance (p=0.05). **(C)** Comparison of sleep disturbance and addiction phenotype prevalence in adults from the DD cohort (n=38) and individuals from UK Biobank (questionnaire n=35-249, ICD10 n=217) and MyCode (n=160). *p≤0.05, Fisher's Exact test. **(D)** Prevalence of developmental and psychiatric phenotypes in children with 16p12.1 deletion from the DD (n=80-151) and SPARK (n=40-51) cohorts. Phenotypes shown were restricted to those present in >20% of probands in either cohort. *p≤0.05, Fisher's Exact test. **(E)** Comparison of SNV (LF) burden in SPARK individuals with 16p12.1 deletion (n=89, left) to age and sex-matched controls without large rare (>500kb) CNVs (n=356, right). P-value from two-tailed t-test. **(F)** Changes in secondary variant burden between probands ("Pro", n=97-99) and carrier parents ("CP", n=54-57) in the DD cohort with 16p12.1 deletion individuals the Estonian Biobank (n=5-8). Blue indicates a depletion in secondary variant burden for Estonian Biobank deletion carriers. One-tailed t-test, *p≤0.05, **Benjamini-Hochberg FDR≤0.05. **(G)** Schematic outlining the proposed relationship between different genetic risk factors in individuals with 16p12.1 deletion across different ascertainments. In cohorts where a majority of participants have a particular disorder, such as autism in SPARK, established risk factors (such as autism PRS) may not show the expected correlations for comorbid features. However, these correlations would be observed in cohorts with different ascertainments (such as the DD cohort). **(H-J)** Forest plots show associations of secondary variants in all genes and constrained genes ("LF Model") with select phenotypes from joint logistic models in **(H)** DD cohort adults, UK Biobank, and MyCode individuals for psychiatric features (n=331); (**I**) UK Biobank and MyCode individuals for clinical phenotypes from EHR data (n=321); and **(J)** children from the DD cohort and SPARK (n=98-125). *p≤0.05. Full results are available in **Table S4G**.

1516

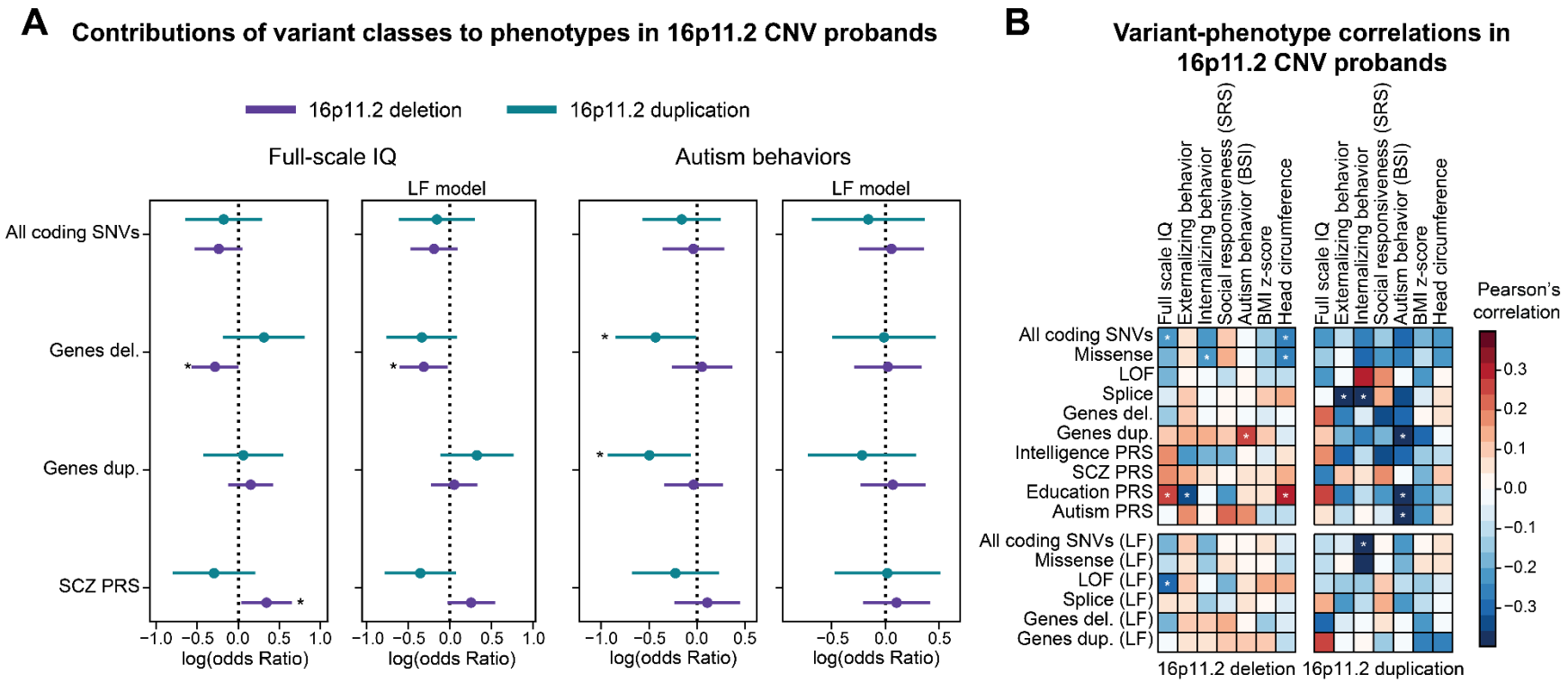


Figure S5. Associations between secondary variants and developmental features of 16p11.2 CNV probands (related to Figure 6). (A) Example forest plots show results from select linear regression models for associations between secondary variant classes and full-scale IQ (left) and autism-related behavior (BSI, right) for probands with the 16p11.2 deletion (n=52-57, purple) and duplication (n=21-25, teal). "LF model" indicates models where rare variants are selected for genes under evolutionary constraint. *p≤0.05. Full results are available in **Table S5A**. **(B)** Pearson's correlations between secondary variant burden and quantitative phenotypes of probands with 16p11.2 deletion (n=58-89) and duplication (n=25-37). *p≤0.05.

1517

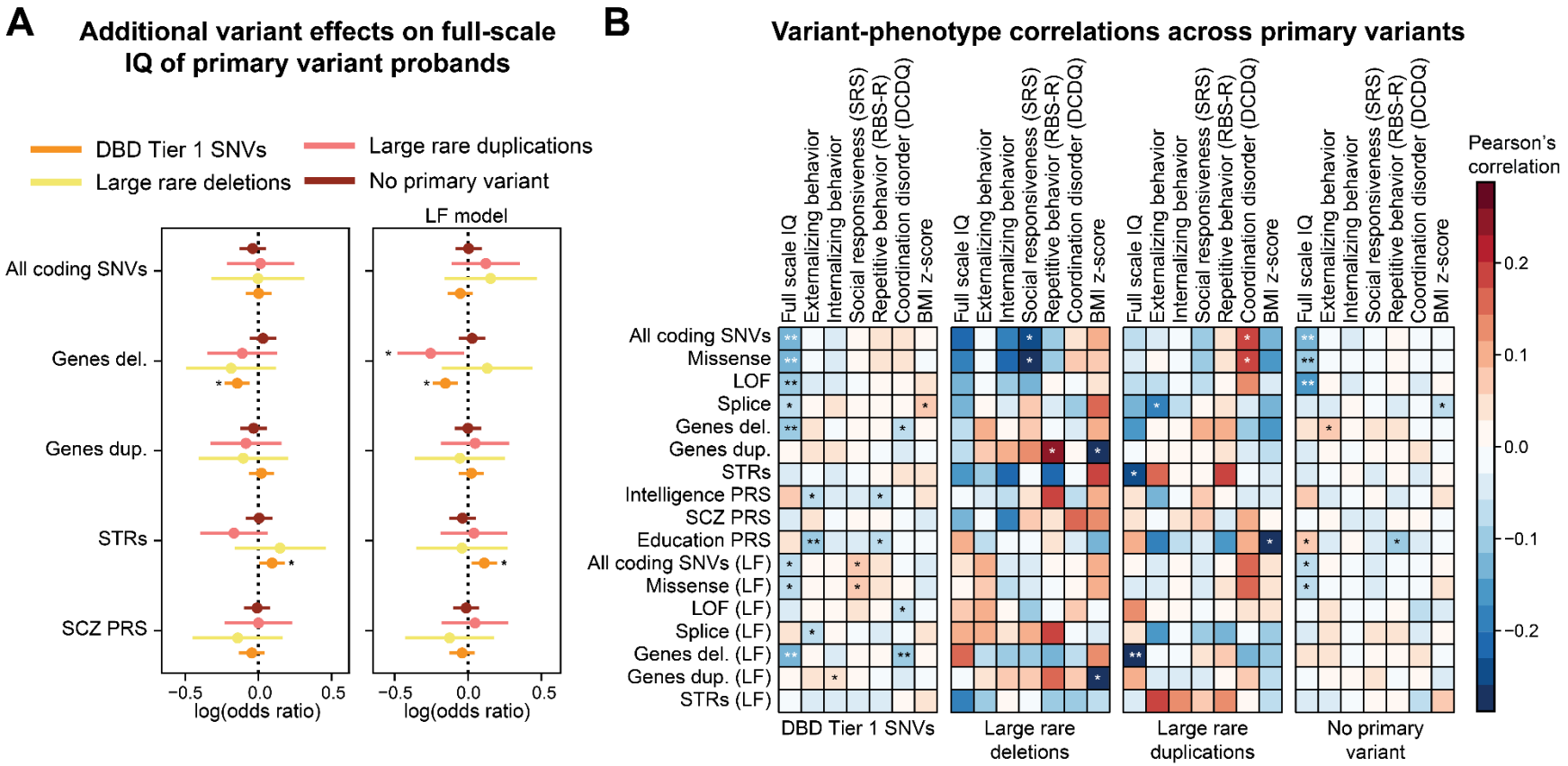


Figure S6. Associations between secondary variants and developmental features of probands with primary variants (related to Figure 6). (A) Example forest plots show results from select linear regression models for associations between secondary variant classes and full-scale IQ for probands with pathogenic SNVs in candidate neurodevelopmental genes (n=660, orange), large, rare deletions (n=51, yellow) and duplications (n=85, pink), and probands without such variants (n=632, red) from the SSC cohort. *p≤0.05. "LF model" indicates models where rare variants are selected for genes under evolutionary constraint. Full results are available in **Table S5A**. **(B)** Pearson's correlations between secondary variant burden and quantitative developmental phenotypes of SSC probands with pathogenic SNVs (n=736-1,236), rare deletions (n=49-78), rare duplications (n=102-148), and probands without such variants (n=671-1,083). *p≤0.05, **Benjamini-Hochberg FDR≤0.05.

SUPPLEMENTAL TABLES

Table S1. Description of the DD cohort and phenotypic data (Excel file; related to Figure 1). **Table S1A** lists all 452 individuals in the DD cohort, including family relationships, age, biological sex, and 16p12.1 deletion status if known. The table also lists secondary variant burden (rare variant counts and PRS), complexity scores for phenotypic domains (child and adult), quantitative measures (BMI, head circumference, IQ, and SRS), and age at developmental milestone achievement for all individuals with available data. **Table S1B** contains the scoring rubric used to calculate complexity scores for phenotypic domains in children. **Table S1C** lists minimum age thresholds used for identifying psychiatric features in pediatric family members. **Table S1D** lists pathogenic variants or deleterious variants in genes associated with disease that were identified in 16p12.1 deletion probands.

Tables S2-S5. Statistical analyses (Excel files). All statistics supplementary tables are linked to the analyses presented in specific figures, which are detailed in the first sheet of each file. The tables list sample sizes, statistic test used, effect sizes/odds ratios, confidence intervals, and p-values with and without multiple testing correction, depending on the analysis. Gene set enrichments and gene lists for specific analyses are listed under separate table headings. Additional data (i.e., GO enrichments) are also provided in some of the files, which are described below.

Table S2. Statistics analysis for Figures 2, 3, S1, and S2 (Excel file).

Table S3. Statistics analysis for Figures 4 and S3 (Excel file). Table S3B lists enriched GO terms for genes with secondary variants among 16p12.1 deletion probands with the five phenotypic domains.

Table S4. Statistics analysis for Figures 5 and S4 (Excel file). Table S4A details how psychiatric phenotypes were matched across questionnaire and EHR (ICD10) datasets across cohorts with different ascertainments.

1549 **Table S5. Statistics analysis for Figures 6, S5, and S6 (Excel file). Tables S6C** and **S6D** list
1550 enriched GO terms for genes with secondary variants among SSC probands with primary SNVs
1551 and CNVs and without primary variants (S6C), and Searchlight probands with 16p11.2 deletions
1552 and duplications (S6D).
1553