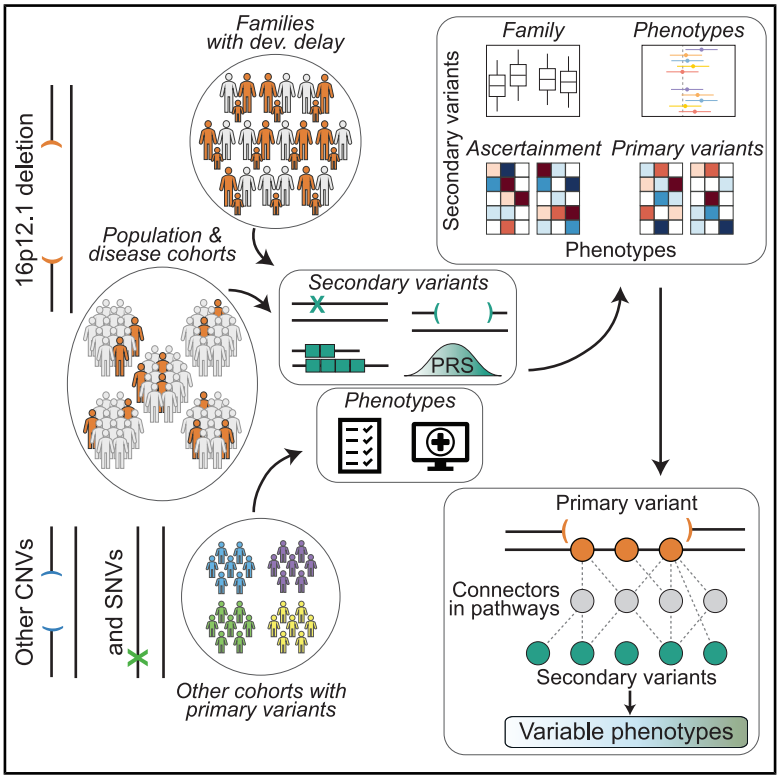
Cell

Genetic modifiers and ascertainment drive variable expressivity of complex disorders

Graphical abstract



Highlights

- Disease-associated variants show extensive phenotypic variability
- The genetic background modifies the expressivity of neurodevelopmental phenotypes
- Modifier effects are disease, population, and primary-variant specific
- Ascertainment bias confounds genotype-phenotype studies

Authors

Matthew Jensen, Corrine Smolen, Anastasia Tyryshkina, ..., Corrado Romano, Joris Andrieux, Santhosh Girirajan

Correspondence

sxg47@psu.edu

In brief

Patterns of secondary variants, influenced by cohort, phenotype, and primary-variant ascertainment, alter phenotypic outcomes in individuals with disease-associated primary variants.

Jensen et al., 2025, Cell *188*, 1–18
December 11, 2025
https://doi.org/10.1016/j.cell.2025.09.012

CellPress

Cell

CellPress
OPEN ACCESS

Article

Genetic modifiers and ascertainment drive variable expressivity of complex disorders

Matthew Jensen,[1,34] Corrine Smolen,[1,34] Anastasia Tyryshkina,[1,34] Lucilla Pizzo,[1,34] Jiawan Sun,[1] Serena Noss,[1] Deepro Banerjee,[1] Matthew Oetjens,[2] Hermela Shimelis,[2] Cora M. Taylor,[2] Vijay Kumar Pounraja,[1] Hyebin Song,[3] Laura Rohan,[1] Emily Huber,[1] Laila El Khattabi,[4] Ingrid van de Laar,[5] Rafik Tadros,[5] Connie R. Bezzina,[6] Marjon van Slegtenhorst,[5] Janneke Kammeraad,[5] Paolo Prontera,[7] Jean-Hubert Caberg,[8] Harry Fraser,[9] Siddharth 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,[14] Kate Pope,[15] Penny Snell,[15] Amber Boys,[15] Paul J. Lockhart,[15,16] Myla Ashfaq,[17] Elizabeth McCready,[18] Margaret Nowacyzk,[18]

(Author list continued on next page)

[1]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA
[2]Autism & Developmental Medicine Institute, Geisinger, Lewisburg, PA 17837, USA
[3]Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA
[4]Paris Brain Institute, Sorbonne Université, Inserm U1127, CNRS UMR 7225, Hôpital Pitié Salpêtrière, 75013 Paris, France
[5]Department of Clinical Genetics, Erasmus MC, University Medical Center Rotterdam, 3000 Rotterdam, the Netherlands
[6]Department of Experimental Cardiology, Amsterdam University Medical Center, University of Amsterdam, 1081 Amsterdam, the Netherlands
[7]Medical Genetics Unit, Hospital Santa Maria della Misericordia, 06156 Perugia, Italy
[8]Centre Hospitalier Universitaire de Liège, Domaine Universitaire du Sart Tilman, 4000 Liège, Belgium
[9]Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester M13 9PL, UK
[10]Manchester Centre for Genomic Medicine, St. Mary's Hospital, Central Manchester University Hospitals, NHS Foundation Trust Manchester Academic Health Sciences Centre, Manchester M13 9WL, UK
[11]Department of Medical Genetics, University and University Hospital Antwerp, 2650 Edegem, Belgium
[12]Greenwood Genetic Center, Greenwood, SC 29646, USA
[13]Department of Human Genetics, Amsterdam Reproduction & Development Research Institute, Amsterdam UMC, Vrije Universiteit Amsterdam, 1105 AZ Amsterdam, the Netherlands
[14]Laboratoire de Genetique Chromosomique et Moleculaire, CHU Dijon, 21000 Dijon, France
[15]Bruce Lefroy Centre, Murdoch Children's Research Institute, Parkville, VIC 3052, Australia
[16]Department of Paediatrics, University of Melbourne, Parkville, VIC 3010, Australia
[17]Department of Pediatrics, McGovern Medical School, University of Texas Health Science Center, Houston, TX 77030, USA
[18]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, ON L8S 4L8, Canada
[19]Research Unit of Rare Diseases and Neurodevelopmental Disorders, Oasi Research Institute-IRCCS, 94018 Troina, Italy

(Affiliations continued on next page)

SUMMARY

Variable expressivity of disease-associated variants implies a role for secondary variants that modify clinical features. We assessed the effects of modifier variants on the clinical outcomes of 2,455 individuals with primary variants. Among 124 families with the 16p12.1 deletion, distinct rare and common variant classes conferred risks for specific developmental features, including short tandem repeats for neurological defects. Network analysis suggested distinct mechanisms involving 16p12.1 genes and secondary variants specific to each proband. Within disease and population cohorts of 976 individuals with the 16p12.1 deletion, we found opposing effects of secondary variants on clinical features across ascertainments. Additional analysis of 1,479 probands with other primary variants, such as the 16p11.2 deletion and *CHD8* variants, and 1,528 probands 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 personalized genomic architecture of complex disorders.

INTRODUCTION

As large-scale sequencing studies uncover increasingly complex links between genomic variants and heritable disorders, identifying genetic etiologies in affected individuals has become more challenging.[1] In contrast to Mendelian disorders caused by single, highly penetrant genes, many disorders show extensive phenotypic heterogeneity, where individuals with the same

CellPress
OPEN ACCESS

Cell
Article

Lucia Castiglia,[19] Ornella Galesi,[19] Emanuela Avola,[19] Teresa Mattina,[20] Marco Fichera,[19,20] Maria Grazia Bruccheri,[19] Giuseppa Maria Luana Mandarà,[21] Francesca Mari,[22] Flavia Privitera,[22] Ilaria Longo,[22] Aurora Curró,[22] Alessandra Renieri,[22] Boris Keren,[23] Perrine Charles,[23] Silvestre Cuinat,[24] Mathilde Nizon,[24] Olivier Pichon,[24] Claire Bénéteau,[24] Radka Stoeva,[25] Dominique Martin-Coignard,[25] Sophia Blesson,[26] Cedric Le Caignec,[27,28] Sandra Mercier,[24] Marie Vincent,[24] Christa L. Martin,[2] Katrin Mannik,[29,30] Alexandre Reymond,[31] Laurence Faivre,[32] Erik Sistermans,[13] R. Frank Kooy,[11] David J. Amor,[15,16] Corrado Romano,[19,20] Joris Andrieux,[33] and Santhosh Girirajan[1,35,*]

[20]Section of Clinical Biochemistry and Medical Genetics, Department of Biomedical and Biotechnological Sciences, University of Catania School of Medicine, 95131 Catania, Italy
[21]Medical Genetics, ASP Ragusa, 97100 Ragusa, Italy
[22]Medical Genetics Unit of the General Hospital of Siena, University of Siena, 53100 Siena, Italy
[23]Department of Genetics, Pitié-Salpêtrière Hospital, Assistance Publique-Hôpitaux de Paris, Sorbonne University, 75013 Paris, France
[24]Medical Genetics Department, CHU Nantes, 44093 Nantes, France
[25]Department of Medical Genetics, Le Mans Hospital, 72037 Le Mans, France
[26]Department of Genetics, Bretonneau University Hospital, 37000 Tours, France
[27]Department of Medical Genetics, CHU Toulouse, 31300 Toulouse, France
[28]Toulouse Neuro Imaging Center, Inserm, UPS, Université de Toulouse, 31300 Toulouse, France
[29]Institute of Genomics, University of Tartu, 50090 Tartu, Estonia
[30]Health2030 Genome Center, Fondation Campus Biotech, 1202 Geneva, Switzerland
[31]Center for Integrative Genomics, Faculty of Biology and Medicine, University of Lausanne, 1015 Lausanne, Switzerland
[32]Center for Rare Diseases and Reference Developmental Anomalies and Malformation Syndromes, CHU Dijon, 21000 Dijon, France
[33]Institut de Genetique Medicale, Hopital Jeanne de Flandre, CHRU de Lille, 59000 Lille, France
[34]These authors contributed equally
[35]Lead contact
*Correspondence: sxg47@psu.edu
https://doi.org/10.1016/j.cell.2025.09.012

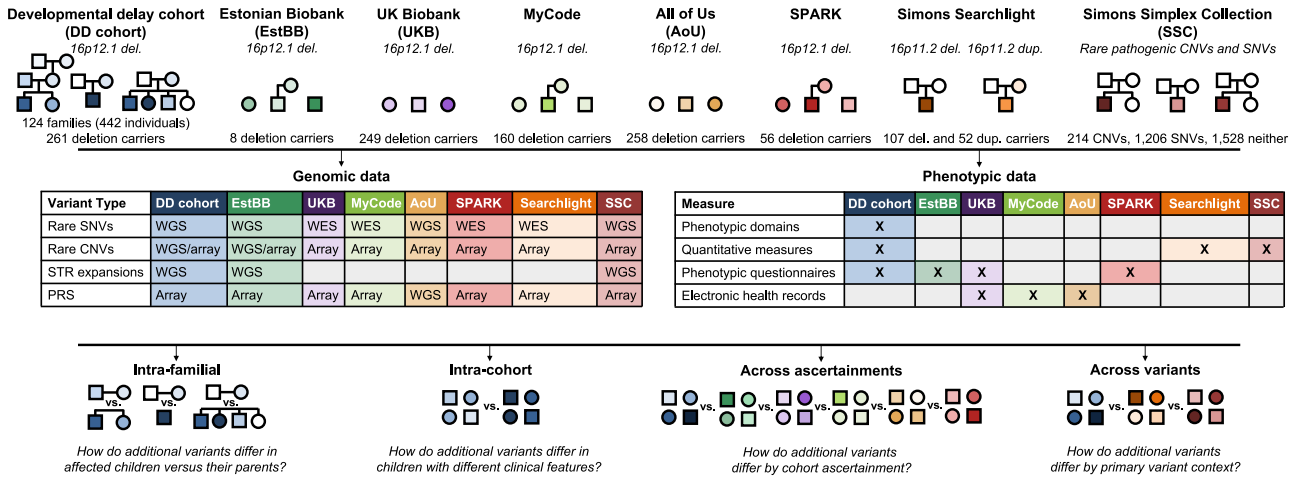
variant exhibit a range of phenotypes with variable penetrance and expressivity.[2,3] Some cases result from multiple genetic diagnoses, where distinct pathogenic variants contribute to independent disorders in the same individual.[4] Variants can also synergistically contribute to new phenotypes, such as seizures among individuals with variants in both *MKS1* (Meckel-Gruber syndrome) and *BBS1* (Bardet-Biedl syndrome).[5] Other cases of heterogeneity occur when the phenotypes of causal variants are modified by secondary variants that do not cause disease themselves.[6] For example, variants in histone modifier genes were enriched among 22q11.2 deletion individuals with variably expressive congenital heart defects.[7] This complexity increases 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 found contributions of polygenic risk to phenotypes of individuals with pathogenic copy-number variants (CNVs),[10,11] such as schizophrenia risk in 22q11.2 deletion individuals.[12–16] Ascertainment may also contribute to variable expressivity, as many pathogenic variants are enriched among individuals across disease ascertainments but lead to other consequences in different populations.[17–21] For example, the autism-associated 16p11.2 deletion[22] is also associated with obesity as well as musculoskeletal, and renal features in the general population.[23,24]

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 genetic modifiers.[3,25] For example, the 16p12.1 deletion is enriched among children with neurodevelopmental features and inherited in >90% of cases from a parent who manifests different psychiatric and cognitive features.[26–28] Phenotypic manifestation among deletion carriers is variable; originally described in children with developmental delay (DD),[26] studies from other populations identified associations with multiple psychiatric and cognitive features.[21,29–31] In fact, severely affected children with the deletion have an increased rare variant burden compared with parents with the deletion, a trend consistent with other primary variants.[26,27,32] This suggests a "multi-hit" model for complex disease etiology, where a primary variant sensitizes an individual for disease, and the clinical outcome is determined by secondary "hits" elsewhere in the genome.[3] However, how specific variant classes modify clinical features across ascertainments and primary variants is not completely understood.

We performed comprehensive analysis of phenotypic and genomic data for 2,455 individuals with primary variants from diverse cohorts (Figure 1). These individuals were selected using a "genetics-first" approach and represent a diverse range of clinical outcomes. We dissected the roles of secondary variant classes on phenotypes in 124 families with the 16p12.1 deletion ascertained for children with DD. We expanded our analysis to uncover phenotypic associations in 976 individuals with the deletion from disease cohorts and healthy populations, 1,479 autism probands who carry other primary variants, and 1,528 autism probands without primary variants. Our results show that variant-phenotype associations depend on both primary and secondary variant context and cohort ascertainment and that the biological pathways disrupted by these secondary variants are unique even among individuals with similar phenotypes. These findings provide a more complete dissection of the genetic etiology of variably expressive pathogenic variants.

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

CellPress
OPEN ACCESS



Variant Type	DD cohort	EstBB	UKB	MyCode	AoU	SPARK	Searchlight	SSC
Rare SNVs	WGS	WGS	WES	WES	WGS	WES	WES	WGS
Rare CNVs	WGS/array	WGS/array	Array	Array	Array	Array	Array	Array
STR expansions	WGS	WGS						WGS
PRS	Array	Array	Array	Array	WGS	Array	Array	Array

Measure	DD cohort	EstBB	UKB	MyCode	AoU	SPARK	Searchlight	SSC
Phenotypic domains	X							
Quantitative measures	X						X	X
Phenotypic questionnaires	X	X	X			X		
Electronic health records			X	X	X			

Figure 1. Secondary-variant-phenotype analyses in 2,455 individuals with primary pathogenic variants

We assessed associations between variant classes and phenotypes in eight cohorts with primary variants. We assessed 124 families with the 16p12.1 deletion (del.), primarily ascertained for children with DD, and 16p12.1 del. carriers from 5 other cohorts, including healthy-biased (UK Biobank [UKB]), clinically-derived (MyCode), population-based (Estonian Biobank and All of Us [AoU]), and single-disorder (SPARK, for autism) ascertainments. We further assessed autism probands with other primary variants, including the 16p11.2 del. or duplication (Simons Searchlight) and large CNVs or rare SNVs in neurodevelopmental genes (SSC). 100 probands in SSC have both primary CNVs and SNVs and are included in both categories (n = 1,320 total probands). Within and across these cohorts, we identified associations for 17 classes of secondary variants with phenotypic features. Figure S1 lists the specific sample sizes used in each analysis. See also Figure S1.

RESULTS

Variability of clinical features in the 16p12.1 deletion

We recruited a cohort of 442 individuals from 124 families with the 16p12.1 deletion ("DD cohort"), including eight three-generation families, primarily ascertained for children with DD (Figures 1 and S1). We analyzed phenotypes from medical records, family interviews, and online assessments for quantitative traits, including non-verbal IQ[33] and social responsiveness scores (SRSs) for autism-related social traits[34] (Table S1A). About 93% of probands (84/90) inherited the deletion, with a slight bias toward maternal inheritance (48/84, 57%), and 70% (87/124) of probands were male. Probands exhibited clinical features grouped across six broadly defined phenotypic domains, including intellectual disability/DD (ID/DD) and behavioral, psychiatric, nervous system, congenital, and growth/skeletal abnormalities (Figure 2A; Tables S1A–S1C). Probands also showed increased childhood developmental and behavioral features (i.e., increased complexity scores; see STAR Methods) compared with their siblings and cousins, while carrier siblings and cousins manifested more features than noncarriers (Figure 2A). Parents who transmitted the deletion ("carrier parents") often manifested psychiatric phenotypes (Figure 2B). Probands had a 1.98 SD decrease in non-verbal IQ ($p = 2.13 \times 10^{-5}$) and a 1.91 SD increase in SRS ($p = 2.59 \times 10^{-7}$) compared with carrier parents (Figure 2C; Table S2A). The average IQ score among 16p12.1 deletion probands was 1.06 SD lower than probands ascertained for autism from the Simons Simplex Collection[33] (SSC) ($p = 0.004$) (Table S2A). 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 ($p = 8.31 \times 10^{-6}$) and 0.38 SD higher than SSC probands ($p = 6.39 \times 10^{-3}$) (Figure 2C; Table S2A). Beyond psychiatric traits, 16p12.1 deletion probands also showed decreased head size ($p = 0.001$) and increased body mass index (BMI, $p = 0.009$) (Figure 2C; Table S2A). Finally, consistent with their ascertainment, probands exhibited significant delays in several developmental milestones[35] compared with their siblings and cousins ($p < 0.05$) (Figure 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.

Patterns of secondary variants within and across families

Using whole-genome sequencing (WGS) and microarray data, we evaluated 17 classes of secondary variants, including rare coding SNVs (missense, splice, and loss-of-function [LoF] variants; see STAR Methods for details), noncoding SNVs (5′ untranslated region [UTR], promoter, and enhancer variants), rare CNVs (deletions and duplications), and short tandem repeat (STR) expansions (defined as repeat length $\geq$ 2 SD than the cohort mean), a subset of which disrupted genes under evolutionary constraint[36] (defined as LOEUF < 0.35 and referred to as "(LF)" variants) (Tables S1A and S1D). We also calculated polygenic risk scores (PRSs) for four cognitive and psychiatric features, including schizophrenia,[37] intelligence,[38] educational attainment,[39] and autism.[40] 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.[41,42] We searched for variants that may contribute to independent genetic diagnoses by assessing whether probands carried additional

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

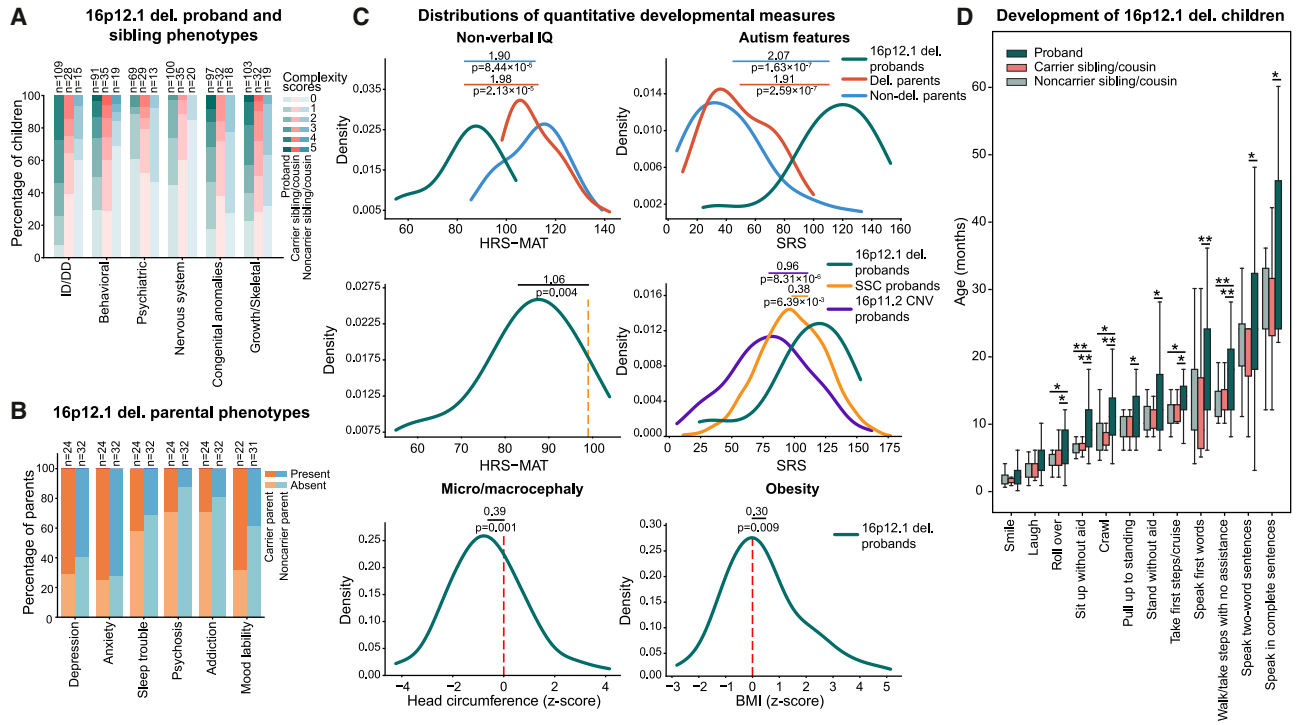


Figure 2. Variably expressive phenotypes of family members with the 16p12.1 del.

(A) Complexity scores for phenotypic domains (see STAR Methods) in 16p12.1 del. probands (n = 69–109), carrier siblings/cousins (n = 28–35), and noncarrier siblings/cousins (n = 13–20) (numbers vary by data availability).

(B) Neuropsychiatric phenotypes in carrier (orange, n = 22–24) and noncarrier parents (blue, n = 31–32) of 16p12.1 del. probands.

(C) Distributions of quantitative phenotypes observed in 16p12.1 del. probands. (Top) Non-verbal IQ (Hansen research services matrix adaptive test [HRS-MAT]) and social responsiveness scores (SRSs) in probands (green, n = 10, 27, respectively) compared with carrier (red, n = 17, 21, respectively) and noncarrier parents (blue, n = 20, 26, respectively). (Middle) Non-verbal IQ and SRS distributions in 16p12.1 probands and SSC (SRS n = 2,844, yellow; HRS-MAT mean from Hansen[33]) and Searchlight probands with 16p11.2 CNVs (n = 139, purple). (Bottom) Head circumference (n = 64) and BMI Z scores (n = 67) in del. probands. Red vertical lines, Z score = 0. p values, Mann-Whitney or one-sample t tests.

(D) Distribution of ages of attainment of developmental milestones in probands (n = 13–33), carrier siblings and cousins (n = 16–18), and noncarrier siblings and cousins (n = 11–15). Boxes, quartiles; bars, range. One-tailed t test, *$p \leq 0.05$, **Benjamini-Hochberg FDR ≤ 0.05. Table S1A lists all data for DD cohort samples.

CNVs with previous associations with disease,[32] variants in ClinVar[43] associated with developmental disorders, or LoF variants in genes from two clinically relevant databases: Developmental Brain Disorder (DBD) database[44] and SFARI Gene.[45] Overall, 31% of probands (31/99) had at least one such variant, including seven probands with ClinVar-defined pathogenic variants (Figure S2A; Tables S1E and S1F). A subset of these cases represented probands with multiple genetic diagnoses.[4] For instance, one proband had an 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[27] as well as ID/DD and craniofacial defects. We found no differences in variant burden between probands with and without additional pathogenic variants, except for splice variants (p = 0.04) and schizophrenia PRS (p = 0.03) (Figure S2B; Table S2C). We further found two families in which the same additional variant, inherited from the carrier parent, was present in the proband and a noncarrier sibling (Figure S2C). In both families, the proband and sibling have different symptoms despite carrying the same secondary variant, highlighting the variable expressivity of these variants. Additionally, 17 probands had STR expansions in spinocerebellar ataxia genes[47] such as *ATXN7* and *CACNA1A*. 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 missense variant in *POLR3E* on the non-deleted allele in a proband with global DD and multiple congenital defects, such as bilateral club feet and natal teeth. Although these variants explained a subset of observed phenotypes, we did not find cases where a single variant accounted for all phenotypes observed in an individual proband.

We next identified patterns of secondary variants in probands compared with their parents (Figure S2D). Probands carried more missense (LF) variants (p = 0.034, false discovery rate [FDR] = 0.235) and a higher schizophrenia PRS than their carrier parents (p = 0.009, FDR = 0.253) and had increased LoF (LF) variants (p = 0.028, FDR = 0.235) compared with their noncarrier parents (Figures 3A and S2E; Table S2D). We found no differences in the burden of secondary variants inherited from either parent (Figure S2F; Table S2D). In three-generation families, we observed increases in variant burden across generations

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

CellPress
OPEN ACCESS

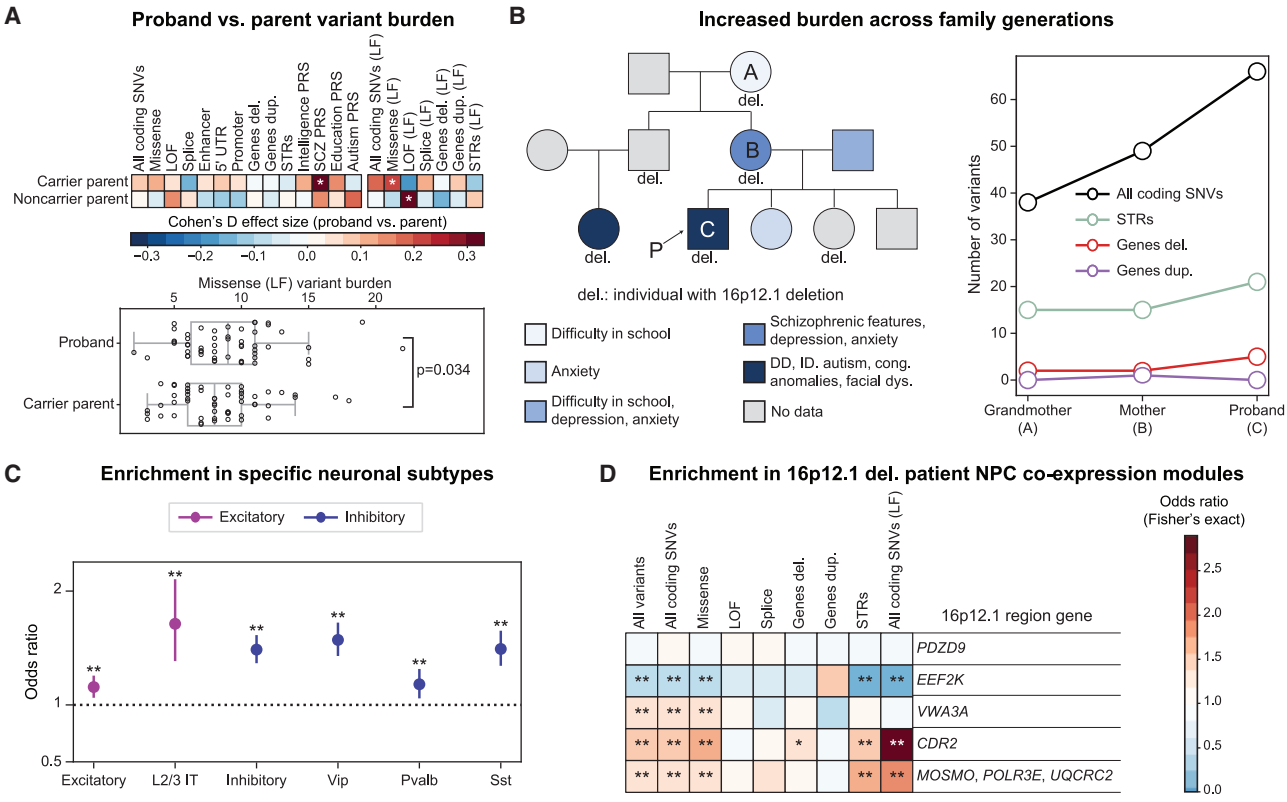


Figure 3. Secondary variants contribute to phenotypic variability within 16p12.1 del. families

(A) Cohen's D (top) for differences in secondary variant burden between probands and parents ($n = 49$–54 pairs). *$p \leq 0.05$, paired one-tailed (rare variants) or two-tailed (PRS) t test. Red, increased burden in probands. (Bottom) Increased burden of missense (LF) variants in probands compared with carrier parents. Box, quartiles; bars, range excluding outliers defined by IQR.

(B) (Left) Phenotypes of individuals with 16p12.1 "del." and family members in family GL_077. (Right) Increased burden of rare variants across generations in family GL_077. Two siblings had no developmental data but did not exhibit psychiatric features.

(C) Enrichment of genes with SNVs in probands for those preferentially expressed in neuronal classes and sub-classes (colored by main class) in the adult motor cortex. Bars, 95% CIs. Full results in Table S2H.

(D) Enrichment of genes with secondary variants in probands for 16p12.1 gene co-expression modules identified from WGCNA of NPC transcriptome data from 16p12.1 del. individuals. Modules are named for their constituent 16p12.1 genes.

In (C) and (D), Fisher's exact test, *$p \leq 0.05$, **Benjamini-Hochberg FDR ≤ 0.05.

See also Figures S2, S3, and S4.

corresponding with increased phenotypic severity among deletion carriers (Figure S3). For instance, the carrier grandmother in family GL_077 had mild cognitive features, while the carrier mother manifested psychiatric features, and the proband had neurodevelopmental features (Figure 3B).

We further profiled putative functions of the secondary variants and found that missense variants were enriched for brain-expressed,[48] constrained,[36] and post-synaptic density genes[48] ($p \leq 1.63 \times 10^{-11}$, FDR $\leq 1.27 \times 10^{-10}$), whereas genes with LoF variants were depleted in these gene sets ($p \leq 0.005$, FDR ≤ 0.016) (Figure S4A; Table S2F). This suggests that LoF variants in essential genes may not be tolerated, particularly with the deletion, while less severe variants in these genes may contribute to neurodevelopmental features seen in probands.[25] Secondary variant genes were also preferentially expressed in several brain regions during early and mid-fetal development,[49] including the ventrolateral frontal cortex ($p = 3.32 \times 10^{-8}$, FDR $= 3.74 \times 10^{-7}$) and hippocampus ($p = 2.33 \times 10^{-7}$, FDR $= 2.19 \times 10^{-6}$) (Figure S4B; Table S2G). SNVs were enriched for genes preferentially expressed across multiple neuronal classes in the adult motor cortex,[50] including excitatory ($p = 1.22 \times 10^{-4}$, FDR $= 1.95 \times 10^{-3}$) and inhibitory ($p = 2.04 \times 10^{-24}$, FDR $= 1.14 \times 10^{-22}$) neurons (Figure 3C; Table S2H). Additionally, secondary variants were enriched in genes co-expressed with 16p12.1 deletion genes in neural progenitor cells (NPCs) derived from 12 individuals in three 16p12.1 deletion families[51] (Figure 3D; Table S2I). For example, variants were enriched among modules of genes co-expressed with *VWA3A* ($p = 0.017$, FDR $= 0.045$); *CDR2* ($p = 1.27 \times 10^{-22}$, FDR $= 1.43 \times 10^{-21}$); and *MOSMO*, *POLR3E*, and *UQCRC2* ($p = 4.17 \times 10^{-7}$, FDR $= 1.88 \times 10^{-6}$) (Figure 3D; Table S2I). This suggests that secondary variants may interact with 16p12.1 genes in developing neurons at a cell stage that represents a convergent point for neurodevelopmental disorders.[52] Overall, our results indicate that a diverse range of biologically relevant modifiers contribute to the variable phenotypes observed in 16p12.1 deletion probands.

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

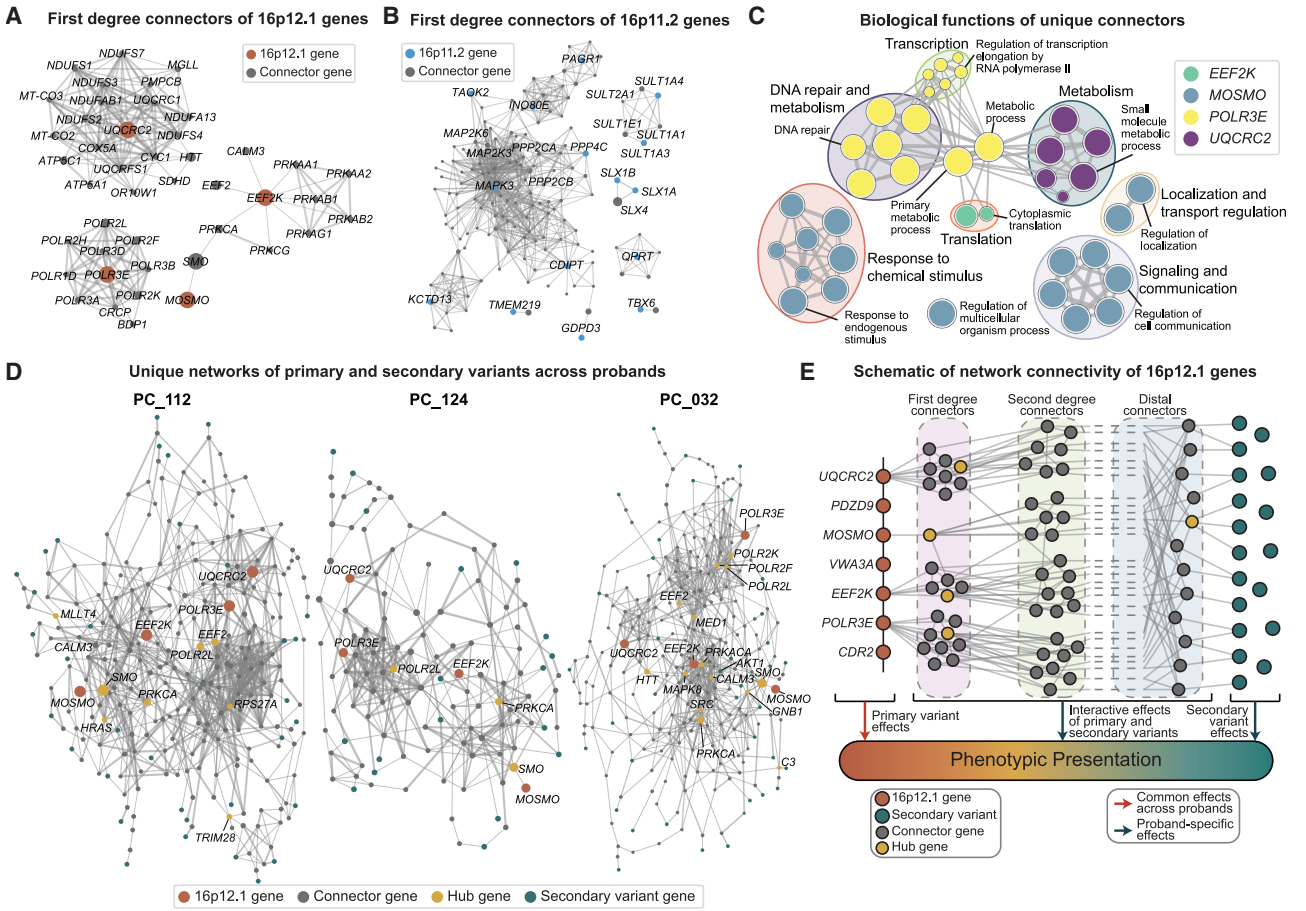


Figure 4. Connectivity of primary and secondary variants in a gene interaction network

(A and B) Network diagrams for first-degree connector genes (gray) of (A) 16p12.1 genes (orange) and (B) 16p11.2 genes (blue). For (B), only 16p11.2 genes and connector genes that interact with multiple 16p11.2 genes are labeled.

(C) Enriched GO biological process terms for unique connector genes of each 16p12.1 gene. Node size, number of genes in the genome annotated for each term; edge width, number of genes shared between terms.

(D) Networks for three 16p12.1 del. probands show genes along shortest paths ("hub genes," yellow; connector genes, gray) between 16p12.1 genes (red) and secondary variant genes (teal); secondary variant genes without connections to 16p12.1 genes are not shown. For (A), (B), and (D), edge widths, confidence of interaction between genes; node size, number of shortest paths containing each gene. Only CNV genes with paths to secondary variant genes are shown.

(E) Schematic detailing connectivity of 16p12.1 del. and secondary variant genes. First- and second-degree connectors are often 16p12.1 gene-specific and shared across probands. Distal connectors are secondary-variant specific and shared across all 16p12.1 genes.

See also Figure S4.

Cellular mechanisms perturbed by the 16p12.1 deletion depend on secondary variant context

We next sought to examine biological mechanisms underlying potential interactions of 16p12.1 genes and secondary variants in probands. Our previous functional studies suggested the seven constituent genes within the deletion region disrupt independent cellular pathways and functions, such as *UQCRC2* for mitochondrial respiration and *POLR3E* for tRNA transcription.[41] To further assess this, we determined the shortest paths between each 16p12.1 gene and all secondary coding variant genes in the STRING protein interaction network.[53] We observed no overlap between first-degree connector genes (0/40 genes) for each 16p12.1 gene and little overlap (5.8%, 17/295 genes) when including second-degree connector genes (Figures 4A and S4C; Table S3A). To understand this in the context of other disorders, we also examined the connectivity of genes within the 16p11.2 CNV region, whose deletion carriers typically exhibit less phenotypic heterogeneity.[25] We found that 16p11.2 genes showed stronger connectivity with each other (Figure 4B; Table S3B), with 4.20% (6/143) first-degree connector genes shared between multiple 16p11.2 genes and a much larger overlap (29.6%, 212/717) when also considering second-degree connector genes (Table S3B).

Expanding on the observation of limited connections between 16p12.1 genes, we assessed Gene Ontology (GO) enrichment for the unique connector genes of each 16p12.1 gene. We observed enrichments for distinct terms consistent with the biological functions of each 16p12.1 gene (Figure 4C; Table S3C). For example, *POLR3E* connectors were enriched for transcription and DNA metabolism functions, consistent with *POLR3E*'s

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

role in RNA polymerase III,[54] and *UQCRC2* connectors were associated with metabolism, in line with *UQCRC2*'s role in the mitochondrial respiratory chain.[55] As secondary variant genes showed fewer connections in the network (Figure S4D; Table S3D), distal connectors were often shared among 16p12.1 genes (Figure S4E), as paths to secondary variants from all 16p12.1 genes traversed the same connector genes in each proband. We propose that the diversity of pathways affected by 16p12.1 genes increases the potential for interactions across a broader genomic landscape, contributing to the phenotypic variability associated with the deletion.

We next assessed network connectivity patterns within individual probands to identify frequent "hub genes" (see STAR Methods) that potentially mediate interactions between 16p12.1 and secondary variant genes (Figure 4D). We found a combination of 16p12.1 gene-specific hub genes shared across probands (such as *SMO*, present in all paths to *MOSMO*) and hub genes that differ across probands. To exemplify this, we compared the networks of PC_112, P1C_124, and PC_032, who have autism spectrum disorder (ASD) without ID/DD, both ID/DD and ASD, and ID/DD without ASD, respectively (Figure 4D). Each of these probands had distinct hubs in their networks. For example, *HRAS*, a Ras pathway member associated with ASD,[56,57] and *MLLT4*, a Ras target involved in cell adhesion,[58] were hubs for PC_112 but not P1C_124. For PC_032, these genes act as connectors but not hubs. Similarly, *AKT1*, part of the neurodevelopment-associated phosphatidylinositol 3-kinase (PI3K)-AKT-mTOR pathway,[59] acts as a hub for PC_032 but is not a connector for the other two probands. These examples represent cases where probands with distinct features (ASD for PC_112 and ID/DD for PC_032) have different hub genes that are not shared, even in a third proband with overlapping features (PC_124 with ID/DD and ASD). This suggests that network paths and topologies mediating the effects of primary and secondary variants may be proband-specific, even across probands with similar phenotypes. Thus, the disruption of distinct functions and pathways by individual-specific secondary variants may underlie the range of phenotypes observed in 16p12.1 deletion patients (Figure 4E).

Distinct secondary variant classes contribute to specific phenotypic outcomes

Given that secondary variants disrupt unique pathways among 16p12.1 deletion carriers, we used logistic regression models to assess the effects of rare variant classes and PRSs on phenotypic domains in probands. Rare coding variants contributed to nervous system (log-odds ratio [logOR] = 0.633, p = 0.034, FDR = 0.518) and growth/skeletal features (logOR = 0.901, p = 0.005, FDR = 0.208) (Figure 5A; Table S4A). Specifically, STRs were associated with nervous system features (logOR = 0.600, p = 0.035, FDR = 0.518), while SNVs were associated with growth/skeletal features (logOR = 0.878, p = 0.007, FDR = 0.208) (Figures 5A and S5A; Table S4A). In contrast, schizophrenia PRS was negatively associated with behavioral phenotypes (logOR = −0.566, p = 0.045, FDR = 0.584) (Figure 5A; Table S4A). Combined variant models explained an average of 9% variance (McFadden's pseudo-R^2; range 5%–13%) for each phenotypic domain and showed better performance than models built using individual variant classes (average of 2% explained variance) (Figure S5B; Table S4A). These estimates 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 (Figure S5B; Table S4A). Orthogonal burden tests identified fewer rare variants in enhancers, promoters, and 5′ UTR elements ($p \leq 0.025$, FDR = 0.385), as well as an increased autism PRS (p = 0.028, FDR = 0.385), among probands with psychiatric features (Figure S5C; Table S4B).

Linear regression models testing the effects of variant classes on quantitative traits revealed negative associations of SNVs (LF) with head circumference Z scores ($\beta = -0.361$, p = 0.037, FDR = 0.553) (Figures 5B and S5A; Table S4A). Secondary CNVs were associated with increased de Vries scores, a quantitative assessment for global developmental features[60] (deletions: β = 0.281, p = 0.015, FDR = 0.447; duplications: β = 0.239, p = 0.035, FDR = 0.553). Correlation analyses revealed that intelligence and educational attainment PRSs were positively correlated with head circumference (education r = 0.318, p = 0.026, FDR = 0.582; intelligence r = 0.287, p = 0.045, FDR = 0.582), while duplications (LF) negatively correlated with social responsiveness deficits (r = −0.605, p = 0.001, FDR = 0.030) (Figure S5D; Table S4C). These results suggest that the modifying roles of secondary variant classes differ across phenotypes.

Additionally, secondary variants in probands with similar phenotypes showed specific enrichments for biological functions and pathways. Secondary variants from probands with growth/skeletal defects were enriched for neurogenesis genes, whereas variants in probands with behavioral features were enriched in Hedgehog signaling pathway genes (Figures 5C and S5E; Table S4D). Some of these enrichments were shared with 16p12.1 genes, such as cell differentiation and Hedgehog signaling for *MOSMO*.[61] Further, probands with specific phenotypes showed an increased burden of rare variants in key neuronal genes, such as candidate autism[45] genes in probands with congenital anomalies (p = 0.003, FDR = 0.182) or nervous system features (p = 0.006, FDR = 0.201) and developmental brain disorder genes[44] in probands with nervous system features (p = 0.043, FDR = 0.316) (Figure 5D; Table S4E). The variety of pathways affected in probands with distinct phenotypes highlights the increased complexity of biological functions disrupted in 16p12.1 deletion carriers.

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 of affected individuals and those drawn from the general population.[62,63] We next compared the phenotypic effects of the 16p12.1 deletion in 976 individuals across multiple cohorts, including 245 pediatric and adult deletion carriers from the DD cohort and four independent cohorts with distinct ascertainments: Simons Powering Autism Research for Knowledge (SPARK) (n = 56), where families were ascertained for probands with autism features,[64] the healthy-biased UK Biobank[65,66] (UKB; n = 249), the hospital-derived Geisinger MyCode[67] (MyCode; n = 160), and the diverse population-based All of Us[68]

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

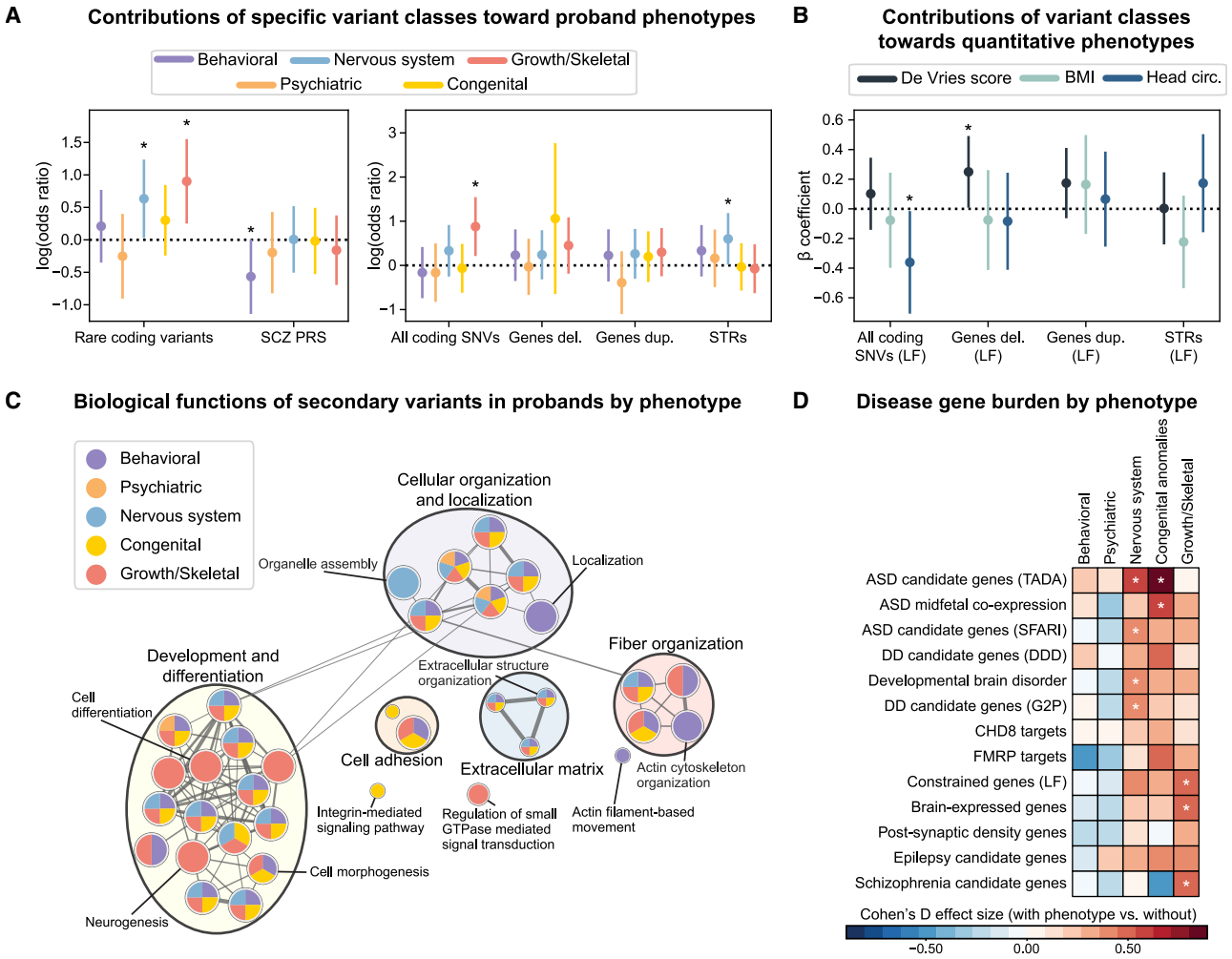


Figure 5. Secondary variant associations for phenotype domains of 16p12.1 del. probands

(A) Log-scaled odds ratios from regression models of secondary variant burden in probands with higher and lower complexity scores for five domains (n = 47–71). Bars, 95% CIs. *$p \leq 0.05$. "LF" model shown in Figure S5A.

(B) β coefficients from regression models for effects of secondary LF variant burden on quantitative phenotypes (n = 43–76). *$p \leq 0.05$. Bars, 95% CIs. Unconstrained model results shown in Figure S5A.

(C) GO biological process terms enriched among secondary variants in probands with each domain. Node size, number of genes annotated for the term; edge width, number of genes shared between terms.

(D) Comparison of secondary variant burden in developmental and neurological genes in probands with (n = 23–67) and without (n = 12–36) phenotypic domains. *$p \leq 0.05$, one-tailed t test.

See also Figure S5.

(AoU; n = 258) (Figures 1 and S1). To compare phenotype prevalence across cohorts, we harmonized electronic health records (EHRs) and clinical questionnaire responses (Table S5A). As expected, the prevalence of neuropsychiatric phenotypes varied across the cohorts (Figures 6A, S6A, and S6B; Table S5B). For example, anxiety symptoms were more common in adults from the DD cohort compared with UKB ($p = 9.80 \times 10^{-5}$) (Figure 6A; Table S5B), likely reflecting ascertainment biases for severely affected family members in the DD cohort compared with healthy volunteers in UKB.[66] Anxiety was also more common in carriers from MyCode ($p = 1.02 \times 10^{-4}$) and AoU ($p = 2.73 \times 10^{-5}$) than those in UKB (Figure 6A; Table S5B).

Assessment of UKB individuals with the deletion showed enrichment for multiple phenotypes, including circulatory, endocrine, and genitourinary features (FDR $\leq$ 0.012) (Figure S6C; Table S5C). PheWAS analysis further revealed enrichment of obesity- and kidney-related features, including type 2 diabetes, obesity, chronic renal failure, and a previously reported association with hypertension[21] ($p \leq 1.78 \times 10^{-6}$) (Figure S6D; Table S5D). Due to these enrichments, we directly assessed the effect of the 16p12.1 deletion on BMI in UKB. The deletion had a relatively large effect on BMI (β = 1.23, p = 0.006) (Figure 6B; Table S5E), confirming previous findings,[69] and showed only additive effects with environmental factors (interaction $p \geq 0.06$), including lifestyle factors

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

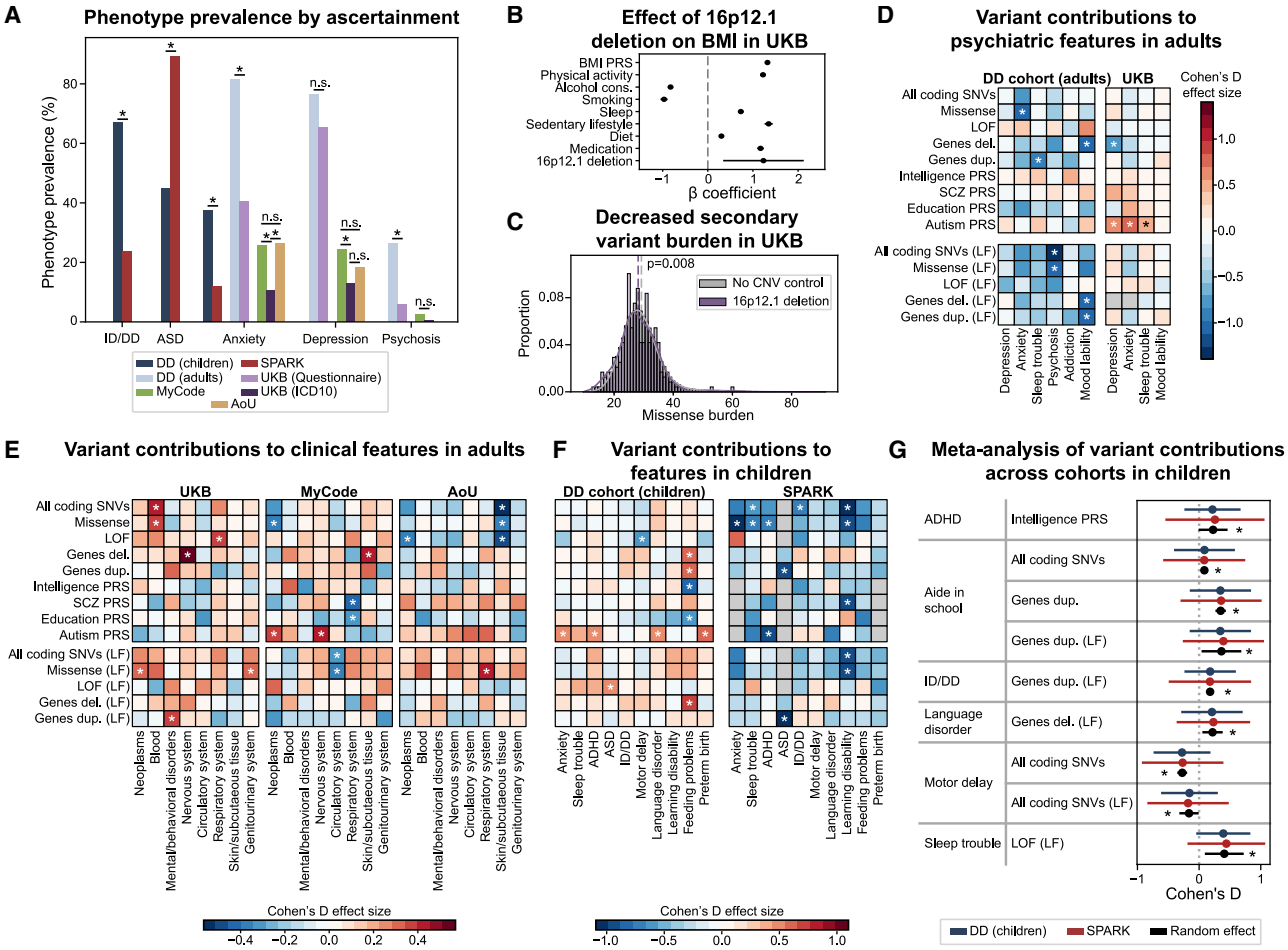


Figure 6. Effects of ascertainment on genotype-phenotype associations of 16p12.1 del.

(A) Phenotype frequency among individuals with 16p12.1 del. from five ascertainments: DD cohort (adults $n = 38$, children $n = 93–151$), SPARK ($n = 51–56$), UKB (questionnaire $n = 52$, ICD10 $n = 229$), MyCode ($n = 160$), and AoU ($n = 185$). Fisher's exact test, $^*p \leq 0.05$. Percentage of AoU samples with psychosis is masked due to low sample size ($n < 20$).

(B) Effect sizes and 95% CIs from regression models of the 16p12.1 del., lifestyle, environmental, and genetic factors on BMI in UKB ($n = 317,978$). Only variables with significant effects ($p \leq 0.05$) shown; full results in Table S5E.

(C) Missense burden in 16p12.1 del. carriers ($n = 239$, purple) and controls without large (>500 kb) rare CNVs ($n = 60,228$, gray) from UKB. p value, one-tailed t test. Dashed line, group mean.

(D–F) Associations of secondary variant burden and phenotypes in 16p12.1 del. individuals from multiple cohorts. Two-tailed t test, $^*p \leq 0.05$.

(D) Psychiatric phenotypes derived from clinical questionnaires in the DD cohort ($n = 24–31$) and UKB ($n = 48–249$).

(E) Select aggregated clinical phenotypes in UKB ($n = 195–229$), MyCode ($n = 143–159$), and AoU ($n = 93–185$).

(F) Developmental phenotypes in the DD cohort ($n = 37–125$) and SPARK ($n = 21–56$).

(G) Meta-analysis of associations from (F) in the DD and SPARK cohorts. Meta-analyses for adult cohorts are shown in Figures S6H and S6I. Bars, 95% CIs. $^*p \leq 0.05$. See also Figure S6.

and antidepressant and antipsychotic medications (Table S5E). This effect is in line with the increased BMI we observed in DD cohort probands (Figure 2C), suggesting that some effects of the deletion may be consistent across ascertainments.

We next investigated how patterns of secondary variants differed across cohorts by comparing the burden of secondary variants in 16p12.1 deletion carriers to matched controls without large CNVs. Deletion carriers in UKB showed decreased burden of missense variants compared with controls ($p = 0.008$) (Figure 6C; Table S5F), while carriers in AoU had an increased burden of SNVs compared with controls ($p = 3.91 \times 10^{-5}$) (Figure S6E; Table S5F). Thus, we observed a higher rare variant burden in deletion carriers compared with controls in cohorts with a disease-biased ascertainment (AoU[70]) and reduced burden in cohorts with healthy-biased ascertainment (UKB[66]). This suggests that deletion carriers in healthy cohorts may require reduced genetic burden to remain unaffected, whereas carriers in disease-biased cohorts may need additional genetic burden to meet disease thresholds. We also directly compared variant burden between deletion carriers in the DD cohort with data from eight deletion carriers in the Estonian Biobank.[71] Estonian Biobank carriers showed a depletion of several variant classes,

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

including missense ($p \leq 0.002$, FDR ≤ 0.011), enhancer ($p \leq 0.009$, FDR ≤ 0.026), and 5′ UTR SNVs ($p \leq 0.007$, FDR ≤ 0.023), compared with probands and carrier parents in the DD cohort (Figure S6F; Table S5G).

We next assessed how the relationship between secondary variants and phenotypes varied by ascertainment by comparing the burden of variant classes in individuals with and without specific phenotypes across cohorts. We found differing trends for psychiatric features from self-reported questionnaires for adults in the DD cohort and UKB, with some similar patterns emerging (Figure 6D; Table S5H). For example, autism PRS was associated with depression, anxiety, and sleep trouble ($p \leq 0.039$, FDR ≤ 0.352) in UKB (Figure 6D; Table S5H). Conversely, rare variants were negatively associated with psychiatric features in both UKB and DD cohort adults, including deletions with depression in UKB ($p = 0.018$, FDR = 0.211) and mood lability in DD (i.e., more secondary deletions decreased the likelihood of psychiatric features) ($p = 0.030$, FDR = 0.223) (Figure 6D; Table S5H). These data suggest potential opposing effects of PRSs and rare variants on psychiatric features in adults with the 16p12.1 deletion. We further compared secondary variant profiles for clinical features represented by ICD10 chapters in UKB, MyCode, and AoU. In UKB, nervous system features were associated with deletions ($p = 5.98 \times 10^{-4}$, FDR = 0.065), while mental health features were associated with duplications (LF) ($p = 0.012$, FDR = 0.612) (Figure 6E; Table S5I). In MyCode, autism PRS was associated with both nervous system features and cancer ($p \leq 0.039$, FDR ≤ 0.843) (Figure 6E; Table S5I). In AoU, we observed negative associations of all coding SNVs and LoF variants with skin/subcutaneous tissue phenotypes ($p \leq 0.001$, FDR = 0.092) and a positive association of missense (LF) variants with respiratory system phenotypes ($p = 0.013$, FDR = 0.814) (Figure 6E; Table S5I). These differences reflect ascertainment differences between the cohorts, potentially due to healthcare system differences, phenotyping modalities, or biases stemming from healthy volunteers versus affected individuals.

We further observed differences in phenotype-variant associations among children with the deletion from the DD and SPARK cohorts. Both rare variants and PRSs were associated with increased risk for neurodevelopmental features in DD children (for example, duplications and deletions for feeding difficulty; $p \leq 0.028$, FDR = 0.439) but decreased risk in SPARK samples (for example, decreased missense variants in individuals with anxiety; $p = 0.016$, FDR = 0.468) (Figure 6F; Table S5J). In fact, individuals with attention-deficit/hyperactivity disorder (ADHD) had an increased autism PRS in the DD cohort ($p = 0.017$, FDR = 0.421) but a decreased autism PRS in SPARK ($p = 0.035$, FDR = 0.517) (Figure 6F; Table S5J). This trend reflects the influence of ascertainment on 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 (Figure S6G). Overall, we found marked differences in secondary variant-phenotype associations between cohorts (Figures 6D–6F), which may explain the variable phenotypic trajectories observed across ascertainments.

We finally performed a random effect meta-analysis using results from all cohorts to identify variant-phenotype associations that are robust against ascertainment bias. We identified nine associations among children in the SPARK and DD cohorts, including intelligence PRS with ADHD ($p = 0.048$, FDR = 0.856) and duplications in constrained genes with ID/DD ($p = 0.006$, FDR = 0.586) (Figure 6G; Table S5K). Across adults in the DD and UKB cohorts, intelligence PRS was associated with anxiety ($p = 0.042$, FDR = 0.719), while autism PRS was associated with sleep trouble ($p = 0.018$, FDR = 0.481) (Figure S6H; Table S5K). Using results from EHR data in UKB, MyCode, and AoU, we found consistent effects of coding SNVs (LF) on nervous system features ($p = 0.023$, FDR = 0.668) and schizophrenia PRS on mental/behavioral disorders ($p = 0.005$, FDR = 0.437) (Figure S6I; Table S5K).

Differing contributions of secondary variants by primary variant ascertainment

To extend our findings beyond the 16p12.1 deletion, we assessed the effects of secondary variants on phenotypes of 1,479 probands ascertained for autism with different rare pathogenic CNVs or SNVs in known neurodevelopmental genes (Figures 1 and S1). We first assessed 126 probands with 16p11.2 deletions ($n = 90$) and duplications ($n = 36$) in the Simon Searchlight cohort[72] and found ten variant-phenotype associations using linear regression models (Figure 7A; Table S6A). Among 16p11.2 deletion probands, schizophrenia PRS was associated with higher full-scale IQ ($\beta = 0.327$, $p = 0.031$, FDR = 0.578) whereas secondary deletions were associated with decreased IQ ($\beta = -0.288$, $p = 0.039$, FDR = 0.578), similar to previous findings[11] (Figures 7A and S7A; Table S6A). In 16p11.2 duplication probands, deletions and duplications were negatively associated with autism behavioral features (BSI; duplications: $\beta = -0.522$, $p = 0.020$, FDR = 0.561; deletions: $\beta = -0.455$, $p = 0.035$, FDR = 0.578), and duplications (LF) were negatively associated with SRS ($\beta = -0.678$, $p = 0.003$, FDR = 0.496) (Figure 7A; Table S6A). Orthogonal correlation analyses revealed several other trends, including the opposing effects of secondary duplications on BSI (16p11.2 deletion: r = 0.239, $p = 0.025$, FDR = 0.585; 16p11.2 duplication: r = −0. 402, $p = 0.020$, FDR = 0.584) (Figure S7B; Table S6B).

We next assessed 214 probands from SSC[73] with a more heterogeneous set of large CNVs, including pathogenic deletions and duplications[32] at 15q13.3, 3q29, and 16p13.11. Among probands with large deletions, linear regression models uncovered negative associations between secondary duplications (LF) and BMI ($ß = -0.275$, $p = 0.049$, FDR = 0.744), while secondary deletions (LF) were associated with decreased IQ in probands with large duplications ($ß = -0.256$, $p = 0.021$, FDR = 0.723) (Figures 7A and S7C; Table S6A). Correlation analyses revealed associations of duplicated genes with repetitive behavior in probands with primary deletions (r = 0.234, $p = 0.039$, FDR = 0.869) and STRs with decreased IQ in probands with primary duplications (r = −0.251, $p = 0.009$, FDR = 0.298) (Figure S7D; Table S6C). We further assessed 1,206 SSC probands with SNVs disrupting canonical neurodevelopmental genes,[44] such as *CHD8*, *DYRK1A*, *SCN1A*, and *PTEN*. We identified a negative association for deletions (LF) with IQ ($ß = -0.157$, $p = 4.01 \times 10^{-5}$, FDR = 0.016), while STRs (LF) were associated with increased IQ ($ß = 0.109$, $p = 0.005$, FDR = 0.254)

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

CellPress
OPEN ACCESS

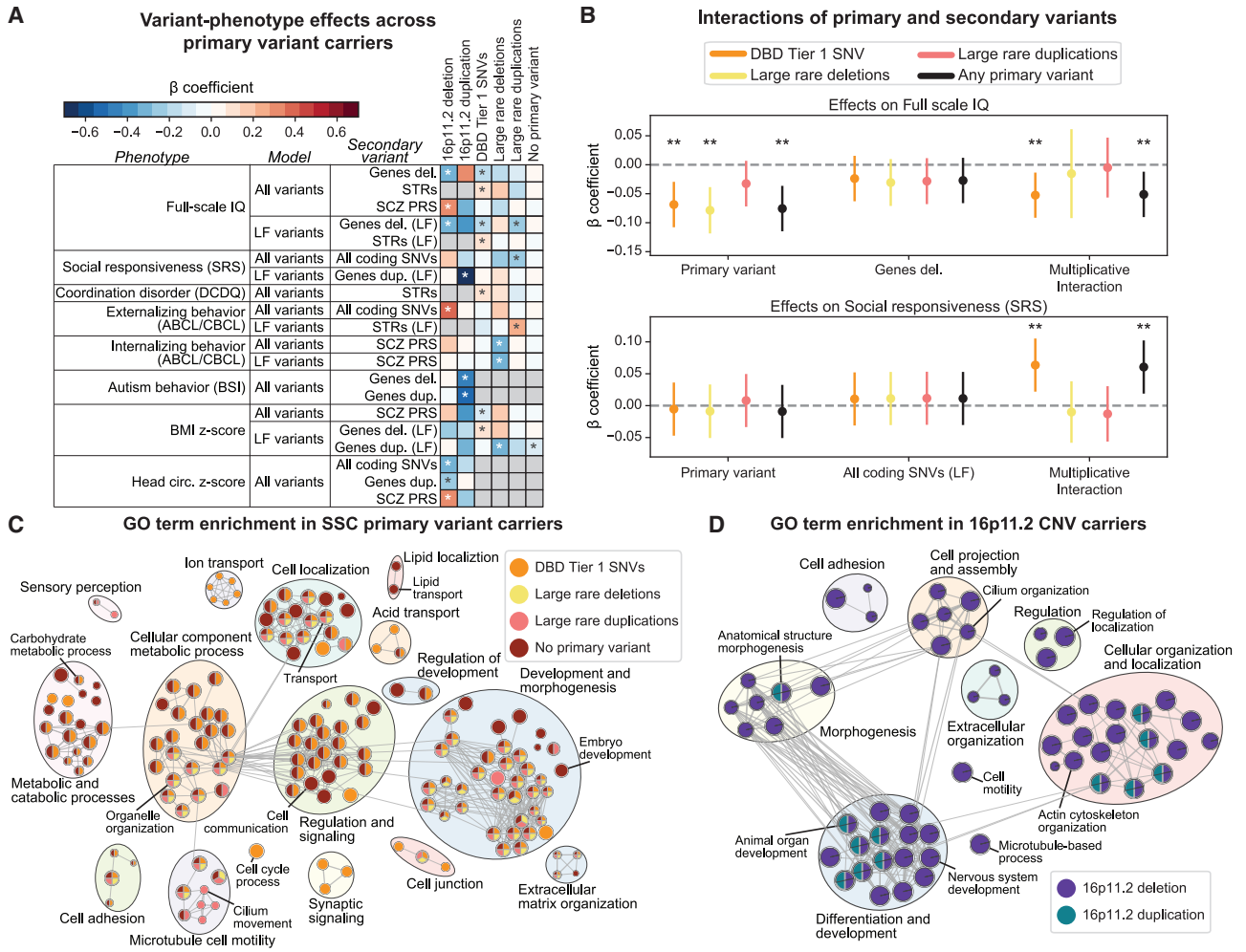


Figure 7. Secondary variant associations in probands with primary variants

(A) β coefficients from select regression models of secondary variant burden on phenotypes in SSC and Searchlight probands with different primary variants (n = 21–660). *$p \leq 0.05$. Full results in Table S6A.

(B) Regression models examining interactive effects of primary (colors) and secondary variants on phenotypes in SSC probands (n = 2,411–2,753). Full results in Table S6D. Bars, 95% CIs. **Benjamini-Hochberg FDR ≤ 0.05.

(C and D) GO biological process terms enriched among secondary variant genes in (C) SSC probands with different primary variants and (D) Searchlight probands with 16p11.2 CNVs. Node size, number of genes annotated for the term; edge width, number of genes shared between terms.

See also Figure S7.

(Figures 7A and S7C; Table S6A). Correlation analyses uncovered negative associations of externalizing behavior (r = −0.114, p = 0.001, FDR = 0.017) and repetitive behavior (r = −0.083, p = 0.017, FDR = 0.133) with educational attainment PRS (Figure S7D; Table S6C). Thus, we found some consistent patterns across primary variants, such as the negative effects of rare deletions on IQ, while other secondary variant effects were primary-variant specific.

We additionally examined secondary variants in 1,528 SSC probands without the above primary variants to assess the role of genetic background outside of a primary variant context. The only observed association from regression analysis was for duplications (LF) and lower BMI (ß = −0.086, p = 0.032, FDR = 0.723) (Figure 7A; Table S6A). The paucity of associations in the absence of primary variants suggests that secondary variant classes mostly exert their effects through interactions with primary variants rather than contributing directly to disease phenotypes. To further assess this, we used multiplicative models to identify the interactive effects of primary and secondary variants on clinical features. We found 13 instances of multiplicative effects, including primary SNVs and secondary deletions for full-scale IQ (ß = −0.052, p = 0.005, FDR = 0.033) and repetitive behavior (ß = −0.057, p = 0.002, FDR = 0.015) and primary SNVs and secondary SNVs (LF) for SRS (ß = 0.064, p = 0.002, FDR = 0.011) (Figures 7B and S7E; Table S6D). Notably, these interactions were often primary-variant specific, further supporting the hypothesis that secondary variant effects are influenced by primary variant context.

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

The relevance of primary variant context was further evident when we assessed the functions of secondary variants (Figures 7C and 7D). Secondary variants in probands with primary SNVs showed specific enrichments for synaptic signaling, while additional variants in probands with primary deletions showed enrichments for cilium movement (Figure 7C; Table S6E). Pathways enriched in secondary variants from these probands underscored the relevance of primary variant context, such as enrichment for carbon and inositol phosphate metabolism in probands with primary SNVs (Table S6E). Secondary variants in 16p11.2 deletion probands were enriched for cell motility and cell adhesion, while variants in both 16p11.2 deletion and duplication probands were enriched for organ development and anatomical structure morphogenesis (Figure 7D; Table S6F). Protein digestion and absorption pathways were specifically enriched in secondary variants from 16p11.2 deletion probands, while oxytocin signaling was enriched in secondary variants from 16p11.2 duplication probands. Several 16p11.2 genes have similar functions (i.e., cell cycle regulation[74] and oxytocin signaling pathway[75] for *MAPK3*), and multiple GO enrichments, including neuronal differentiation and projection, were shared among genes differentially expressed in animal models of 16p11.2 genes[76] (Figure 7D; Table S6F). We therefore posit that modifier variants influence developmental features by acting additively or synergistically in pathways disrupted by primary variants.[77]

DISCUSSION

Our analysis of 2,455 individuals with primary variants from diverse cohorts uncovered evidence that modifier variants confer distinct risks toward different phenotypes. These effects vary by 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 secondary variants in multiple contexts to unravel the etiology of heterogeneous clinical features in individuals with primary variants.

Several of our results expand on previous work and refine the roles of modifier variants in complex disorders. First, we identified roles for a broad range of rare variants in modulating features of the 16p12.1 deletion, including noncoding variants and STRs, expanding previous definitions of secondary variants.[27,32] Second, we expanded the role of PRSs in phenotypes of individuals with primary variants. These findings are in line with studies that identified roles for PRSs as modifiers of pathogenic CNVs, such as BMI for 16p11.2 CNVs.[10] Third, we observed 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.[27,42] This phenomenon could be attributed to assortative mating; we recently reported that 16p12.1 deletion spouses show correlations for psychiatric disorders, potentially leading to increased genetic risk over generations.[78] Contributions of secondary variants may depend on each variant's effects on gene function and molecular pathways. Inherited variants, in particular, may show differential effects due to genetic background or environmental factors. Fourth, although disease-relevant secondary variants contributed to multiple genetic diagnoses in 16p12.1 deletion probands, no variant accounted for all phenotypes in a proband, suggesting additive or interactive effects involving the deletion and secondary variants. This potentially explains the incomplete penetrance and variable expressivity often observed for disease-associated variants. Efforts such as the Genes to Mental Health (G2MH) Consortium, which aims to characterize secondary variant effects in pathogenic CNVs,[9,14] could further delineate the effects of genetic background on expressivity of primary variants.

Gene interaction networks showed that cellular mechanisms disrupted in individuals with primary variants vary by each person's unique set of secondary variants. Thus, if the primary variant affects multiple discrete pathways, as in the 16p12.1 deletion, more diverse secondary variants may interact with the primary variant. Primary variants that affect overlapping pathways, such as the 16p11.2 deletion,[74,76] may instead show additive effects of secondary variants on phenotypic outcomes.[77] This may explain the increased penetrance and decreased phenotypic variability of 16p11.2 deletion carriers compared with 16p12.1 deletion carriers. The unique mechanisms disrupted in each CNV carrier necessitate individual-level assessments to uncover the cellular etiology of associated clinical features.

Ascertainment bias can preclude a thorough assessment of primary variants, as deeper evaluations are typically restricted to individuals with specific disorders.[71,79] Although the 16p12.1 deletion contributed to clinical outcomes across cohorts, specific phenotypic trajectories and influences of secondary variants differed across ascertainments. Management of clinical effects of primary variants may therefore differ in individuals evaluated for severe developmental features versus those with other medical features.[18] Thus, a shift in treatment focus from the primary variant to all variants in an affected individual could allow for more effective management. The observed variability among variant-phenotype patterns in each cohort limits the generalizability of genetic associations within a single cohort. Meta-analyses across cohorts can be used to overcome this bias, as we found several associations across cohorts. Across primary variants, we observed general trends for PRS effects on psychiatric features, including autism PRS and ADHD in DD children, and rare variant effects on cognitive features, including deletions with full-scale IQ in 16p11.2 deletion probands. Both trends mirror variant-phenotype associations in individuals with autism outside of a primary variant context.[80] Exceptions to these patterns exist; for example, 16p11.2 deletion probands and 16p12.1 deletion carriers from SPARK show significant effects of schizophrenia PRS on cognitive features. Therefore, studies describing genotype-phenotype associations in a primary variant context should assess independent cohorts to determine how ascertainment may bias their results.

Much of the etiology for clinical features of pathogenic variants is not accounted for in our study. Several other factors could contribute, including environmental factors, inversions or complex variants, and population- or sex-specific effects. Another under-studied source of variance is non-additive interactions between variants. Only a few synergistic variant interactions have been identified in complex genomic disorders,[81] and large cohorts are required to quantify the effects of these interactions

on phenotypes.[82] Molecular studies could unravel mechanisms by which modifier variants interact with primary variants to influence their phenotypes. Although the overall effects of such interactions could account for 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.[25]

Overall, we identified family-, phenotype-, ascertainment-, and primary-variant-specific patterns of secondary variants influencing the expressivity of the 16p12.1 deletion and other primary variants. Our study emphasizes the complexity of neurodevelopmental disorders even when a causal variant is identified, suggestive of an oligogenic model for pathogenicity.[83] The complexity of the 16p12.1 deletion and other primary variants calls for personalized medicine approaches that account for individual-level phenotypic presentation and genome-wide variant profiles for counseling, management, and treatment.

Limitations of the study

This study represents one of the largest cohorts of individuals with the same pathogenic variant, but it is under-powered for identifying enrichments of individual variants or genes for specific phenotypes. Additionally, although our study captures themes regarding variable expressivity, some associations have only nominal significance and should be interpreted with caution until replicated. Larger cohorts will allow for further study of these trends and may uncover specific mechanisms and features of individuals with the same primary variant, although our results suggest extensive inter-individual variability. Additionally, some variant classes, including STRs and other structural variants, have limited accuracy when called from short-read sequencing or array data. Additional technologies, such as long-read sequencing, could more accurately assess the contributions of these variants. Finally, although we analyzed data from 976 16p12.1 deletion carriers across multiple ascertainments, differences in genotyping methods precluded burden comparisons between cohorts.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to, and will be fulfilled by, the lead contact, Santhosh Girirajan (sxg47@psu.edu).

Materials availability

The iPSC lines from consented donors can be obtained from the lead contact with a completed materials transfer agreement. All reagents are listed in the key resources table.

Data and code availability

WGS and microarray data for the DD cohort will be available from the lead contact and at NCBI dbGaP: phs002450. RNA sequencing (RNA-seq) data from 16p12.1 deletion NPCs will be available at NCBI dbGaP: phs002403. Genetic and phenotypic data from other cohorts are available from their respective biobanks (see STAR Methods). Bioinformatic pipelines and analysis code generated for this project are available at https://www.github.com/girirajanlab/16p12.1-Deletion-WGS and https://doi.org/10.5281/zenodo.16928624. All software and resources used for this project are listed in the key resources table. Statistical analyses and experimental results are available in Tables S2, S3, S4, S5, and S6.

ACKNOWLEDGMENTS

This work was supported by NIH R01-GM121907, NIH R21-NS122398, and resources from the Huck Institutes of the Life Sciences to S.G. This project received funding through the Oak Ridge Associated Universities under an agreement with the National Library of Medicine for the analysis of AoU data. M.J. and C. Smolen were supported by NIH T32-GM102057. A.T. was supported by NIH T32-LM012415. L.P. was supported by Fulbright Commission Uruguay-ANII. A. Reymond was supported by grants from the Swiss National Science Foundation 31003A_182632. S. Banka was supported by the NIHR Manchester Biomedical Research Centre (NIHR203308). We thank Craig Praul for assistance with designing the WGS strategy; Abby Hare-Harris for assistance with RedCap; Veera Rajagopal and Bertrand Isidor for comments on the manuscript; Johnathan Ray for assistance with processing transcriptome data; and Jianyu Yang, Sarah Dwiekat, and Edmundo Torres-Rodriguez for assistance with curating variant annotations. We are grateful to all of the individuals in each cohort as well as clinical sites and staff. 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, and E. Wijsman) and the Simons Searchlight Consortium. We appreciate obtaining access to genomic and phenotypic data on SFARI Base. Approved researchers can obtain Simons datasets used here 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/. We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We thank the NIH's *All of Us* Research Program for making available the data examined here.

AUTHOR CONTRIBUTIONS

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

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Human Participants
  - Cell Lines
- METHOD DETAILS
  - Phenotypic analysis
  - DNA extraction and whole-genome sequencing
  - Identification of single-nucleotide variants
  - Copy-number variants and short tandem repeats
  - Polygenic risk score calculations
  - RNA isolation and sequencing
  - Variant enrichment and pathogenicity analysis
  - Network analysis

- 16p12.1 deletion samples by ascertainment
- Samples with other neurodevelopmental disorders
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - 16p12.1 deletion DD cohort analysis
  - Ascertained 16p12.1 deletion cohort analysis
  - Multicohort meta-analysis
  - Neurodevelopmental disease cohort analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2025.09.012.

Received: August 19, 2024
Revised: July 4, 2025
Accepted: September 15, 2025

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–R187. 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. Banerjee, D., and Girirajan, S. (2023). Pathogenic Variants and Ascertainment: Neuropsychiatric Disease Risk in a Health System Cohort. Am. J. Psychiatry *180*, 11–13. https://doi.org/10.1176/appi.ajp.20220934.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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. N. Engl. J. Med. *358*, 667–675. https://doi.org/10.1056/NEJMoa075974.
23. Auwerx, C., Moix, S., Kutalik, Z., and Reymond, A. (2024). Disentangling mechanisms behind the pleiotropic effects of proximal 16p11.2 BP4-5 CNVs. Am. J. Hum. Genet. *111*, 2347–2361. https://doi.org/10.1101/2024.03.20.24304613.
24. 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.

25. 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.
26. 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.
27. 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.
28. Girirajan, S., Pizzo, L., Moeschler, J., and Rosenfeld, J. (2018). 16p12.2 Recurrent Deletion. In GeneReviews, M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J. Bean, K.W. Gripp, G.M. Mirzaa, and A. Amemiya, eds. (University of Washington).
29. 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.
30. 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.
31. 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.
32. 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.
33. 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.
34. 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.
35. Zubler, J., and Whitaker, T. (2022). CDC's Revised Developmental Milestone Checklists. Am. Fam. Physician *106*, 370–371.
36. 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.
37. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427. https://doi.org/10.1038/nature13595.
38. 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.
39. 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.
40. 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.
41. Pizzo, L., Lasser, M., Yusuff, T., Jensen, M., Ingraham, P., Huber, E., Singh, M.D., Monahan, C., Iyer, J., Desai, I., et al. (2021). Functional assessment of the "two-hit" model for neurodevelopmental defects in Drosophila and X. laevis. PLoS Genet. *17*, e1009112. https://doi.org/10.1371/journal.pgen.1009112.
42. 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.
43. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862–D868. https://doi.org/10.1093/nar/gkv1222.
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. 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.
50. 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*, e64875. https://doi.org/10.7554/eLife.64875.
51. Sun, J., Noss, S., Banerjee, D., Bhavana, V.H., Smolen, C., Das, M., Giardine, B., Prabhu, A., Amor, D.J., Pope, K., et al. (2025). An integrated framework for functional dissection of variable expressivity in genetic

disorders. Preprint at medRxiv. https://doi.org/10.1101/2025.07.22.25331885.

52. Ernst, C. (2016). Proliferation and Differentiation Deficits are a Major Convergence Point for Neurodevelopmental Disorders. Trends Neurosci. *39*, 290–299. https://doi.org/10.1016/j.tins.2016.03.001.

53. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., et al. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. *51*, D638–D646. https://doi.org/10.1093/nar/gkac1000.

54. Hu, P., Wu, S., Sun, Y., Yuan, C.-C., Kobayashi, R., Myers, M.P., and Hernandez, N. (2002). Characterization of human RNA polymerase III identifies orthologues for Saccharomyces cerevisiae RNA polymerase III subunits. Mol. Cell. Biol. *22*, 8044–8055. https://doi.org/10.1128/MCB.22.22.8044-8055.2002.

55. Duncan, A.M., Ozawa, T., Suzuki, H., and Rozen, R. (1993). Assignment of the gene for the core protein II (UQCRC2) subunit of the mitochondrial cytochrome bc1 complex to human chromosome 16p12. Genomics *18*, 455–456. https://doi.org/10.1006/geno.1993.1500.

56. Hérault, J., Petit, E., Martineau, J., Perrot, A., Lenoir, P., Cherpi, C., Barthélémy, C., Sauvage, D., Mallet, J., and Müh, J.P. (1995). Autism and genetics: clinical approach and association study with two markers of HRAS gene. Am. J. Med. Genet. *60*, 276–281. https://doi.org/10.1002/ajmg.1320600404.

57. Comings, D.E., Wu, S., Chiu, C., Muhleman, D., and Sverd, J. (1996). Studies of the c-Harvey-Ras gene in psychiatric disorders. Psychiatry Res. *63*, 25–32. https://doi.org/10.1016/0165-1781(96)02829-6.

58. Yamamoto, T., Harada, N., Kano, K., Taya, S., Canaani, E., Matsuura, Y., Mizoguchi, A., Ide, C., and Kaibuchi, K. (1997). The Ras target AF-6 interacts with ZO-1 and serves as a peripheral component of tight junctions in epithelial cells. J. Cell Biol. *139*, 785–795. https://doi.org/10.1083/jcb.139.3.785.

59. Wang, L., Zhou, K., Fu, Z., Yu, D., Huang, H., Zang, X., and Mo, X. (2017). Brain Development and Akt Signaling: the Crossroads of Signaling Pathway and Neurodevelopmental Diseases. J. Mol. Neurosci. *61*, 379–384. https://doi.org/10.1007/s12031-016-0872-y.

60. De Vries, B.B.A., White, S.M., Knight, S.J., Regan, R., Homfray, T., Young, I.D., Super, M., McKeown, C., Splitt, M., Quarrell, O.W., 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.

61. Pusapati, G.V., Kong, J.H., Patel, B.B., Krishnan, A., Sagner, A., Kinnebrew, M., Briscoe, J., Aravind, L., and Rohatgi, R. (2018). CRISPR Screens Uncover Genes that Regulate Target Cell Sensitivity to the Morphogen Sonic Hedgehog. Dev. Cell *44*, 113–129.e8. https://doi.org/10.1016/j.devcel.2017.12.003.

62. 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.

63. 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.

64. SPARK Consortium. Electronic address: pfeliciano@simonsfoundation.org, and SPARK Consortium. (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.

65. 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.

66. 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.

67. 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.

68. All of US Research Program Genomics Investigators (2024). Genomic data in the All of Us Research Program. Nature *627*, 340–346. https://doi.org/10.1038/s41586-023-06957-x.

69. Owen, D., Bracher-Smith, M., Kendall, K.M., Rees, E., Einon, M., Escott-Price, V., Owen, M.J., O'Donovan, M.C., and Kirov, G. (2018). Effects of pathogenic CNVs on physical traits in participants of the UK Biobank. BMC Genomics *19*, 867. https://doi.org/10.1186/s12864-018-5292-7.

70. Zeng, C., Schlueter, D.J., Tran, T.C., Babbar, A., Cassini, T., Bastarache, L.A., and Denny, J.C. (2024). Comparison of phenomic profiles in the All of Us Research Program against the US general population and the UK Biobank. J. Am. Med. Inform. Assoc. *31*, 846–854. https://doi.org/10.1093/jamia/ocad260.

71. 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.

72. Simons; Vip Consortium (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.

73. 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.

74. 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.

75. Jurek, B., and Neumann, I.D. (2018). The Oxytocin Receptor: From Intracellular Signaling to Behavior. Physiol. Rev. *98*, 1805–1908. https://doi.org/10.1152/physrev.00031.2017.

76. 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.

77. 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.

78. Smolen, C., Jensen, M., Dyer, L., Pizzo, L., Tyryshkina, A., Banerjee, D., Rohan, L., Huber, E., El Khattabi, L.E., Prontera, P., et al. (2023). Assortative mating and parental genetic relatedness drive the pathogenicity of variably expressive variants. Preprint at medRxiv. https://doi.org/10.1101/2023.05.18.23290169.

79. 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.

80. 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.

81. 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.

82. 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.

83. 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.

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. 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*, 317–330. https://doi.org/10.1038/nature14248.

87. Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Hum. Mutat. *32*, 894–899. https://doi.org/10.1002/humu.21517.

88. 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.

89. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

90. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

91. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van Der Auwera, G.A., 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. Preprint at bioRxiv. https://doi.org/10.1101/201178.

92. 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.

93. 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.

94. 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.

95. 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.

96. 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.

97. Pedersen, B.S., Layer, R.M., and Quinlan, A.R. (2020). smoove: structural-variant calling and genotyping with existing tools. Version 0.2.8. GitHub. https://github.com/brentp/smoove.

98. 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.

99. 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*, i333–i339. https://doi.org/10.1093/bioinformatics/bts378.

100. 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.

101. 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.

102. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

103. 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.

104. 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.

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

106. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527. https://doi.org/10.1038/nbt.3519.

107. Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., and Peterson, H. (2023). g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. *51*, W207–W212. https://doi.org/10.1093/nar/gkad347.

108. Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. In Proceedings of the 7th Python in Science Conference, G. Varoquaux, T. Vaught, and J. Millman, eds., pp. 11–15. https://doi.org/10.25080/TCWV9851.

109. 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.

110. 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.

111. 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.

112. 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.e12. https://doi.org/10.1016/j.cell.2017.08.047.

113. 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.

114. Bozaoglu, K., Gao, Y., Stanley, E., Fanjul-Fernández, M., Brown, N.J., Pope, K., Green, C.C., Vlahos, K., Sourris, K., Bahlo, M., et al. (2019). Generation of seven iPSC lines from peripheral blood mononuclear cells suitable to investigate Autism Spectrum Disorder. Stem Cell Res. *39*, 101516. https://doi.org/10.1016/j.scr.2019.101516.

115. Deshpande, A., Yadav, S., Dao, D.Q., Wu, Z.-Y., Hokanson, K.C., Cahill, M.K., Wiita, A.P., Jan, Y.-N., Ullian, E.M., and Weiss, L.A. (2017). Cellular Phenotypes in Human iPSC-Derived Neurons from a Genetic Model of Autism Spectrum Disorder. Cell Rep. *21*, 2678–2687. https://doi.org/10.1016/j.celrep.2017.11.037.

116. 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 *246*, 1–190.

117. 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.

118. 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.

119. Pedersen, B.S., Bhetariya, P.J., Brown, J.M., Kravitz, S.N., Marth, G., Jensen, R.L., Bronner, M.P., Underhill, H.R., and Quinlan, A.R. (2020). Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. Genome Med. *12*, 62. https://doi.org/10.1186/s13073-020-00761-2.

120. 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.

121. 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.

122. 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.

123. 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.

124. International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58. https://doi.org/10.1038/nature09298.

125. Zhang, Y., Parmigiani, G., and Johnson, W.E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom. Bioinform. *2*, lqaa078. https://doi.org/10.1093/nargab/lqaa078.

126. 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.

127. 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.

128. Wu, Y., Yao, Y.-G., and Luo, X.-J. (2017). SZDB: A Database for Schizophrenia Genetic Research. Schizophr. Bull. *43*, 459–471. https://doi.org/10.1093/schbul/sbw102.

129. 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.

130. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J. M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.

131. Gene Ontology Consortium, Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., et al. (2023). The Gene Ontology knowledgebase in 2023. Genetics *224*, iyad031. https://doi.org/10.1093/genetics/iyad031.

132. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

133. Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G.D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One *5*, e13984. https://doi.org/10.1371/journal.pone.0013984.

134. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

135. Kucera, M., Isserlin, R., Arkhangorodsky, A., and Bader, G. (2016). AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations. F1000Res *5*, 1717.

136. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559. https://doi.org/10.1186/1471-2105-9-559.

137. Langfelder, P., and Horvath, S. (2012). Fast R Functions for Robust Correlations and Hierarchical Clustering. J. Stat. Softw. *46*, i11. https://doi.org/10.18637/jss.v046.i11.

138. Observational Health Data Sciences and Informatics (OHDSI) (2024). OHDSI Vocabulary Search Tool (Observational Health Data Sciences and Informatics).

139. 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.

140. 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.

141. Banerjee, D., and Girirajan, S. (2025). Discovery of novel obesity genes through cross-ancestry analysis. Preprint at medRxiv. https://doi.org/10.1101/2024.10.13.24315422.

142. 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.

143. 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.

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
QIAamp DNA Blood Maxi	Qiagen	Cat. #51104
TruSeq DNA PCR-free Library Prep Kit	Illumina	Cat. #20015962
OmniExpress 24 v.1.1	Illumina	Cat. #20062061
Cytotune-iPS 2.0 Sendai Reprograming kit	Invitrogen	Cat. #A16517
mTESR1 medium	Stem Cell Technologies	Cat. #85850
mTESR plus medium	Stem Cell Technologies	Cat. #100-0276
Penicillin/Streptomycin	Sigma-Aldrich	Cat. #P4333
Geltrex	Gibco	Cat. #A1413302
Rock inhibitor Y-27632	Stem Cell Technologies	Cat. #72304
Collagenase Type IV	Stem Cell Technologies	Cat. #07909
DMEM/F-12 medium	Gibco	Cat. #10565018
N-2 Supplement	Gibco	Cat. #17502001
MEM Nonessential amino acids	Gibco	Cat. #11140050
Heparin	Stem Cell Technologies	Cat. #07980
SB431542	Stem Cell Technologies	Cat. #72232
LDN193189	Stem Cell Technologies	Cat. #72147
Neurobasal medium	Gibco	Cat. #21103049
GlutaMax supplement	Gibco	Cat. #35050061
B-27 supplement, minus vitamin A	Gibco	Cat. #12587010
Accutase	Gibco	Cat. #A1110501
STEMdiff neural progenitor medium	Stem Cell Technologies	Cat. #05833
TRIzol	Invitrogen	Cat. #15596026
PureLink RNA Mini Kit	Invitrogen	Cat. #12183018A
TURBO DNase	ThermoFisher Scientific	Cat. #AM1907
NEBNext Ultra II RNA Library Prep Kit	New England Biolabs (NEB)	Cat. #E7770S
Deposited data		
Raw WGS/microarray data (DD Cohort)	This paper	dbGaP: phs002450
Raw NPC RNA-seq data	This paper	dbGaP: phs002403
Exome sequencing data (UK Biobank)	UK Biobank	RRID:SCR_012815
CNV data (UK Biobank)	UK Biobank	Data Fields 22437, 22431; RRID: SCR_012815
BMI polygenic risk score	UK Biobank	Data Field 26216; RRID:SCR_012815
Phenotypic data (UK Biobank)	UK Biobank	Data Fields 41270, 20446, 20441, 1200, 1920, 20241, 20425, 20549, 20418, 20407, 20413, 20456, 20463, 20471, 20468; RRID: SCR_012815
Sequencing/phenotype data (SPARK)	Simons Foundation	https://base.sfari.org
Sequencing/phenotype data (Searchlight)	Simons Foundation	https://base.sfari.org
Sequencing/phenotype data (SSC)	Simons Foundation	https://base.sfari.org; RRID:SCR_004644
GnomAD (allele freq., LOEUF scores)	Lek et al.[84]	https://gnomad.broadinstitute.org; RRID: SCR_014964
GENCODE v19 (gene definitions)	Frankish et al.[85]	https://www.gencodegenes.org/human/release_19.html; RRID: SCR_014966

(*Continued on next page*)

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ClinVar (gene pathogenicity)	Landrum et al.[43]	https://www.ncbi.nlm.nih.gov/clinvar/; RRID:SCR_006169
SFARI Gene (NDD gene annotations)	Abrahams et al.[45]	https://gene.sfari.org; RRID:SCR_001872
DBD Genes (NDD gene annotations)	Gonzalez Mantilla et al.[44]	https://dbd.geisingeradmi.org
Roadmap Epigenomics (chromatin state)	Roadmap Epigenomics Consrtium[86]	https://egg2.wustl.edu/roadmap/web_portal/; RRID:SCR_008924
Psychiatric trait GWAS summary statistics	Schizophrenia Working Group of the Psychiatric Genomics Consortium[37], Savage et al.[38], Lee et al.,[39] Grove et al.[40]	https://pgc.unc.edu/for-researchers/download-results/; RRID: SCR_004495
BrainSpan Atlas (Brain tissue expression)	Miller et al.[49]	https://www.brainspan.org/static/download.html; RRID: SCR_008083
Motor cortex single-cell expression	Bakken et al.[50]	https://portal.brain-map.org/atlases-and-data/rnaseq/comparative-lgn
STRING protein interaction network	Szlarczyk et al.[53]	https://string-db.org/; RRID:SCR_005223
dbNSFP v.4	Liu et al.[87,88]	https://www.dbnsfp.org; RRID: SCR_005178
Experimental models: Cell lines		
Human: Patient-derived iPSCs and NPCs	Sun et al.[51]	N/A
Software and algorithms		
Bioinformatic pipelines and analysis code	This paper	https://doi.org/10.5281/zenodo.16928624
BWA v.0.7.13	Li and Durbin[89]	https://github.com/lh3/bwa; RRID: SCR_010910
Samtools v.1.9	Li et al.[90]	https://github.com/samtools/samtools; RRID:SCR_002105
GATK v.3.8	Poplin et al.[91]	https://gatk.broadinstitute.org; RRID: SCR_001876
ANNOVAR	Wang et al.[92]	https://www.openbioinformatics.org/annovar/annovar_download_form.php; RRID:SCR_012821
Vcfanno	Pedersen et al.[93]	https://github.com/brentp/vcfanno; RRID: SCR_024372
PennCNV	Wang et al.[94]	https://penncnv.openbioinformatics.org/; RRID:SCR_002518
CNVNator v.0.4.1	Abyzov et al.[95]	https://github.com/abyzovlab/CNVnator; RRID:SCR_010821
Lumpy-sv v.0.2.13	Layer et al.[96], Pedersen et al.[97]	https://github.com/arq5x/lumpy-sv; RRID: SCR_003253
Manta v.1.6.0	Chen et al.[98]	https://github.com/Illumina/manta; RRID: SCR_022997
Delly v.1	Rausch et al.[99]	https://github.com/dellytools/delly; RRID: SCR_004603
GangSTR v.2.5	Mousavi et al.[100]	https://github.com/gymreklab/GangSTR
TRTools	Mousavi et al.[101]	https://trtools.readthedocs.io/
PLINK	Purcell et al.[102]	https://www.cog-genomics.org/plink/; RRID:SCR_001757
TopMed Imputation Server v.r2	Das et al.[103]	https://imputation.biodatacatalyst.nhlbi.nih.gov; RRID:SCR_015677
Peddy v.0.4.8	Pedersen and Quinlan[104]	https://github.com/brentp/peddy; RRID: SCR_017287
LDPred-2	Privé et al.[105]	https://privefl.github.io/bigsnpr/

(*Continued on next page*)

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Kallisto v0.50.0	Bray et al.[106]	https://github.com/pachterlab/kallisto; RRID:SCR_016582
gProfiler	Kolberg et al.[107]	https://biit.cs.ut.ee/gprofiler/gost; RRID: SCR_006809
NetworkX	Hagberg et al.[108]	https://networkx.org/; RRID:SCR_016864
Variant Effect Predictor	McLaren et al.[109]	https://github.com/Ensembl/ensembl-vep; RRID:SCR_007931
PheWAS	Denny et al.[110]	https://github.com/PheWAS/PheWAS; RRID:SCR_003512
Other		
HRS-MAT (non-verbal IQ assessment)	Hansen[33]	https://www.hrs-mat.com
SRS (autism-related behavior assessment)	Constantino et al.[34]	https://www.wpspublish.com/srs-2-social-responsiveness-scale-second-edition

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human Participants

We analyzed genomic and phenotypic data from a cohort of 442 individuals belonging to 124 families with the 16p12.1 deletion (hg18/NCBI36 when originally reported; currently maps to 16p12.2 in hg19/GRCh37), which we refer to as the Developmental Delay cohort ("DD cohort") (Figure 1). These families include single probands (n=14), parent-child pairs (n=13), complete trios (n=97), and extended families, including eight families with three generations (Figure S3) and 22 families with multiple affected children. Individuals ranged from 0.5 to 75 years of age and included 237 males and 205 females; a list of all individuals in the DD cohort, including demographic information, familial relationships, and 16p12.1 deletion status, is available in Table S1A. The deletion was identified through prior clinical diagnostic tests for ID/DD or other developmental disorders in probands. Many, but not all, probands in the DD cohort have ID/DD phenotypes, while many of the family members also manifest neurodevelopmental or psychiatric features, likely due to proband ascertainment. However, presentation of clinical features was not a condition for ascertainment in this cohort. 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.[71] 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). Power calculations for detecting changes in rare variant burden within deletion family members (Figure S2D) 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.[27,111,112]

In addition to the DD cohort, we assessed genotype-phenotype associations from individuals carrying 16p12.1 deletions derived from four additional cohorts, each representing a distinct ascertainment. Individuals in these cohorts were identified from analysis of microarray data. Individuals in the Simons Powering Autism Research for Knowledge (SPARK) cohort (n=56) were ascertained for families with autism,[64] the Geisinger MyCode Community Health Initiative (MyCode) (n=160) represents a health care-based cohort,[67] the UK Biobank[65] (UKB) (n=249) consists of individuals with a "healthy volunteer" bias,[66] and individuals from All of Us[68] (AoU) (n=258) represent a diverse group of adults from the United States. SPARK, MyCode, and UKB are composed of primarily European individuals, while AoU represents a more diverse cohort. Combined with samples from the DD cohort (after excluding samples with incomplete phenotypic information), we assessed a total of 976 deletion carriers. 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). Individuals from the MyCode cohort were recruited during primary care or specialty clinic visits to Geisinger Health System locations, independent of condition, diagnosis, or demographic characteristic. Written informed consent was obtained from adult patients and from the parents or guardians of pediatric patients. The study was conducted with approval from the Geisinger institutional review board. Data from the UK Biobank was accessed under application 45023. AoU data were accessed from the All of Us Researcher Workbench.

We further assessed genomic and phenotypic data from individuals with other disease-associated primary variants. Specifically, we assessed probands ascertained for 16p11.2 deletions and duplications from the Simons Searchlight project,[72] and probands ascertained for simplex cases of autism in the Simons Simplex Collection (SSC).[73] Probands from these cohorts are primarily European.

Within the Simons Searchlight cohort, we assessed 159 probands with the 16p11.2 duplication (n=52) or deletion (n=107). Within the SSC cohort, we assessed genomic data of 2,848 total probands, and classified probands with the following primary variant classes for downstream analysis: (i) 1,206 probands with rare, deleterious variants (<0.1% gnomAD and ancestry-specific gnomAD frequency,[84] loss-of-function or missense variants with CADD[113] Phred-like scores >25) in Tier 1 genes from the Developmental Brain Disorder Gene Database, which represent genes with well-documented connections to neurodevelopmental disease[44]; (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) 1,528 probands who did not carry any of these variants or any other known pathogenic CNVs.[32] We note that groups with primary variants have overlapping samples, such that a total of 1,320 SSC probands with primary variants were assessed. Additionally, we assessed phenotypic data for an additional 28 SSC probands to compare SRS distributions with 16p12.1 deletion probands. A total of 2,876 total SSC probands and 159 Simons Searchlight probands were used in our 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 summary, we assessed secondary variants and phenotypes in 2,455 individuals with primary variants from seven cohorts: DD (n=245), Estonian Biobank (n=8), SPARK (n=56), MyCode (n=160), UKB (n=249), AoU (n=258), Searchlight (n=159), and SSC (n=1,320). We also assessed data from 1,528 SSC probands without primary variants (Figures 1 and S1). All individuals were identified using a genetics-first approach, defined by the presence or absence of a primary variant. Data from an additional 406,276 control individuals was also included, including noncarrier samples from DD (n=95), age and sex-matched controls without CNVs from AoU (n=58,560), PheWAS controls and controls without large CNVs from UKB (n=347,593), and SSC probands without genetic data but with SRS data (n=28) (Figure S1).

Cell Lines

Induced pluripotent stem cell (iPSC) generation was performed on 12 individuals in three 16p12.1 deletion families[51] using a standard iPSC generation protocol.[114] Briefly, peripheral blood mononuclear cells (PBMCs) were first isolated from peripheral blood samples, and reprogrammed into iPSCs using the Cytotune-iPS 2.0 Sendai Reprograming kit (Invitrogen). The reprogrammed iPSCs were validated by immunofluorescence (NANOG, OCT4, SOX2 and SSEA4) and flow cytometry (EPCAM, TRA-1-81, SSEA4 and CD9), and grown in mTESR1 and mTESR plus medium (Stem Cell Technologies) supplemented with 1% Penicillin/Streptomycin (Sigma-Aldrich) on Geltrex (Gibco)-coated dishes. The iPSCs were then passaged using 0.5mM EDTA with Rock inhibitor Y-27632 (Stem Cell Technologies) to improve cell survival. All cells were incubated at 37°C and 5% CO_2. One clone from each individual was used to generate three replicates for differentiation and analysis.

Neural progenitor cells (NPCs) represent a cell stage that may be a convergent point for neurodevelopmental disorders[52] and are particularly relevant for the 16p12.1 deletion based on previous studies showing a role for 16p12.1 gene homologs in proliferation and apoptosis.[41] Because of this, we differentiated iPSCs into NPCs using a standardized protocol with modification.[115] Briefly, on day 0, iPSCs were treated with Collagenase Type IV (Stem Cell Technologies). Cells were then scraped and transferred to a 60mm dish with mTESR1 media with 10μM Y-27632 for embryoid body (EB) formation. After suspension culture for 48 hours, cells were supplied with neural differentiation media (N2 media) containing DMEM/F-12 (Gibco) with 1% N-2 Supplement (Gibco) and 1% MEM Nonessential amino acids (Gibco), 2μg/ml Heparin (Stem Cell Technologies), 1% Penicillin/Streptomycin (Sigma-Aldritch), 5μM SB431542 (Stem Cell Technologies) and 0.25μM LDN193189 (Stem Cell Technologies). EBs were cultured with N2 media for 4 days, with media changes every two days. On day 6, EBs were seeded on Geltrex-coated plates and incubated for rosette formation. Rosettes were cultured in N2 media during days 6 to 8, and in N2B27 media containing 1:1 DMEM/F-12 and Neurobasal Medium (Gibco) with GlutaMax supplement (Gibco); 1% N-2 Supplement; 2% B-27 supplement, minus vitamin A (Gibco); 1% MEM Nonessential amino acids; 2μg/ml Heparin; and 1% Penicillin/Streptomycin during days 8 to 14. On day 14, rosettes were treated with Collagenase Type IV and transferred to an uncoated 60mm dish with N2B27 media for neurosphere (NS) formation in suspension culture for 12 days. During NS formation, media was changed every two days. On day 26, NSs were dissociated using Accutase (Gibco) and then plated on Geltrex-coated plates. NPCs were maintained in STEMdiff neural progenitor medium (Stem Cell Technologies).

METHOD DETAILS

Phenotypic analysis

Individual-level phenotypic data described below are available in Table S1A.

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 complete 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, and also measured specific quantitative features (see "Assessment of quantitative phenotypes" below).

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

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 (Figure 2A; Tables S1A–S1C). 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 did not add 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.

Assessment of quantitative phenotypes

We performed online quantitative assessments using the Hansen Research Services Matrix Adaptive Test (HRS-MAT) for non-verbal IQ[33] and Social Responsiveness Scale (SRS) for autism-related social behavior[34] (Figure 2C). HRS-MAT was self-administered to participants through an online platform, while 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.[116,117] We further obtained SRS, BMI, and head circumference z-scores for probands in the SSC and Simons Searchlight cohorts (see below), while the mean HRS-MAT score in SSC probands was obtained from Hansen.[33] 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). We note that these tests did not undergo multiple testing correction due to the small number of tests (nine) in this analysis.

Developmental milestones

We assessed the achievement of developmental milestones in children from the DD cohort based on CDC guidelines.[35] 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).

Additional phenotypes

We additionally assessed adults and children in the DD cohort for the presence of several clinical phenotypes. For adults, presence of six phenotypes (depression, anxiety, sleep trouble, psychosis, addiction, and mood lability) was assessed by interpreting responses to mental health questionnaires. A complete list of the questions and positive responses is available in Table S5A. For children, phenotypes were assessed by clinician-, guardian-reported presence of the following clinical features: birth/pregnancy complications, preterm birth, microcephaly, macrocephaly, strabismus, depression, anxiety, sleep trouble, seizures, heart defects, hearing loss, vision problems, feeding problems, obesity, ID/DD, motor delay, speech delay, language disorder, aide in school, learning disability, ASD, ADHD, OCD, schizophrenia, bipolar disorder, and pervasive developmental delay. All phenotypes for individuals in the DD cohort are available in Table S1A.

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 submitted 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 35.7× coverage, or 716.2 M reads/sample, with 94.9% of reads mapping to the human genome. After processing for quality control using Trimmomatic[118] (leading:5, trailing:5, and slidingwindow:4:20 parameters), sequences were aligned to the hg19 reference genome using BWA v.0.7.13,[89] and sorted and indexed using Samtools v.1.9.[90] We note that sequencing data from 163 individuals was described previously.[42,78]

We used SNP microarrays for copy-number variant validation and genotyping experiments (i.e. CNV calling and polygenic risk score calculation) in 368 individuals. Samples were run on Illumina OmniExpress 24 v.1.1 microarrays by the Northwest Genomics Center at the University of Washington (Seattle, WA, USA). We note that microarray data of 208 individuals in this study was described previously.[27,42,78]

Identification of single-nucleotide variants

We identified SNVs and small indels using the GATK Best Practices pipeline,[91] followed by quality control and extensive variant and gene-level annotations. Duplicate read 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[93] to annotate variants with gnomAD frequency.[84] All variant 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. Calls were then additionally annotated for rarity (frequency ≤0.1%) using ancestry-specific frequencies from gnomAD. Genetic ancestry for each sample was calculated using somalier[119] and combined with self-reported ancestry to identify calculated or self-reported European, Ashkenazi Jewish, East Asian, Admixed American, or African ancestry. Variants for each sample were then filtered for those with ancestry-specific gnomAD frequency ≤0.1% for any calculated or self-reported ancestry.

We annotated coding and noncoding variants within genes from GENCODE[85] v19 using ANNOVAR[92] and Vcfanno[93] 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,[113] presence in ClinVar database,[43] and the gene-level pathogenicity metric LOEUF (v.2.1.1).[36] 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[86] (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 in the DD cohort are listed in Table S1A and average burden for the whole cohort is listed in Table S1D. We determined inheritance of variants in probands by checking if the same variant was called in either parent. If it was not called in either parent, we checked the unfiltered calls from GATK. If at least 20% of the reads in either parent support the alternate allele, the variant was determined to be inherited from that parent.

Copy-number variants and short tandem repeats

CNVs were called from both microarray data using PennCNV[94] and WGS data using CNVnator v.0.4.1,[95] Lumpy-sv v.0.2.13[96] with Smoove v.0.2.5,[97] Delly v.1,[99] and Manta v.1.6.0.[98] 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.[32] 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.[120] Known pathogenic CNVs[32] were excluded from this filter. All CNVs were then filtered for GnomAD-SV frequency[121] <0.1% and intracohort frequency ≤10, and >50 kb CNVs were additionally filtered for <0.1% frequency in a control cohort.[122] All CNVs were finally filtered for those intersecting at least one protein-coding gene, using gene annotations from GENCODE v19.[85] For large CNVs, including the 16p12.1 deletion, inheritance was determined directly from parental genetic data without the need for haplotype phasing in probands. Proband CNVs were determined to be inherited if there was a CNV with 50% reciprocal overlap present in either parent.

We identified STR expansions from WGS data using TRTools pipelines[101] with GangSTR[100] (reference file v.13.1). 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. We defined STR expansions as STR variants with lengths >2 SD higher than the average of STR lengths among all individuals in the DD cohort 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 with LOEUF scores.[36] STRs were determined to be inherited if only one parent has the same STR call in the final filtered call set. The number of genes affected by rare CNVs and the number of STR expansions for all individuals with available WGS data are listed in Table S1A and the average burden for all DD cohort samples is listed in Table S1D.

Polygenic risk score calculations

Using microarray data, we calculated polygenic risk scores for educational attainment,[39] intelligence,[38] schizophrenia,[37] and autism[40] among the samples in the DD cohort, based on standardized bioinformatics pipelines for quality control.[123] We first downloaded summary statistics from four recent GWAS studies of neuropsychiatric traits, and filtered SNPs for imputation INFO scores >0.8 and removed duplicate and ambiguous SNPs. We then merged SNP genotype data from different microarray batches together using PLINK.[102] Initial quality control removed SNPs with minor allele frequency <0.05, Hardy-Weinberg equilibrium $<1.0\times10^{-6}$, and genotype rate <0.01, along with samples missing >1% of genotypes. We used the HRC-1000 Genomes Imputation toolkit (https://www.well.ox.ac.uk/~wrayner/tools) to process PLINK files into individual chromosomes for imputation, and VcfCooker (https://genome.sph.umich.edu/wiki/VcfCooker) to convert PLINK files to VCF files. Microarray-based SNPs were imputed using the TOPMed v.r2 imputation server using Eagle v2.4 for phasing.[103] After imputation, VCF files were converted back to PLINK format, and SNPs were again filtered with identical QC filters. Additional QC filters included removing samples with >±3SD of the mean heterozygosity rate and removing individuals with non-European ancestry, based on imputed genetic ancestry (calculated using Peddy v.0.4.8[104] with 1000 Genomes-based population panel) or self-reported ancestry. To calculate PRS, we used standardized pipelines

for the LDPred2 software package, which uses Bayesian approaches to optimize parameters for PRS calculation.[105] Briefly, we filtered the four sets of GWAS summary statistics for SNPs present in the HapMap3 dataset,[124] and used 1000 Genomes datasets to calculate linkage disequilibrium matrices for the SNPs. After regressing betas or odds ratios of GWAS SNPs according to linkage disequilibrium, we used the LDPred2-auto model to calculate the four PRS values for all samples with available genotype data. PRS for all available individuals are listed in Table S1A.

RNA isolation and sequencing

Total RNA was isolated from NPCs[51] (passage 5) using TRIzol (Invitrogen) and the PureLink RNA Mini Kit (Invitrogen). Isolated RNA was treated with TURBO DNase (ThermoFisher Scientific). RNA quality was assessed using the Agilent TapeStation 4200 (Agilent Technologies) and samples with an RNA integrity number score (RIN) $\geq$6.0 and purity (A260/280)>1.8 were processed for sequencing. RNA sequencing libraries were prepared by Genewiz from Azenta Life Sciences (South Plainfield, NJ) using the NEBNext Ultra II RNA Library Prep Kit (NEB) for 150bp paired-end sequencing using the Illumina NovaSeq 6000 platform for an average of 34.7 million reads per sample. Trimmomatic v0.39[118] (leading:3, trailing:3, slidingwindow:4:15, and minlen:36 parameters) was used to trim adapters and to remove low-quality reads. Transcript abundance was quantified using kallisto v0.50.0[106] with n=100 bootstrap samples (index built with GRCh38.p14 cDNA). Batch effects were corrected using ComBat-seq[125] within the R package *sva* v.3.35.2 in R v.4.3.1.

Variant enrichment and pathogenicity analysis

For all enrichments, Benjamini-Hochberg multiple testing correction was performed using the scipy v.1.13.1 *false_discovery_control* function. 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[36]; referred to as "(LF)"). Sample sizes, test statistics, p-values, and FDR values for enrichments are listed in Tables S2, S3, and S4.

Gene set enrichment

We assessed enrichment of genes with secondary variants among sets of neurodevelopmental disease genes and genes with neuronal function from several previously published resources.[36,44,45,48,126–129] We identified enrichment of variants in these gene sets by performing Fisher's Exact tests against the whole genome for each gene list and calculated odds ratios and p-values for genes with variants in the DD cohort using the *contingency.odds_ratio* function from scipy v.1.13.1 (Table S2F).

Gene ontology and pathway enrichment

Gene ontology (GO) and biological pathway enrichment was performed using gProfiler[107] Python API with g:SCS multiple testing correction method applying significance threshold of 0.05. Enrichment was assessed for the GO:BP[130,131] and KEGG[132] annotation datasets (Tables S3C, S4D, S6E, and S6F). The EnrichmentMap[133] plugin for Cytoscape[134] was used to plot enriched GO terms across different conditions. The AutoAnnotate plugin[135] was used for initial clustering of related GO terms, followed by manual arrangement to create groups of related terms.

Spatio-temporal brain expression

We assessed variant enrichment in genes preferentially expressed in specific brain tissues using the BrainSpan Atlas[49] and in genes preferentially expressed in specific cell types using single-cell RNA-seq expression data in the M1 motor cortex.[50] We previously defined preferentially expressed genes as those with expression >2SD higher than the median expression across all tissues or all cell types for that gene.[42] We used Fisher's exact tests as described above to find the odds that a gene both carries a variant in the DD cohort and is expressed in a specific brain region or cell type (Tables S2G and S2H).

16p12.1 co-expressed genes

We identified gene co-expression modules using weighted gene correlation network analysis (WGCNA) in induced pluripotent stem cell-derived neural progenitor cells from 12 individuals in three families with the 16p12.1 deletion using the *WGCNA* package v1.73[136,137] in R v4.2.3. After importing gene expression counts, we log-transformed the counts and selected the 20,000 most variable genes for downstream analysis. We used the *pickSoftThreshold* function to identify the optimal power and the *blockwiseModules* function to build a "signed hybrid" network, identifying 20 co-expression modules in the network. For our secondary variant enrichment analysis, we focused on five modules that contained the 16p12.1 genes. Secondary variants were identified from a "genetics-first" approach, where gene expression profiles were not considered during variant calling. Thus, they are unlikely to be biased towards genes co-expressed with 16p12.1 genes. We used Fisher's exact tests as described above to identify enrichments of secondary variants in the 16p12.1 deletion co-expression modules (Table S2I).

Pathogenic variant analysis

We defined pathogenic SNVs as those that are "Pathogenic" or "Likely pathogenic" from multiple submitters in ClinVar,[43] not annotated with autosomal recessive inheritance, and associated with neurodevelopmental phenotypes, or loss-of-function variants in genes that (a) are a Tier S SFARI gene, which represent strong candidate autism genes,[45] or (b) are in the Tier 1 gene list from the Developmental Brain Disorder Gene Database[44] (Table S1E). Complete lists of genes and variants meeting these criteria can be found in their respective databases. Pathogenic CNVs were identified from 50% reciprocal overlap with a previously published list of CNVs[32] and probands with pathogenic CNVs are listed in Table S1F.

Network analysis

We assessed the connectivity of primary and secondary variant genes within the STRING DB protein interaction network.[53] After mapping proteins present in the network to ENSEMBL gene IDs, we re-calculated interaction scores using nine of the thirteen available lines of evidence (fusion, co-occurrence, homology, coexpression, coexpression-transferred, experiments, experiments-transferred, database, and database-transferred). For all downstream analyses, we restricted the network to only the highest confidence gene interactions (minimum interaction score$\geq$0.9). We used NetworkX[108] to identify the shortest paths between 16p12.1 and 16p11.2 genes (''primary variant genes'') and secondary variant genes in 16p12.1 deletion and 16p11.2 deletion probands, respectively. Paths were weighted as 1/interaction score, such that gene pairs with more confident interactions were closer to each other within the network. We defined connector genes as any gene on a shortest path between a primary and secondary variant gene and defined ''hub genes'' as connector genes scoring in $\geq$95th percentile for betweenness centrality. GO and pathway enrichment analysis for unique connector genes of each 16p12.1 gene was performed as described above (Table S3C).

We also used this framework to assess the connectivity of secondary variants in DD cohort probands. We binned network genes into four quartiles based in the number of connections each gene had (1st: 0 connections, 2nd: 1-2 connections, 3rd: 3-9 connections, 4th: $\geq$10 connections) and counted the number of secondary variant genes within each quartile. To calculate empirical p-values, we compared those values to 1000 permutations in which we randomly selected the same number of genes in the genome and counted the number within each quartile (Table S3D).

16p12.1 deletion samples by ascertainment

Phenotype analysis

We assessed phenotypic data in the ascertainment-specific cohorts using ICD10 codes derived from electronic health records (EHR) (MyCode and UKB), EHR SNOMED CT codes (AoU), and self-reported questionnaire responses (UKB and SPARK) (Table S5A). Phenotypic information from SPARK was downloaded from the Simons Foundation through the SFARI Base portal (https://www.base.sfari.org). EHR data was available from participants in MyCode.[60] EHR data from UKB were identified from Data Field 41270 (ICD10 codes), while questionnaire data was identified from additional Data Fields (Table S5A). EHR data for AoU participants is available in the Controlled Tier data in the Researcher Workbench. For harmonization of phenotypic data across cohorts, SNOMED codes and ICD10 codes were matched to phenotypes from questionnaires, details of which are provided in Table S5A. For investigations of broader sets of EHR phenotypes grouped by ICD10 Chapters, SNOMED codes were mapped to ICD10 codes using standardized vocabularies distributed through Athena.[138]

SNV calling from sequencing data

SNVs for the MyCode, UKB, and SPARK cohorts were identified from whole exome sequencing (WES) data, while SNVs in AoU were identified from WGS data.

NimbleGen (SeqCap VCRome) and xGEN probes from Integrated DNA Technologies (IDT) were used for target sequence capture in the MyCode cohort.[139,140] Sequencing was performed by paired end 75bp reads on an Illumina NovaSeq or HiSeq at >20$\times$ coverage for >80% of the targeted bases. Alignments and variant calling were based on GRCh38 human genome reference sequence. Variants were called with the WeCall variant caller version 1.1.2 (https://github.com/Genomicsplc/wecall). Whole exome VCFs for SPARK samples were downloaded through the SFARI Base portal (https://www.base.sfari.org). VCFs from both cohorts were processed using the same pipeline described above for the DD cohort.

To identify SNVs from UKB individuals,[141] we accessed WES data available as multi-sample project VCFs[142] in the UK Biobank Research Analysis Platform. After splitting multi-allelic records, we applied the following set of quality control filters using Hail in the DNAnexus platform: (a) variant call rate >90%, (b) Hardy Weinberg equilibrium p-value >10-15, (c) minimum read depth > 10, and (d) at least one sample passing the allelic balance threshold of 0.2. Next, we removed variants with an intracohort frequency >0.1% and present in less than two samples. The remaining variants were then annotated using Variant Effect Predictor[109] (VEP) v.109 and dbNSFP[87,88] 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.[87,88]

To identify SNVs from WGS data in 414,830 individuals from AoU,[141] we obtained exome-overlapping variant calls as a Hail matrix table from the AoU Researcher Workbench. Preliminary quality control was performed by the AoU consortium; variants were further filtered for those with GQ $\geq$ 20, DP $\geq$ 10 (DP $\geq$ 5 for haploid calls), and allele balance between 0.2 and 0.8 for heterozygous sites. Samples flagged by the AoU team for failing quality control were excluded, along with samples identified as potential duplicates (kinship coefficient > 0.354) or those with sex at birth not reported as male or female. Rare variants (MAF <0.1%) were retained and annotated using Nirvana annotations, available through the AoU platform, classifying their predicted functional impact using the same criteria as for UKB: predicted LOF or missense (deleterious in at least seven tools). Using gnomAD[84] genome frequency annotations for five superpopulations (AFR, AMR, EAS, SAS, EUR), each variant was assigned an alternate allele frequency based on the maximum observed frequency across the AoU cohort and the gnomAD populations. All AoU analyses were performed in the AoU Researcher Workbench.

Please cite this article in press as: Jensen et al., Genetic modifiers and ascertainment drive variable expressivity of complex disorders, Cell (2025), https://doi.org/10.1016/j.cell.2025.09.012

CNV calling from sequencing data

Carriers of the 16p12.1 deletion in each cohort were identified based on CNV calls from microarray data. Samples from MyCode were genotyped using the Illumina Global Screening Array and Illumina OmniExpressExome-8 Kit. SNP log-r ratio and b-allele frequencies for SPARK samples were downloaded through the SFARI Base portal (https://www.base.sfari.org). Signals for the UK Biobank were accessed through Data Fields 22437 and 22431, and signal data from AoU samples was accessed through the Controlled Tier of data in the Researcher Workbench. CNVs for all cohorts were called using the PennCNV[94] pipeline described above for the DD cohort. Additionally, 60,228 and 58,560 additional samples without any large (>500kb), rare (<0.1% population frequency) CNVs were identified as controls for additional genetic analysis from the UK Biobank and AoU, respectively.

PRS calculations

Autism, intelligence, educational attainment, and schizophrenia PRS was calculated from microarray data for samples with reported or imputed European ancestry in MyCode, UKB, and SPARK using the same pipelines as described above for the DD cohort. BMI PRS was available for UKB samples from Data Field 26216. For AoU, PRS was calculated using common variants identified from WGS obtained from a Hail matrix table available in the AoU Researcher Workbench. These variants were restricted to only those in European samples and in SNPs present in the GWAS summary statistics. We then used Hail to sum the effect sizes for each SNP per sample. These scores were then normalized such that all samples would have a mean score of zero and standard deviation of one for each PRS.

Samples with other neurodevelopmental disorders

Exome sequencing VCFs and raw microarray data for Searchlight cohorts, whole genome sequencing VCFs and STR calls for SSC,[111] 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[111] 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,[143] while CNV calls from microarrays for the Searchlight cohort were processed using PennCNV.[27,94] 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.[40] 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).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses and experimental results for the data presented in Figures 2, 3, 4, 5, 6, and 7 and associated supplementary figures, including sample sizes (number of individuals), test statistics, confidence intervals, p-values, and FDR values, are available in Tables S2, S3, S4, S5, and S6. All statistical comparisons and model analyses were conducted using Python v.3.7 unless specifically noted below. For all analyses, Benjamini-Hochberg multiple testing correction was performed using the scipy v.1.13.1 *false_discovery_control* function, unless otherwise stated. FDR values reported in the text are corrected for multiple testing, while p-values are not corrected for multiple testing.

16p12.1 deletion DD cohort 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 noncarrier parents. Burden analysis (paired and independent t-tests) and Pearson's correlation analyses were calculated using the scipy v1.13.1 *ttest_rel, ttest_ind*, or *pearsonr* functions, respectively. We note that paired t-tests for PRS burden were two-tailed, due to the dual directionality of PRS for different phenotypes, while other t-tests were one-tailed.

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. For all models, we assessed the contribution of the burden of multiple variant categories in a proband to their phenotypic complexity score (binarized to create two groups of approximately equal size, logistic models) or quantitative phenotypic score (linear models). 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 (R^2 for linear models and McFadden's pseudo-R^2 for logistic models) was calculated for all models. Sample

sizes, odds ratios, p-values, FDR values, confidence intervals, and variance statistics for all models used in the DD cohort are available in Table S4A.

Ascertained 16p12.1 deletion cohort analysis

Comparisons of secondary variant burden between 16p12.1 deletion carriers and age- and sex-matched controls in the UK Biobank and AoU were assessed using one-tailed t-tests. These tests were not corrected for multiple testing due to the small number of tests (six tests per cohort). The effects of the 16p12.1 deletion and environmental factors on BMI in UKB were assessed using a linear model using the *OLM* function in statsmodels v.0.14.2. This model included 16p12.1 deletion carrier status, BMI PRS, and multiple environmental and lifestyle factors (physical activity, alcohol consumption, smoking status, sleep, diet, and use of anti-psychotic or anti-depressant medications) as input variables, and sex, age, age-squared, and ten genetic principal components as covariates. The model also assessed multiplicative interactions between 16p12.1 deletion carrier status and each environmental or lifestyle factor. The p-values from this model were not corrected for multiple testing because only a single model was run. The relationship of secondary variant burden and phenotypes in all ascertainment-specific cohorts (UK Biobank, MyCode, All of Us, and SPARK), and comparisons with adults and children in the DD cohort, were assessed using two-tailed t-tests. We note that phenotypes in the ascertainment 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 available as controls. Logistic regression was performed on main and secondary ICD10 codes in UKB, collapsed to "Chapter", with age and sex included as covariates. PheWAS was performed using the *PheWAS* v.0.99.6-1 package in R[110] on all available samples from the UK Biobank using Phecodes derived from ICD10 codes, while correcting for sex, age, and four genetic principal components. Sample sizes, test statistics, p-values, FDR values, confidence intervals, and variance statistics for all analyses are available in Table S5.

Multicohort meta-analysis

To examine consistent associations of secondary variants and phenotypes across cohorts, we performed a meta-analysis of the t-test results from (a) DD cohort children and SPARK, (b) DD cohort adults and UKB results using clinical questionnaires, and (c) MyCode, AoU, and UKB results using EHR data. We used the *metacont* function from the R package *meta* v.8.0-2 to calculate a random effect estimate for each variant-phenotype association for each comparison group. The random effect estimates, confidence intervals, p-values, and FDR values for all analyses are available in Table S5K.

Neurodevelopmental disease cohort analysis

Linear regression models for assessing variation in quantitative developmental phenotypes among probands in the Simons Searchlight and SSC cohorts 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). All 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, p-values, FDR values, confidence intervals, and variance statistics for all analyses are available in Table S6.

CellPress
OPEN ACCESS

Supplemental figures

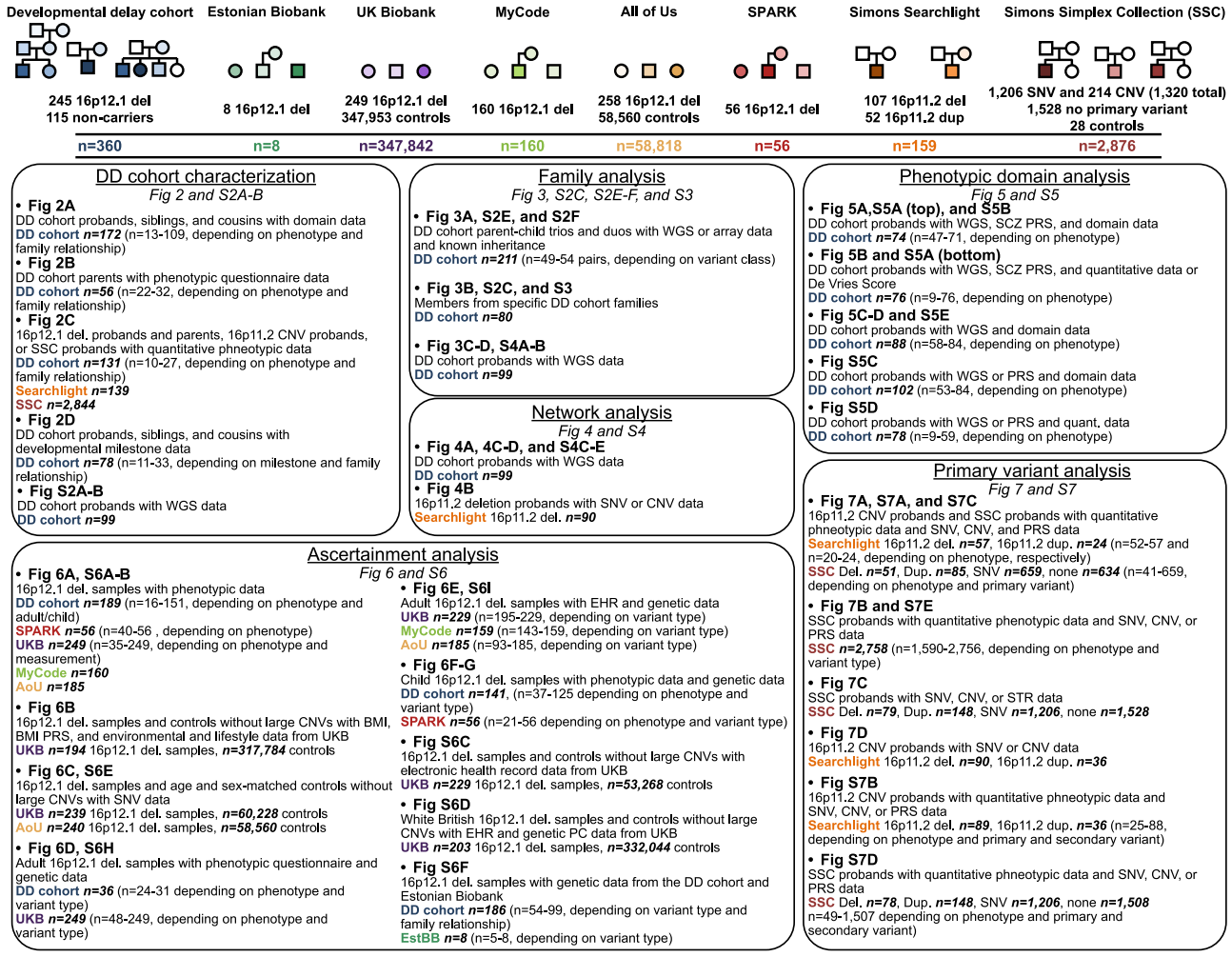


Figure S1. Analyses in 2,455 individuals with primary pathogenic variants, related to Figure 1

Descriptions of cohorts, data types, and samples used in all analyses. The number of samples with primary variants and controls are listed for each cohort, as well as the total number of samples considered in each cohort. Related analyses are grouped into broad categories and annotated with the figure(s) where each data type is used. The bolded sample sizes indicate the total number of samples used in each analysis from a single cohort, while the sample sizes in parentheses indicate the range of samples sizes for specific comparisons within a given analysis. In some cases, the total number of samples used is greater than the maximum sample size in the range as comparisons may consider non-overlapping groups of samples.

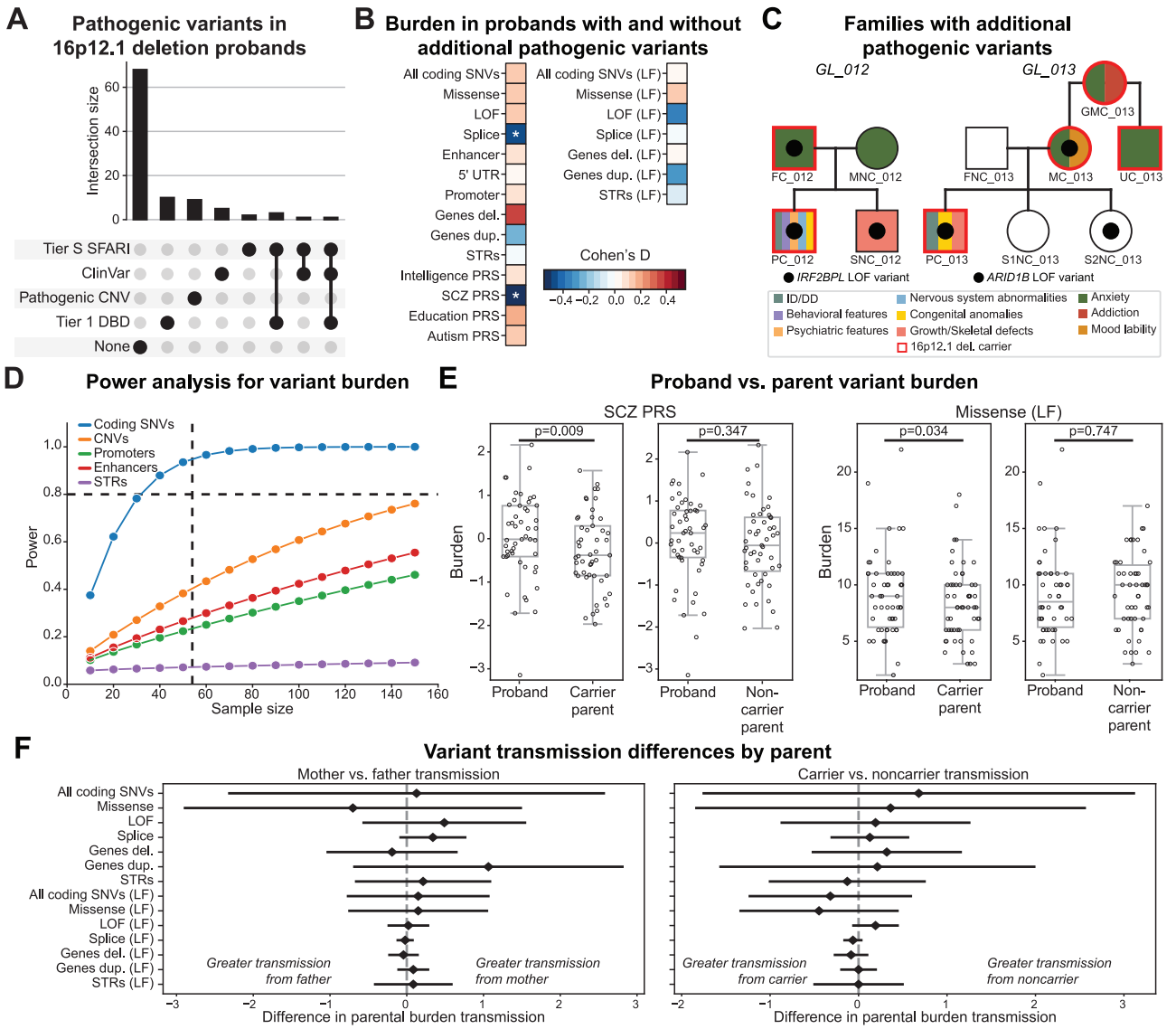


Figure S2. Disease association, statistical power, and burden of secondary variants, related to Figure 3

(A) UpSet plot shows the number of 16p12.1 del. probands with secondary variants in one or more disease-associated categories, potentially indicative of multiple genetic diagnoses. Analysis restricted to probands with WGS data.

(B) Two-tailed *t* tests comparing the variant burden of probands with and without additional pathogenic variants. $^{*}p < 0.05$.

(C) Pedigrees from GL_012 (left) and GL_013 (right) show inheritance of LoF variants in *IRF2BPL* (left) and *ARID1B* (right) in multiple family members, each with distinct phenotypes. Black circles, carriers of LoF variants; colors, phenotypes. White indicates no phenotype.

(D) Power analysis for detecting changes in burden of rare variant classes among individuals with the 16p12.1 del. (see STAR Methods). Dashed horizontal line indicates 80% power; dashed vertical line represents the sample size of proband-carrier parent pairs ($n = 54$ in the DD cohort).

(E) Changes in burden of schizophrenia (SCZ) PRS (left) and missense (LF) variants (right) between 16p12.1 del. probands and their carrier (left, $n = 49, 54$) and noncarrier parents (right, $n = 51, 50$). Boxes, quartiles; bars, range excluding outliers defined by IQR. *p* values from one-tailed (rare variants) or two-tailed (SCZ PRS) *t* tests.

(F) Effect sizes and 95% confidence intervals from paired *t* tests comparing the transmission of rare variant classes from each parent for probands in the DD cohort ($n = 47$).

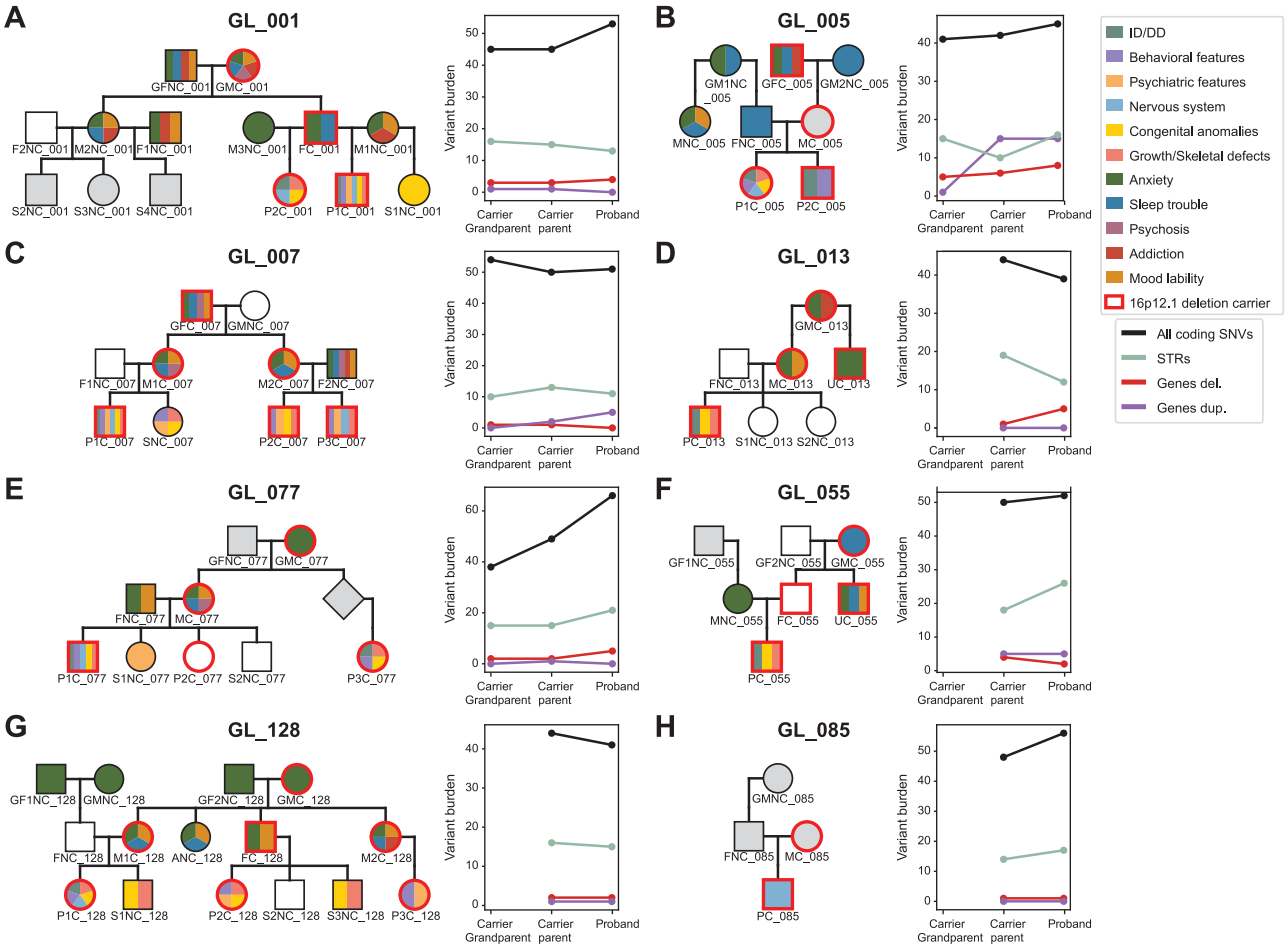


Figure S3. Pedigrees of three-generation families, related to Figure 3

(A–H) Pedigrees and changes in secondary variant burden for eight three-generation families from the DD cohort. Red outline, 16p12.1 del. carrier; colors, phenotypes. White, no identified phenotypes; gray, no data.

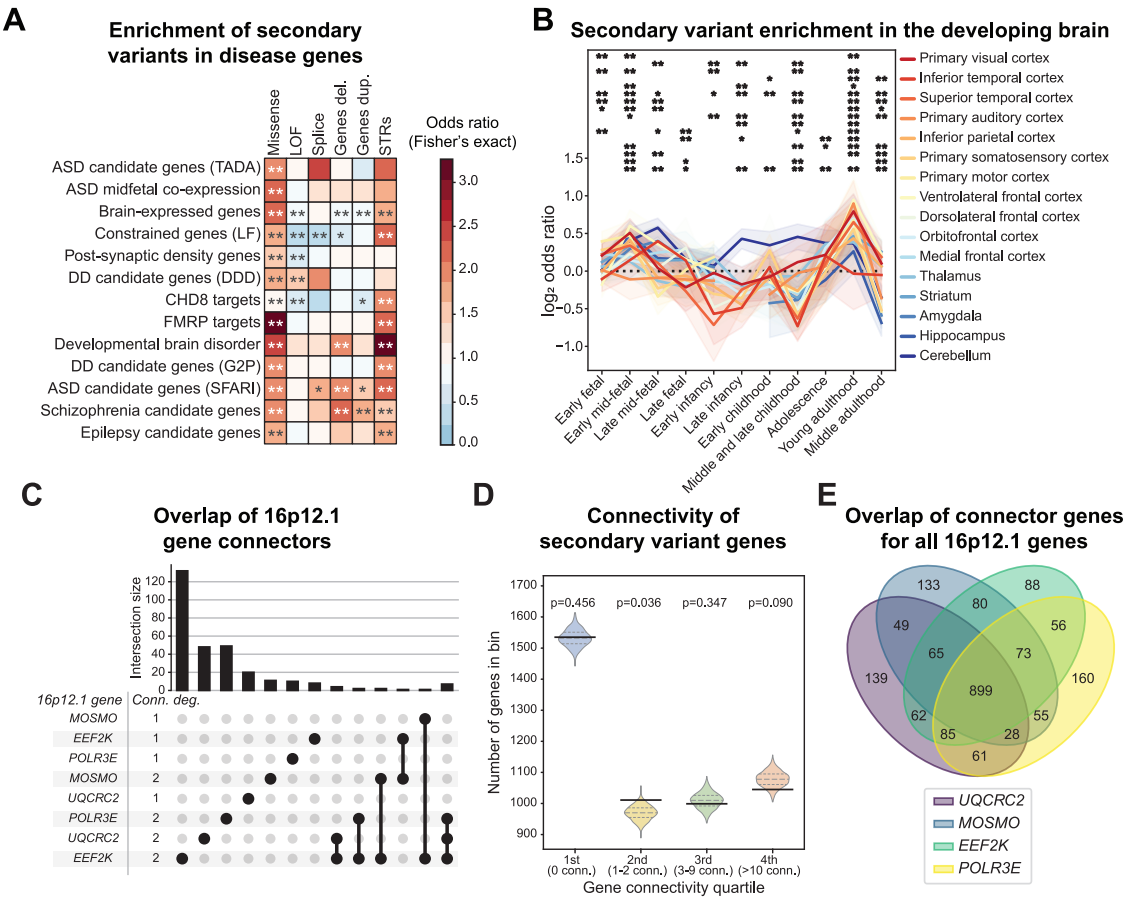


Figure S4. Functional effects and network connectivity of secondary variants observed in 16p12.1 del. probands, related to Figures 3 and 4

(A) Enrichment of secondary variant classes in 16p12.1 del. probands for sets of genes involved with neurodevelopmental disease and related functions. Fisher's exact test, *$p \leq 0.05$, **Benjamini-Hochberg FDR ≤ 0.05.

(B) Enrichment (log-odds ratios with 95% confidence intervals; *y* axis) of secondary variants in 16p12.1 del. probands among genes preferentially expressed in 16 brain tissues (colored lines) over 11 developmental time points (*x* axis). Fisher's exact test, *$p \leq 0.05$, **Benjamini-Hochberg FDR ≤ 0.05.

(C) UpSet plot for the overlap of first- and second-degree connector genes of 16p12.1 genes. Conn. deg., connectivity degree.

(D) Violin plot showing the connectivity of secondary variant genes in the STRING network compared with distributions of connectivity from 1,000 random permutations of genes in the genome. Black line, true number of secondary variant genes in each quartile; gray lines, median and first and third quartiles of distributions.

(E) Overlap of all connector genes for each 16p12.1 gene.

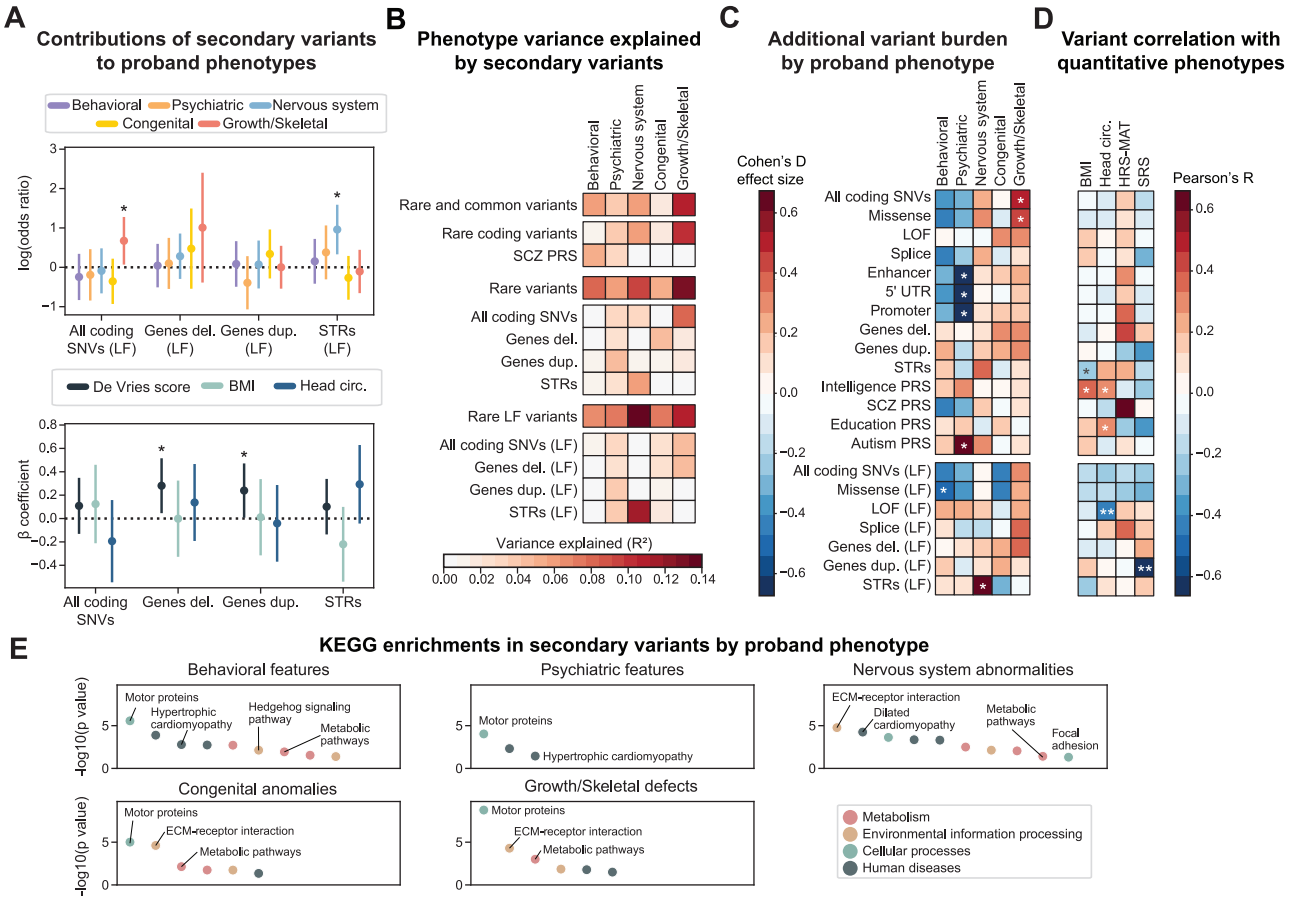


Figure S5. Secondary variant associations with 16p12.1 del. phenotypic domains, related to Figure 5

(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. *$p \leq 0.05$. (Bottom) β coefficients from linear regression models for quantitative phenotypes in probands (n = 43–76). Bars, 95% CIs. *$p \leq 0.05$.

(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 models that include multiple classes. Full regression results are available in Table S4A.

(C) Comparisons of secondary variant burden between 16p12.1 del. probands (n = 53–84) with higher and lower complexity scores for each phenotypic domain. Two-tailed t test, *$p \leq 0.05$, **Benjamini-Hochberg FDR ≤ 0.05.

(D) Pearson correlations between quantitative phenotypes and secondary variant burden in del. probands (n = 9–59). *$p \leq 0.05$, **Benjamini-Hochberg FDR $\leq$ 0.05.

(E) KEGG pathway enrichments of secondary variants from probands within each phenotypic domain. Colors indicate broad KEGG classifications (x axis) from the KEGG BRITE functional hierarchy.

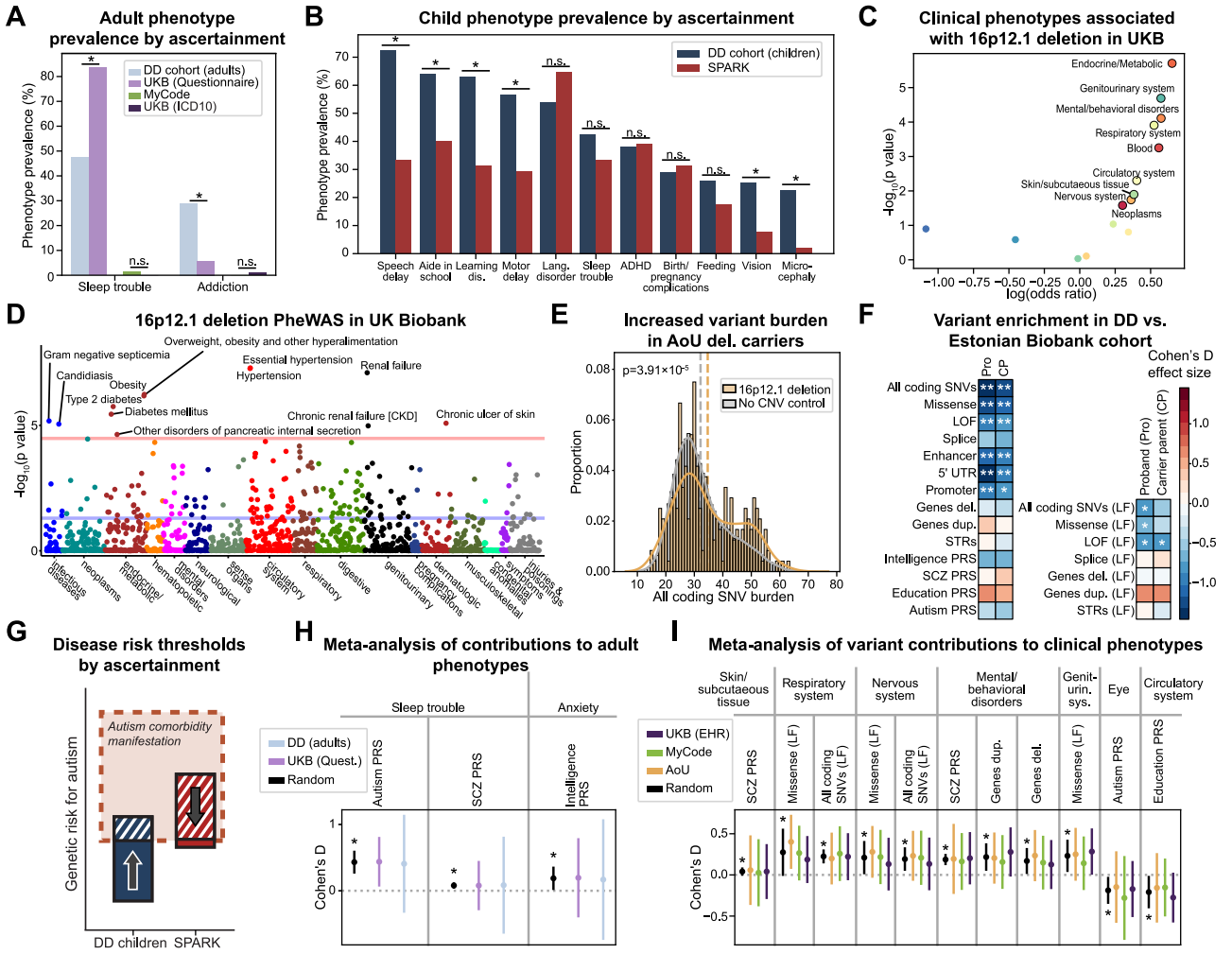


Figure S6. Effects of ascertainment on phenotypes and secondary variant associations in 16p12.1 del. carriers, related to Figure 6

(A) Comparison of sleep disturbance and addiction phenotype prevalence in adults from the DD cohort (n = 38) and individuals from UKB (questionnaire n = 35, 249, respectively; EHR n = 229) and MyCode (n = 160). *$p \leq 0.05$, Fisher's exact test. Phenotype frequencies are not reported for AoU as they were present in <20 AoU samples.

(B) Prevalence of developmental and psychiatric phenotypes in children with 16p12.1 del. from DD (n = 80–151) and SPARK (n = 40–51) cohorts. Phenotypes shown were restricted to those present in >20% of probands in either cohort. *$p \leq 0.05$, Fisher's exact test. Full results are available in Table S5B.

(C) Enrichment of select ICD10 chapters from logistic regression models in UKB 16p12.1 del. carriers (n = 229) compared with controls without large rare CNVs (n = 53,268). Labeled points indicate phenotypes with Benjamini-Hochberg FDR ≤ 0.05. Full results are available in Table S5C.

(D) PheWAS analysis for the 16p12.1 del. in UKB (n = 117,611–332,247). Colored points, groups of related phenotypes; red line, phenome-wide significance (Bonferroni $p = 0.05$); blue line, nominal significance ($p = 0.05$). Full results are available in Table S5E.

(E) Comparison of SNV burden in AoU individuals with 16p12.1 del. (n = 240, yellow) to age and sex-matched controls without large rare (>500 kb) CNVs (n = 58,560, gray). p value from one-tailed t test. Dashed lines indicate group means. Full results are available in Table S5F.

(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 del. individuals from the Estonian Biobank (n = 5–8). Blue indicates a depletion in secondary variant burden among Estonian Biobank del. carriers. One-tailed t test for rare variants, two-tailed t test for PRS. *$p \leq 0.05$, **Benjamini-Hochberg FDR ≤ 0.05.

(G) Schematic outlining the proposed relationship among genetic risk factors in individuals with 16p12.1 del. 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 associations with comorbid features. However, these associations may be observed in cohorts with different ascertainments (such as the DD cohort).

(H and I) Random-effects meta-analyses of associations of secondary variant burden with psychiatric questionnaire (H) and clinical EHR (I) phenotypes in 16p12.1 del. adults from the DD (blue), UKB (purple), MyCode (green), and AoU (yellow) cohorts. Bars, 95% CIs. *$p \leq 0.05$. Full meta-analyses results are available in Table S5K.

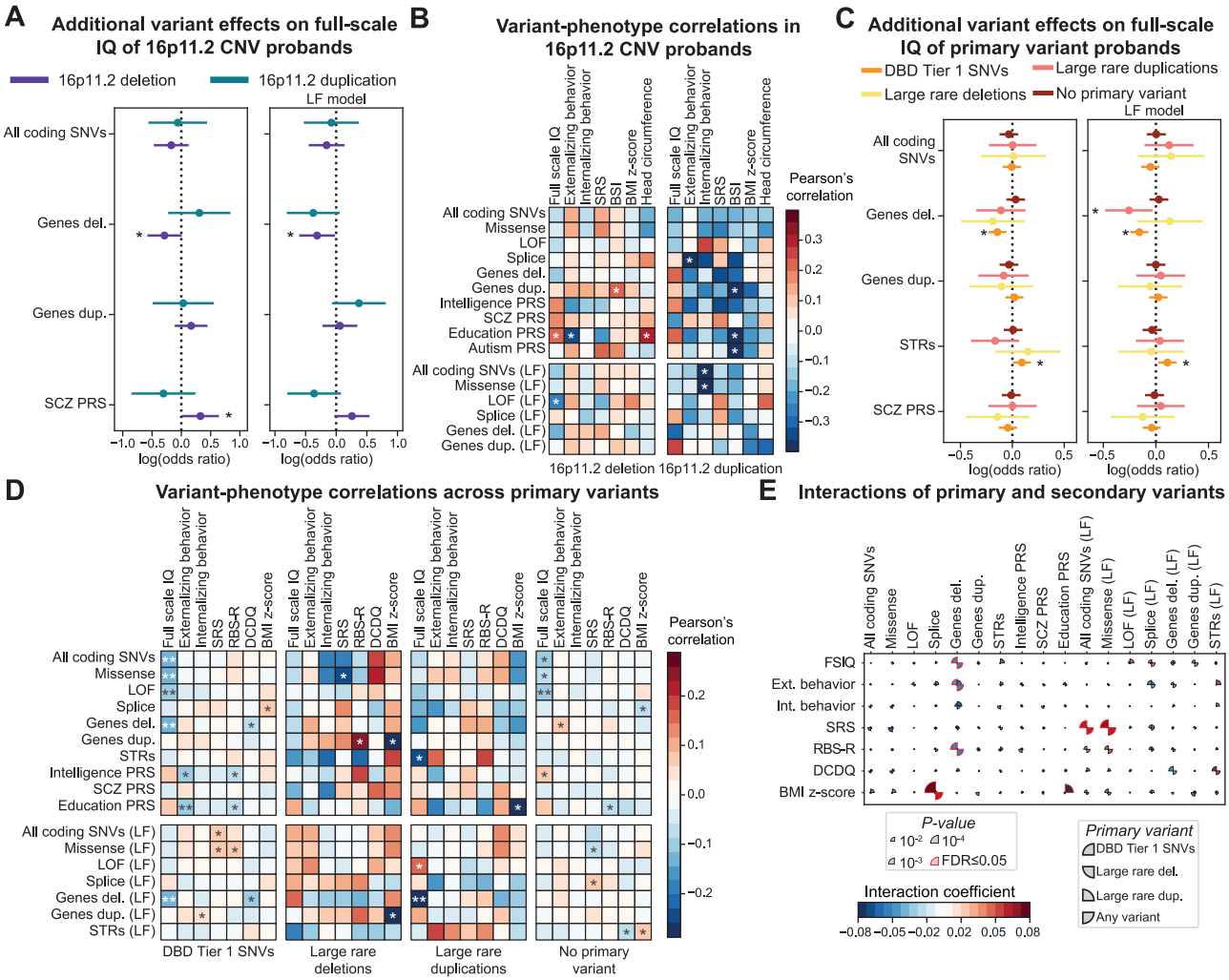


Figure S7. Associations between secondary variants and developmental features of 16p11.2 CNV probands, related to Figure 7

(A) Example forest plots show results from select linear regression models for associations between secondary variant classes and full-scale IQ for probands with the 16p11.2 del. (n = 57, purple) and duplication (n = 23, teal). Bars, 95% CIs. *$p \leq 0.05$. Full results are available in Table S6A.

(B) Pearson's correlations between secondary variant burden and quantitative phenotypes of probands with 16p11.2 del. (n = 58–88) and duplications (n = 25–34). *$p \leq 0.05$.

(C) 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 = 659, orange), large rare del.s (n = 51, yellow), and duplications (n = 85, pink), as well as probands without such variants (n = 633, red) from the SSC cohort. Bars, 95% CIs. *$p \leq 0.05$. Full results are available in Table S6A. In (A) and (C), "LF model" indicates models where rare variants are selected for genes under evolutionary constraint.

(D) Pearson's correlations between secondary variant burden and quantitative developmental phenotypes of SSC probands with pathogenic SNVs (n = 735–1,206), rare del.s (n = 49–78), rare duplications (n = 102–148), and probands without such variants (n = 717–1,507). *$p \leq 0.05$, **Benjamini-Hochberg FDR ≤ 0.05.

(E) β coefficients from linear regression models examining interactive effects of primary variants (wedges) and specific secondary variant classes (x axis) on quantitative phenotypes (y axis) in SSC probands (n = 1,590–2,756). Wedge color indicates interaction coefficients and wedge size indicates p value for strength of interaction coefficient. Red outlines indicate Benjamini-Hochberg FDR ≤ 0.05. Full results are available in Table S6D.