https://doi.org/10.1038/s41588-018-0288-4

nature genetics

Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity

Bradley P. Coe[1], Holly A. F. Stessman[2], Arvis Sulovari[1], Madeleine R. Geisheker[1], Trygve E. Bakken[3], Allison M. Lake[4], Joseph D. Dougherty[4], Ed S. Lein[3], Fereydoun Hormozdiari[5], Raphael A. Bernier[6] and Evan E. Eichler[1,7]*

We combined de novo mutation (DNM) data from 10,927 individuals with developmental delay and autism to identify 253 candidate neurodevelopmental disease genes with an excess of missense and/or likely gene-disruptive (LGD) mutations. Of these genes, 124 reach exome-wide significance ($P<5\times10^{-7}$) for DNM. Intersecting these results with copy number variation (CNV) morbidity data shows an enrichment for genomic disorder regions (30/253, likelihood ratio (LR) +1.85, $P=0.0017$). We identify genes with an excess of missense DNMs overlapping deletion syndromes (for example, *KIF1A* and the 2q37 deletion) as well as duplication syndromes, such as recurrent *MAPK3* missense mutations within the chromosome 16p11.2 duplication, recurrent *CHD4* missense DNMs in the 12p13 duplication region, and recurrent *WDFY4* missense DNMs in the 10q11.23 duplication region. Network analyses of genes showing an excess of DNMs highlights functional networks, including cell-specific enrichments in the D1$^+$ and D2$^+$ spiny neurons of the striatum.

The importance of DNMs underlying neurodevelopmental disorders (NDDs) has been recognized for many years. Some of the strongest genome-wide evidence came from early CNV studies, which consistently showed an excess of de novo and large private CNVs in patients with autism, developmental delay (DD), and epilepsy[1–4]. Significance based on CNV recurrence was more readily achieved from smaller sample sizes because of elevated mutation rates in regions flanked by segmental duplications[5] or hotspots of recurrent rearrangement near telomeres[6]. In many cases, the individual genes underlying the genomic disorders remain unknown.

The advent of next-generation sequencing and exome sequencing rapidly accelerated our ability to specify genes associated with potentially pathogenic de novo single-nucleotide variants (SNVs) for both DD and autism[7,8], although recurrent mutations occurred more rarely[9,10]. Different statistical models for discovery of disease-associated genes based on recurrent SNV mutation have been developed, including those based on chimpanzee–human divergence[11], trinucleotide mutation context[12], and clustering of DNMs[13–15]. Despite extensive CNV analyses of nearly 45,000 patients with autism and DD[16–18], few attempts have been made[18,19] to integrate the wealth of CNV data with recent exome sequencing results despite a common mutational model of dosage imbalance.

In this study, we perform an integrated meta-analysis combining DNMs from exome sequence data from individuals with autism spectrum disorder (ASD), intellectual disability (ID) and/or DD (hereafter referred to as ID/DD)[20] with CNV morbidity data. Because of the significant comorbidity between ID and ASD[21,22] and the fact that cases of autism with a severe DNM are enriched in DD[23], we overlay these data with known genomic disorders. The goals of this study are threefold: (1) provide an integrated list of candidate NDD genes based on multiple lines of DNM and CNV evidence; (2) compare different models of recurrent mutation; and (3) identify the most likely genes underlying pathogenic microdeletion and microduplication CNVs associated with DD.

Results

Genes enriched for de novo SNV mutation and model comparisons. We compiled de novo variation identified from exome sequencing of 10,927 cases with NDDs from the denovo-db v.1.5 database release[20]. This includes 5,624 cases with a primary diagnosis of ASD and 5,303 cases with a diagnosis of ID/DD collected from 17 studies[11,23–38] (Supplementary Table 1). We consider all protein-altering and LGD mutations, including frameshifts, splice donor or acceptor mutations, start losses, and stop gains. The combined set of 12,172 DNMs includes 2,357 LGD and 9,815 missense mutations.

We initially applied two statistical models. The first incorporates locus-specific transition, transversion, and indel rates and chimpanzee–human coding sequence divergence[11] to estimate the number of expected DNMs, hereafter referred to as the chimpanzee–human divergence model or the CH model. The second model, denovolyzeR[12], estimates mutation rates based on trinucleotide context and incorporates exome depth and divergence adjustments based on macaque–human comparisons over a ±1-Mb window and accommodates known mutational biases, such as CpG hotspots. Both models apply their underlying mutation rate estimates to generate prior probabilities for observing a specific number and class of mutations for a given gene. While both models incorporate LGD

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [2]Department of Pharmacology, Creighton University Medical School, Omaha, NE, USA. [3]Allen Institute for Brain Science, Seattle, WA, USA. [4]Department of Genetics, Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA. [5]Department of Biochemistry and Molecular Medicine, University of California, Davis, Davis, CA, USA. [6]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA. [7]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. *e-mail: eee@gs.washington.edu

and missense probabilities, we recently modified the CH model to incorporate CADD scores[39,40], allowing us to also specifically test for enrichment for the missense subset of the predicted most severe 0.1% of mutations (that is, CADD scores over 30, or MIS30). Such missense mutations are more likely to be functionally equivalent to an LGD mutation and have been shown to be significantly enriched in NDD cases compared with controls[14]. To account for the sensitivity biases, we applied an upper bound baseline mutation rate assumption of 1.8 DNMs (derived from high-coverage genome sequencing data) to the CH model, which exceeds the overall DNM rate of this cohort.

Combined, the two models (union set) identify 253 candidate NDD genes with evidence of excess DNM at a false discovery rate (FDR) <5% and at least two mutations for at least one mutational category (Tables 1 and 2). This includes 145 genes with excess LGD mutations and 123 genes with excess missense mutations. Among these genes, 29 demonstrate evidence of both LGD and missense mutations (Fig. 1, Table 1, and Supplementary Table 2). In general, both models highlight similar genes (Fig. 1 and Supplementary Note), particularly for LGD events where 73.1% (106/145) of genes are shared. This finding stands in contrast with recurrent missense DNMs, where only 51.2% (63/123) of the genes overlap between the models, suggesting that additional model refinement is required to more accurately predict pathogenic missense mutations. A more stringent application of the exome-wide Bonferroni family-wise error rate (FWER) identifies a union of 124 genes (Supplementary Table 2 and Supplementary Note).

We identified additional evidence of disease association, from a comprehensive database and PubMed literature search, for 204/253 genes (Methods), indicating that 49/253 union genes (10/124 FWER union genes) presented here are novel associations (Supplementary Tables 2 and 3). Of these novel genes, 61% (30/49) demonstrate DNM in both ASD and ID/DD patients. We wish to note that neither FDR nor FWER, however, are metrics of pathogenicity, but rather thresholds of significance to identify genes for further investigation.

Because it has been well established that NDD genes are less tolerant to mutation, we categorized the 253 genes into different functional groups and compared their tolerance to mutation in the general population using three metrics: residual variance to intolerance score (RVIS), probability of loss-of-function intolerance (pLI), and missense constraint scores (missense Z scores). For both pLI and missense Z scores, we utilized the ExAC subset with known neuropsychiatric cohorts removed (45,376 individuals)[41]. For genes with an enrichment of LGD variants, we observed a significant increase in pLI scores compared to all other genes ($P = 8.3 \times 10^{-58}$, two-tailed Wilcoxon rank-sum test, ROC AUC = 0.90) (Fig. 1e). We also observed a significant increase in missense constraint (missense Z scores) among genes with enrichment for missense variation ($P = 2.0 \times 10^{-44}$, two-tailed Wilcoxon rank-sum test, ROC AUC = 0.87) (Fig. 1f). Similarly, we observed a significant RVIS depletion for all categories in which at least two genes were identified (Fig. 1g).

Examination of a combination of constraint metrics is particularly valuable, as a small number of genes demonstrate conflicting results, such as the LGD- and missense-enriched gene *MEF2C*, which is involved in severe ID when disrupted by deletions or mutations[42] and demonstrates constraint by RVIS but not by pLI (RVIS = 18.97, pLI = 2.4×10^{-3}, missense Z score = 4.47). Interestingly, among genes without pLI or RVIS support, we identified established ASD and ID genes in addition to the expected potential false positives. Among the union genes, 82.6% (209/253) have no detected LGD or missense DNM in controls ($n = 2{,}278$ controls; Supplementary Note). The detection of control events may represent incomplete penetrance, variable expressivity, undiagnosed or subclinical controls, or benign variation (primarily in the case of missense variation). None of the recurrent control DNM genes reach exome-wide FDR significance. Although some of these are plausible candidates, disease significance should be considered with caution until additional functional and clinical data establish their role.

ASD versus ID/DD genes. We investigated the distribution of LGD and missense mutations between the ASD and ID/DD cohorts. The majority of this NDD gene set (68.4% (173/253) of the union and 72.3% (107/148) of the intersection show evidence of DNM in both ASD and ID/DD cohorts, highlighting the utility of joint analyses. Although a small number of genes are specific or enriched for a diagnosis of ID/DD ($q < 0.01$, one-tailed Fisher's exact test), none are statistically enriched for ASD yet (Supplementary Fig. 1). A few genes (*WDFY3*, *DSCAM*, and *CHD8*) trend toward ASD diagnosis. To eliminate potential ascertainment bias in discovery, we repeated the overlap analysis considering gene discovery independently in each cohort (Supplementary Note). Considering all 253 genes and the full set of NDD patients, we calculate that 17.7% (1,932/10,927) of the samples have at least one de novo event in this gene set.

The proportion of patients with a DNM is significantly higher for ID/DD (26.8% or 1,421/5,303 patients) compared to ASD (9.1% or 511/5,624 patients) (OR = 3.66, $P = 1.62 \times 10^{-133}$, two-tailed Fisher's exact test). While this approach may be partially biased by the differences in DNM sensitivity between exome studies, this observation exceeds the baseline 1.58-fold excess of LGD DNM in ID/DD cohorts and thus reflects differences in heterogeneity between the disorders. Further supporting this bias of LGD events

Table 1 | Recurrent DNM gene summary and model comparison

					ID/DD only	ASD only
Variant category	CH model count	denovolyzeR count	Union count (FDR (FWER))	Intersection count (FDR (FWER))	Union count (FDR (FWER))	Union count (FDR (FWER))
LGD and MIS30 and missense	14	NA	14 (5)	*NA*	4 (1)	0 (0)
LGD and MIS30	1	NA	1 (3)	*NA*	0 (0)	0 (0)
LGD and missense	13	22	15 (6)	21 (8)	19 (10)	3 (1)
LGD	92	108	115 (62)	85 (49)	95 (55)	31 (14)
MIS30 and missense	28	NA	31 (18)	*NA*	9 (7)	0 (0)
MIS30	16	NA	14 (10)	*NA*	0 (0)	0 (0)
Missense	46	63	63 (20)	42 (22)	56 (25)	7 (3)
Total	210	193	253 (124)	148 (79)	183 (98)	41 (18)

The number of genes reaching statistical significance for DNM enrichment in $n = 10{,}927$ independent samples by the CH model and denovolyzeR as well as the unions and intersections of these gene sets are shown for each mutation category (LGD, missense, MIS30). Also shown are the union counts for the ASD ($n = 5{,}624$ independent samples) and ID/DD ($n = 5{,}303$) only analyses. Counts represent genes passing an FDR q-value threshold of 0.05. Bracketed numbers represent genes passing a Bonferroni FWER correction ($P < 5 \times 10^{-7}$).

Table 2 | Genes enriched for de novo variation in 10,927 ASD/ID/DD patients in denovo-db v.1.5

Significance category	Genes
LGD	*ADNP*[a*], *AHDC1*[a*], *ANK2*[a], *ANKRD11*[a*], *ANP32A*[*], *ARID1B*[a*], *ARID2*[a], *ASH1L*[a*], *ASXL1*[a], *ASXL3*[a*], *AUTS2*[a], *BCL11A*[a*], *BRPF1*[a], *CAPRIN1*[a*], *CASZ1*[a], *CDC42BPB*[a], *CDKL5*[a], *CHAMP1*[a*], *CHD7*[a], *CHD8*[a*], *CLTC*[a], *CNKSR2*[a*], *CNOT3*[a], *CTNNB1*[a*], *CUL3*[a*], *DLG4*[a*], *DSCAM*[a], *DVL3*[a], *EBF3*[a], *EHMT1*[a*], *ENO3*[a*], *EP300*[a*], *FAM200A*[a], *FAM200B*[a], *FOSL2*, *FOXP2*[a], *GATAD2B*[a*], *HIST1H1E*[a], *HIVEP2*[a], *HIVEP3*[a*], *HNRNPD*[a], *IRF2BPL*[a], *KANSL1*[a*], *KAT6A*[a*], *KAT6B*[a*], *KCNS3*[a], *KDM5B*[a*], *KDM6A*[a], *KDM6B*[a], *KIAA2022*[a*], *KIF11*[a], *KMT2A*[a*], *KMT2C*[a], *LARP4B*[a], *LEO1*[a*], *MBD5*[a*], *MEIS2*[a*], *MSL3*[a*], *NAA15*[a*], *NFE2L3*[a], *NONO*[a*], *NSD1*[a*], *ODC1*[a], *PDHA1*[a], *PHF12*[a], *PHF21A*[a], *PHF3*[a], *PHIP*[a], *POU3F3*[a], *PPM1D*[a*], *PRR12*[a], *PTCHD1*[a], *QRICH1*[a*], *RAI1*[a], *RPL26*[a], *SET*[a*], *SETBP1*[a*], *SETD2*[a], *SETD5*[a*], *SHANK3*[a*], *SIN3A*[a], *SKIDA1*[a], *SMC1A*[a*], *SON*[a], *SOX5*[a*], *SPAST*[a], *SPEN*[a], *SPRY2*[a], *SRCAP*[a*], *SRRM2*, *SRSF11*[a], *STARD9*[a], *SUV420H1*[a*], *SYNCRIP*[a*], *SYNGAP1*[a*], *TAB2*[a], *TBR1*[a], *TCF12*[a], *TCF20*[a*], *TNRC6B*, *TRA2B*[a], *TRIP12*[a*], *UPF3B*[a*], *USP9X*[a*], *VEZF1*, *WAC*[a*], *WDFY3*[a], *WDR45*[a*], *WDR87*[a], *WHSC1*[a], *YTHDF3*[a], *ZBTB18*[a*], *ZBTB7A*[a], *ZC4H2*[a], *ZNF292*[a]
LGD and missense	*CHD2*[a*], *CREBBP*[a*], *DYRK1A*[a*], *FBXO11*[a], *FOXG1*[a*], *FOXP1*[a*], *HNRNPU*[a*], *MEF2C*[a*], *MYT1L*[a*], *NFIX*[a*], *POGZ*[a*], *PTEN*[a*], *PURA*[a*], *TLK2*[a*], *WDR26*[a*]
LGD and MIS30	*TCF7L2*[a]
LGD, missense and MIS30	*CASK*[a*], *DDX3x*[a*], *HDAC8*[a*], *IQSEC2*[a*], *MECP2*[a*], *MED13L*[a*], *PPP2R5D*[a*], *PUF60*[a*], *SATB2*[a*], *SCN2A*[a*], *SLC6A1*[a*], *STXBP1*[a*], *TBL1XR1*[a*], *TCF4*[a*]
Missense	*ABI2*[a], *ACHE*[a], *ADAP1*[a*], *AGAP2*[a], *AGO1*[a], *AGO4*[a*], *AQP10*[a], *BRAF*[a*], *BTF3*[a], *C2orf42*[a], *CABP7*[a], *CAPN15*[a*], *CBL*[a], *CHD4*[a], *CLASP1*[a], *DEAF1*[a], *DLX3*[a*], *DNM1*[a], *EGLN2*[a], *GABRB2*[a*], *GABRB3*[a], *GLRA2*[a], *GNAI1*[a*], *HMGXB3*[a], *HUWE1*[a], *ITPR1*[a*], *KCNC1*[a], *KCNJ6*[a], *MAPK3*[a], *MTF2*[a], *MYO1E*[a], *PBX1*[a], *PLAC8L1*[a], *PLK5*[a], *PPP1CB*[a*], *PRKCA*[a], *PRKD1*[a], *PRPF18*[a], *PSMG4*[a], *PTPN11*[a*], *RAC1*[a*], *RFX8*[a], *RRP8*[a], *RYR2*[a], *SETD1B*[a], *SF3B1*[a], *SHISA6*[a], *SMAD4*[a], *SMARCD1*[a*], *SMC3*[a], *SNAPC5*[a*], *SNX5*[a*], *SUSD4*[a], *SYT1*[a], *TAOK1*[a], *TMEM178A*[a*], *TMEM42*[a], *TNPO3*[a], *TRAF7*[a], *TRRAP*[a], *UNC80*[a], *VAMP2*[a], *WDFY4*[a], *YWHAG*[a]
MIS30	*ACTC1*[a], *AGO3*[a], *CACNA1E*[a], *FAM104A*[a], *HIST1H2AC*[a], *KIF5C*[a], *PACS2*[a*], *PAPOLG*[a], *PDK2*[a], *SEPT10*[a], *STC1*[a], *TAF1*[a*], *TNPO2*[a*], *U2AF2*[a]
Missense and MIS30	*CDK13*[a*], *CHD3*[a*], *COL4A3BP*[a*], *CSNK2A1*[a*], *CTCF*[a*], *DNMT3A*[a*], *DYNC1H1*[a*], *EEF1A2*[a*], *EFTUD2*[a*], *GNAO1*[a*], *GRIN2B*[a*], *HECW2*[a*], *KCND3*[a], *KCNH1*[a*], *KCNQ2*[a*], *KCNQ3*[a*], *KIF1A*[a*], *MAP2K1*[a*], *NAA10*[a*], *NR2F1*[a], *NR4A2*[a], *PACS1*[a*], *PIK3CA*[a*], *PPP2R1A*[a*], *RAB11A*[a], *SCN8A*[a*], *SMARCA2*[a*], *SMARCA4*[a*], *TRIO*[a*], *ZMYND11*[a*]

[a] Gene also in the intersection set. *Gene in FWER exome-wide significance ($P<5\times10^{-7}$) set. Listing of genes reaching significance for excess of DNM in $n=10{,}927$ independent samples at an FDR of 5% by either the denovolyzeR or CH model (union) for each mutational category (LGD, missense, MIS30).

to ID/DD, Simons Simplex Collection (SSC) autism probands carrying an LGD DNM in one of the LGD genes were less likely to be high-functioning. Instead, there was an over-representation of ID (≤70 IQ, expected 753/2,445, observed 42/95) and low-to-normal IQ ranges (70<IQ<100, expected 1,018/2,445, observed 45/95) compared to the high-IQ probands (≥100 IQ, expected 674/2,445, observed 8/95) ($P=4.0\times10^{-5}$, two-tailed likelihood ratio test) (Supplementary Table 4). Among the autism mutation carriers, we observe a nominal enrichment for increased severity of repetitive behavior (RBS-R ($t(18)=3.12$, corrected $P=0.048$), two-tailed independent samples *t* test) but were overall underpowered to further disentangle associated phenotypic features.

Network enrichment and patterns of brain expression. Examination of the 253 genes identified by the union of both statistical approaches suggests that our set is strongly enriched for functionally related networks of genes. The STRING database, for example, identifies a highly significant 1.8-fold enrichment (1,067 edges vs. 573 expected) in interactions among the 253 union genes ($P<1.0\times10^{-16}$, one-tailed hypergeometric test). Given this high level of interconnectivity, we applied MAGI[43], a gene network discovery tool, to identify potential gene clusters, functional enrichments, and additional candidate interactions and highlight the top four protein−protein interaction (PPI) and coexpression networks (modules 1–3 at $P<0.01$, module 4 at $P<0.05$, one-tailed permutation test) and their associated PANTHER functional enrichments (Fig. 2a–d and Supplementary Table 5–7).

Module 1 (20 genes) delineates 'regulation of transcription from RNA polymerase II promoter' ($P=0.0269$, one-tailed Bonferroni adjusted binomial test) (Fig. 2a and Supplementary Tables 5–7) and contains 15 significant genes in addition to three candidates that do not yet reach significance (*CREB1*, *RBBP5*, *CBX5*) and two genes (*SREK1*, *SMARCB1*) with no DNM in our current data set. Module 2 highlights multiple functions relating to neurotransmitter signaling ($P=0.0358$; one-tailed Bonferroni adjusted binomial test) and synaptic signaling ($P=8.91\times10^{-5}$, one-tailed Bonferroni-adjusted binomial test) (Fig. 2b and Supplementary Tables 5–7) and contains nine significant genes, ten genes that do not reach significance (*DLG2*, *HTT*, *AP2A2*, *AP2M1*, *KCNJ4*, *KCNB1*, *STX1A*, *GRIN2A*, *GRIN1*, *CAMK2A*) and one gene with no DNM (*PRKCB*). Module 3 highlights the 'transmembrane receptor protein serine/threonine kinase signaling pathway' ($P=0.002$; one-tailed Bonferroni-adjusted binomial test) (Fig. 2c and Supplementary Tables 5–7) and contains eight significant genes, 21 genes that do not reach significance (*SMURF2*, *SMURF1*, *CDC73*, *RNPS1*, *RBBP4*, *UBE3A*, *CUL1*, *CHMP1A*, *FBXW11*, *VCP*, *VPS4A*, *PPP5C*, *PRPF38A*, *SKIL*, *HSPA4*, *PSMD3*, *UIMC1*, *GAPVD1*, *NLGN2*, *GTF3C1*, *NRXN1*) and six genes with no DNM (*ING2*, *SRSF4*, *FAF1*, *UBC*, *HSP90AB1*, *YWHAB*). Finally, module 4 highlights c-Jun N-terminal kinase (JNK) ($P=4.65\times10^{-5}$, one-tailed Bonferroni-adjusted binomial test) and mitogen-activated protein kinase (MAPK) ($P=3.13\times10^{-6}$, one-tailed Bonferroni adjusted binomial test) cascades (Fig. 2d and Supplementary Tables 5–7) and contains two significant genes in addition to 15 genes that do not reach significance (*RPS6KA3*, *RASGRF1*, *MAPK8IP1*, *SMAD3*, *DUSP3*, *MAPK9*, *SPTBN1*, *ACTN4*, *CAMK2G*, *TFE3*, *PRKAR1A*, *SNAP25*, *MAPK8IP2*, *MAPK8IP3*, *PRKAR1B*) and three genes with no DNM (*SYN1*, *MAPK1*, *PRKACA*). Among these nonsignificant genes are many previously identified NDD candidate genes (for example, *NRXN1*, *GRIN2A*, and *CAMK2A*)[44–46], suggesting this group as a potential target for future screening and disease gene discovery.

In addition to enrichment in the PPI context, we also noted functional cell-specific and tissue-specific enrichment analyses (CSEA and TSEA, respectively)[47,48] obtained primarily from mouse expression data sets. As expected, the 253 gene set is enriched for the brain expression with a bias toward early to mid-fetal gene expression in the cortex, striatum, and amygdala (Fig. 2e). Among these, the greatest specificity is observed for early to early mid-fetal cortical

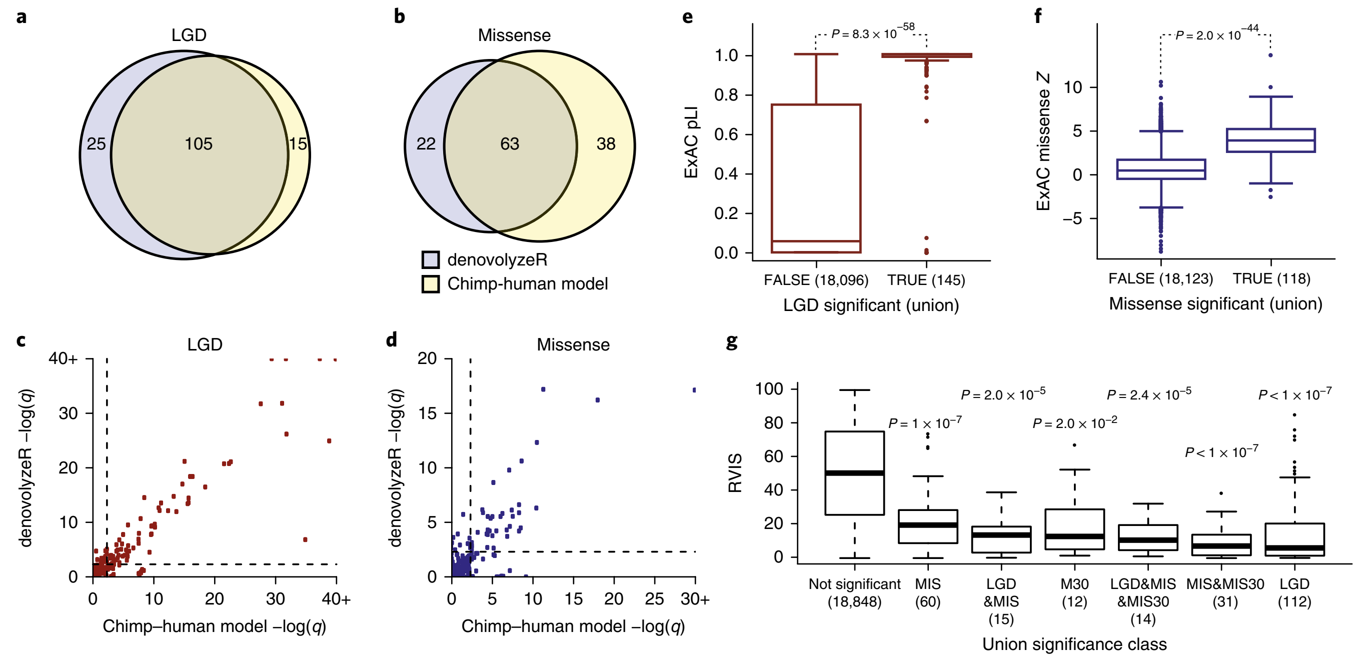

Fig. 1 | De novo–enriched genes and their characteristics. a–d, Results of applying both the chimpanzee–human (CH) divergence model and denovolyzeR to de novo variation in $n = 10{,}927$ independent individuals with ASD/ID/DD. The two models show considerable gene overlap (**a**,**b**) with correlated significance values (LGD Pearson $r^2 = 0.94$, missense $r^2 = 0.74$) (**c**,**d**). CH model LGD outliers include *NONO*, *MEIS2*, *LEO1*, *WDR26*, and *CAPRIN1*, and denovolyzeR LGD outliers include *ZBTB18* and *FAM200B* (**c**). CH model missense (MIS) outliers include *CAPN15*, *SNAPC5*, *DLX3*, *TMEM178A*, *ADAP1*, *SNX5*, *SMARCD1*, *WDR26*, and *AGO4*, and denovolyzeR missense outliers include *ITPR1*, *RAC1*, *SETD1B*, *WDFY4*, and *UNC80* (**d**). **e**, Recurrent mutated LGD genes (TRUE, $n = 145$ with pLI scores) are highly enriched for genes intolerant to mutation as defined by ExAC pLI score (two-tailed Wilcoxon rank-sum test). **f**, Genes significantly enriched for missense DNMs ($n = 118$ with missense Z scores) are outliers by the ExAC missense depletion Z scores (two-tailed Wilcoxon rank-sum test). **g**, Similarly, all subcategories of significant genes (n shown below each category name) are intolerant to mutation (RVIS percentile) compared to nonsignificant genes (Tukey HSD test; P values are corrected for all possible group comparisons). Boxplot edges represent quartiles 1–3; midlines indicate medians. Whiskers span from Q1 − 1.5× interquartile range (IQR) to Q3 + 1.5× IQR.

development. By CSEA, the de novo gene set shows enrichment in both classes of medium spiny neurons within the striatum (striatum D1$^+$ and D2$^+$ medium spiny neurons Benjamini–Hochberg (BH)-corrected $P = 0.013$ and $P = 0.011$, one-tailed Fisher's exact test) at a pSI (specificity index P value) threshold of 0.05. Additionally, we observe nominal significance for D1$^+$ and D2$^+$ spiny neurons (uncorrected $P = 0.027$ and $P = 0.023$, one-tailed Fisher's exact test) at a pSI of 0.01 (Supplementary Fig. 2a), and through a further analysis of available single-cell sequencing data, two pyramidal neuron subtypes (S1PyrL5, BH $P = 0.008$; one-tailed Fisher's exact test) and the hippocampus (CA1Pyr1, BH $P = 0.046$; one-tailed Fisher's exact test) (Supplementary Fig. 2b and Supplementary Note).

As a final analysis, we also assessed the expression patterns of the union gene set using human RNA-seq data sets. The gene set, irrespective of class of mutation, is significantly enriched for a pan-neuronal pattern of expression when compared to control sets selected based on synonymous DNM in cases and genes with recurrent DNM in controls samples (Fig. 3 and Supplementary Fig. 3). This analysis does not highlight specific cell types compared to control genes ($P = 0.52$, two-tailed corrected Wilcoxon rank-sum test) but does show slightly higher expression across neuronal cell types ($P = 0.0001$; two-tailed corrected Wilcoxon rank-sum test) even after controlling for gene length (Supplementary Fig. 3). The 253 union gene set shows a strikingly broad expression profile across adult human cortical neuron types, including GABAergic (inhibitory) and glutamatergic (excitatory) neurons, compared to control genes (Fig. 3a, Supplementary Fig. 3, and Supplementary Note).

Projected rates of gene discovery. Based on the number of genes that reach significance for DNM in our cohort of 10,927 cases, we estimate the potential yield by mutational class and the CH model. To this end, we subsampled smaller populations from our set 10,000 times each and tested for how many genes would reach significance using the CH model in a resampled cohort. We assessed logistic growth models for each mutation class and selected the best fitting model by Bayesian information criteria (BIC) to predict future performance. For genes with excess LGD DNMs, we observe what appears to be a rapid upcoming plateau in gene discovery, with an asymptote at 216 genes (95% CI 208–225) (ΔBIC linear model, asymptotic regression model = 259) (Fig. 4). Similarly, for genes with an excess of MIS30 DNMs, the model predicts an asymptote of only 65 genes (95% CI 63–67) (ΔBIC linear model, Weibull model = 250) (Fig. 4). By contrast, genes with an excess of recurrent missense mutations cannot yet be projected (Supplementary Note).

CNV intersection. In order to identify potentially dosage-sensitive genes underlying pathogenic CNVs, we intersected the 253-candidate gene set with a list of 58 genomic disorders based on previous CNV morbidity maps and the DECIPHER database (Fig. 5, Table 3 and Supplementary Table 2). Considering all genes with a de novo variant ($n = 6{,}886$), we find that 30 of 253 significant genes intersect a genomic disorder region, compared to 426 of 6,633 nonsignificant genes intersecting a disorder. This represents a significant enrichment ($P = 0.0017$, two-tailed Fisher's exact test; LR + 1.85 (95% CI 1.38–2.43)) compared to expectations supporting the notion that neurodevelopmental CNVs and DNMs converge on a common genetic etiology of gene-dosage imbalance. While we are underpowered to detect enrichment for any specific mutational and CNV class interactions by post hoc testing, as expected, LGD-significant genes and deletion disorders demonstrated the strongest

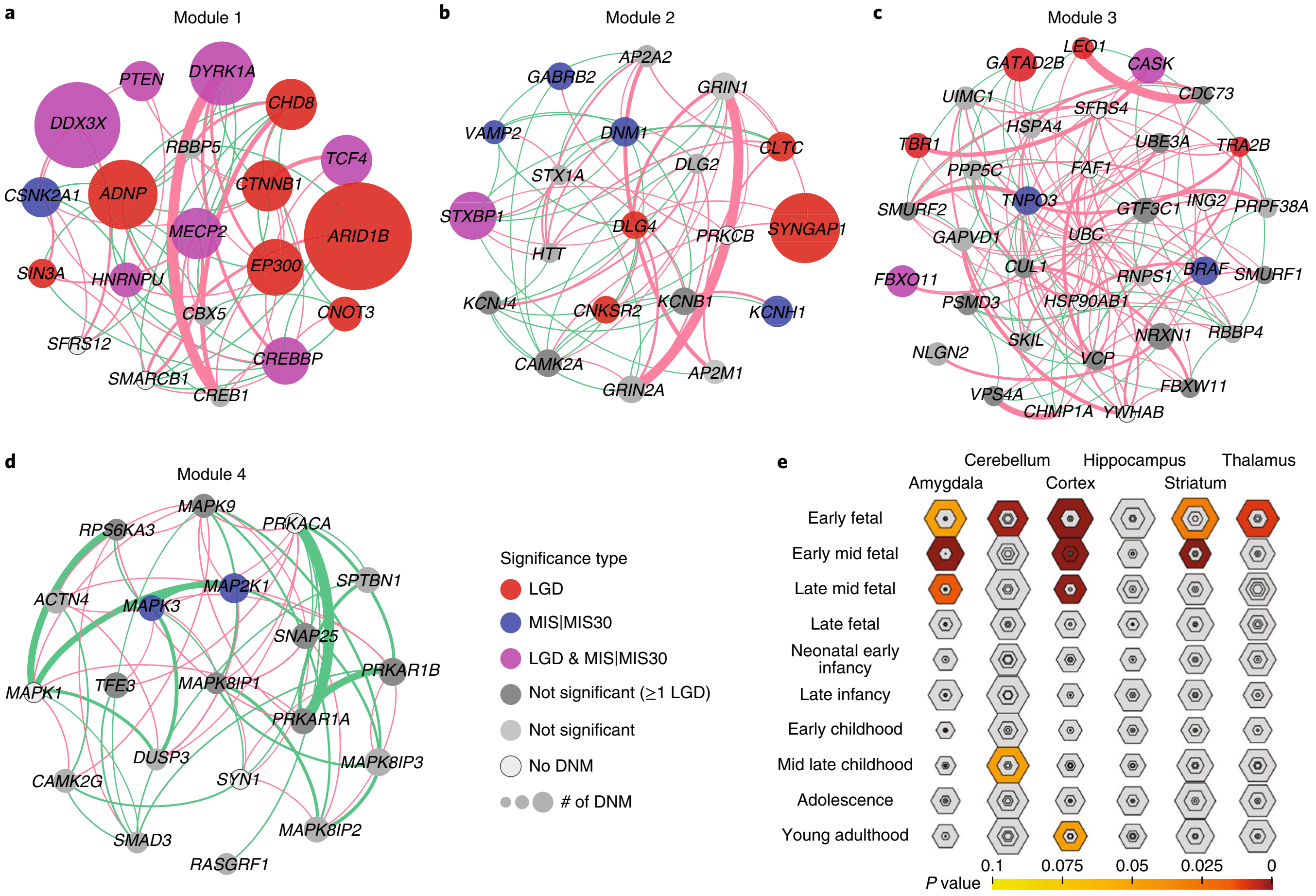

Fig. 2 | Gene expression and protein-interaction networks. a–d, MAGI[43] analysis of the union set ($n = 253$ independent genes) highlights the top four modules of coexpression and PPI, including genes significant for DNM enrichment by denovolyzeR (FDR-adjusted Poisson test) or the CH model (FDR-adjusted binomial test) (colored circles) and new candidate genes with DNM that do not yet reach significance (dark gray). The size of the circle represents the relative number of individuals with DNMs within this cohort. Edges depict PPIs (pink arcs) and coexpression (green arcs) scaled by their scores from geneMANIA[100]. **e**, Tissue-specific enrichment analyses (TSEA) of the union set ($n = 253$ independent genes) highlight a strong bias to various developing parts of the brain with the strongest signal in early to mid-fetal development (color corresponds to FDR-adjusted one-tailed Fisher's exact test *P* values; shaded regions closer to the center of each hexagon indicate increasing tissue specificity).

enrichment among the four combinations ($P = 0.1$, two-tailed Fisher's exact test, LR + 1.77 (95% CI 1.2–2.54)). Given the known complexity of gene regulation in CNVs[49], we highlight candidates representing all interaction types (Table 3 and Supplementary Table 2), with the expectation that the strongest candidates will correspond to a simple model of haploinsufficiency.

Many genomic disorders intersect with a single DNM-enriched gene, confirming a known CNV gene association, including *KANSL1* (Koolen–de Vries)[50], *SHANK3* (Phelan–McDermid)[51], *RAI1* (Smith–Magenis)[52], *NSD1* (Sotos)[53], *WHSC1* (Wolf–Hirschhorn)[54], *BCL11A* (2p15–16.1 microdeletion)[55], *EHMT1* (9q34 deletions/Kleefstra syndrome)[56], and *CREBBP* (Rubinstein–Taybi)[57] (Table 3 and Supplementary Table 2). Additionally, this analysis also highlights genes that have been implicated as candidates by case reports, functional studies, or smaller CNVs (Table 3 and Supplementary Table 2). Among these, we identify an excess of recurrent missense mutation in *MAPK3* mapping to the 16p11.2 microdeletion/microduplication region associated with ASD and ID[58]. Recurrent LGD mutations in *PHF21A*, a gene previously implicated by translocations and focal CNVs[59,60], map to the Potocki–Shaffer deletion region, and an excess of missense mutations in *KIF1A* correspond to the 2q37 deletion syndrome region[61,62]. Recurrent LGD DNMs in *SIN3A*, a REST and MECP2 interactor, map to the 15q24 deletion region[63–65]. *PPM1D* has been linked to ID[66] and is located in the 17q23.1q23.2 deletion region. *CLTC* has been linked to multiple malformations, and DD and is located in the 17q23 deletion region[67]. Genes enriched for recurrent missense DNM, *YWHAG*, and *GABRB3*, colocalize to the Williams–Beuren distal and Prader–Willi deletion/duplication regions, respectively[68–70].

Finally, we also consider as part of this analysis the 14 regions identified as significant for CNV burden[18] and identify five candidate intersections (Table 3). These include *SATB2* in the 2q33.1 region[18,71]; *MEF2C*, which demonstrated focal deletions, functions at cortical synapses and has been independently linked to hyperkinesis and epilepsy[72,73]; *CHD4* in the 12p13 duplication region, which has been linked by both a genome-wide association study (GWAS) and DNM to an ID syndrome[74,75]; and *WDFY4* in the 10q11.23 duplication region, which appears to be a novel finding at this time.

A second category of CNVs are those that intersect more than a single gene (Table 3 and Supplementary Table 2). The 2q33.1 region contains several potentially high-impact candidates, including *HECW2*, which has been linked to neurodevelopmental delay, ID and epilepsy by missense mutations[27,76]; *SATB2*, which has been independently identified by focal CNVs[18,71]; *ABI2*, which is a candidate for autosomal recessive ID[77]; and *SF3B1*, which interacts directly with the ID gene *PQBP1* (ref. [78]). Similarly, in the 2p15–p16.1 deletion region we identify both the primary gene *BCL11A*[55] as well as a second candidate gene in the minimal

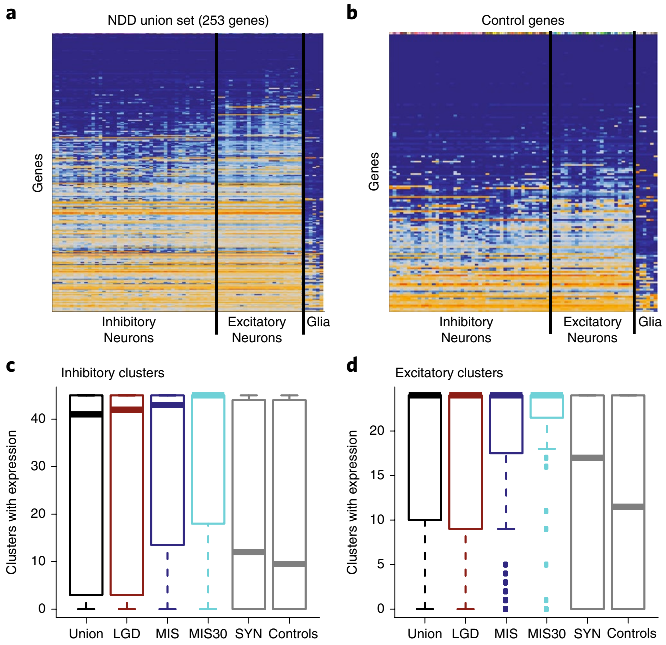

Fig. 3 | Expression in human cortical neurons. a,**b**, Heatmaps demonstrating a broad pattern of inhibitory and excitatory neuronal expression (median $\log_2$ (CPM + 1)) in the union gene set ($n = 253$ independent genes) compared to control genes ($n = 156$ independent genes). Expression level is indicated by a color gradient from low expression (dark blue) to high (orange). Rows represent individual genes and are ordered by the number of clusters (transcriptomic defined cell types) with expression (median CPM > 1), and columns represent 41 inhibitory neuronal, 24 excitatory neuronal, and 6 glial transcriptional clusters, each representing a distinct cell type. **c**,**d**, The number of inhibitory and excitatory clusters with expression in NDD genes (union $n = 253$, LGD $n = 145$, MIS $n = 123$, MIS30 $n = 59$) compared to controls (synonymous (SYN) $n = 101$, control $n = 156$ independent genes). The signal is strongest for NDD genes with the most severe missense mutations (MIS30). Boxplot edges represent quartiles 1– 3; midlines indicate medians. Whiskers span from Q1 – 1.5 IQR to Q3 + 1.5 IQR.

critical region, *PAPOLG*[79], which is enriched for severe missense DNM. Finally, in the 2q11.2q13 deletion region, we identified both *POU3F3*, which has been linked to ID and dysmorphic features by focal deletions[80], and *RFX8*, which has limited functional information in the literature.

Discussion

Exome sequencing of parent–child trios is a particularly powerful tool for the identification of genes, which when disrupted lead to pediatric NDD. The use of two DNM models, the CH model and denovolyzeR, identifies a high-confidence intersection ($n = 148$ genes) and a comprehensive union set (253 genes), which reach significance by one or both models. An advantage of using both models is that we identify high-risk candidate genes unique to each model ($n = 75$ additional genes), including several in which DNMs have already been associated with neurodevelopmental disease. Examination of this gene set in the context of the general population (that is, ExAC) confirms that this set list is enriched for genes that are constrained in the general population (LGD pLI $P = 8.3 \times 10^{-58}$, missense Z score $P = 2.0 \times 10^{-44}$, two-tailed Wilcoxon rank-sum test, RVIS $P < 1 \times 10^{-7}$ to 2.0×10^{-2}, Bonferroni adjusted Tukey HSD test) (Fig. 1).

While intolerance metrics are useful to enrich for pathogenic genes, our analysis suggests caution in strict application of a specific cutoff or even a single metric. Several known pathogenic genes are borderline by only one intolerance score, whereas several are poorly constrained by both metrics (for example, *MECP2* and Rett syndrome; RVIS = 32.4, pLI = 0.66). For example, we identified 20 genes that are intolerant to mutation by pLI but not RVIS

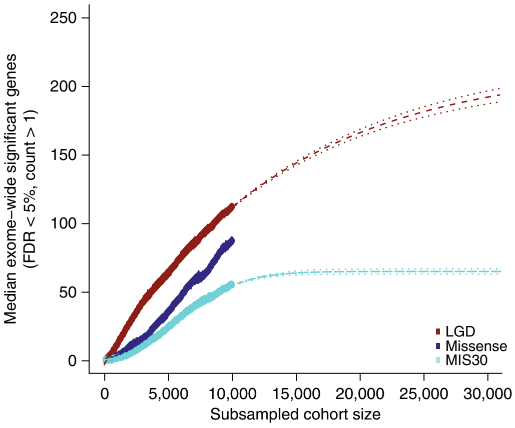

Fig. 4 | Estimation of gene discovery rates in future cohorts. We estimate the number of genes reaching significance under the CH model at varying population sizes subsampled from the total cohort of 10,927 individuals. Both the number of significant genes with recurrent LGD and MIS30 DNMs appear to be saturating with limited new gene discovery as sample sizes grow. De novo missense variants (including MIS30), however, as a more general class, demonstrate a more complex growth pattern with no best-fit line and thus are likely to represent the most important reservoir for new gene discovery as sequence data are generated from additional ASD and DD cohorts.

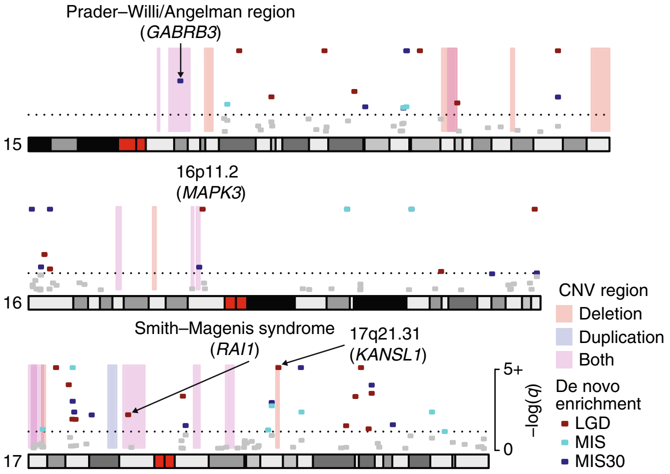

Fig. 5 | Integration of de novo SNVs and CNV morbidity map. Shown are examples of pathogenic CNVs (blue, red, and purple shading) associated with genomic disorders from chromosomes 15, 16, and 17, which intersect with genes that show a significant excess of DNM in $n = 10{,}927$ independent patients (red, turquoise, and blue points representing the minimum q value from either denovolyzeR or CH model; the dashed line represents a q value of 0.05). The analysis confirms known associations, such as *RAI1*, and *KANSL1* and candidate association for *MAPK3*. Recurrent severe missense mutations of *GABRB3* have been associated with autism and may be relevant to the recurrent 15q11 duplication. We note that mutations and deletions of the imprinted genes *SNRPN* (no DNM in our data set) and *UBE3A* (1 LGD and 1 missense DNM in our data set) are known to cause the core phenotype of Prader–Willi and Angelman syndromes, respectively, but do not reach significance in this analysis.

Table 3 | Intersection between pathogenic CNVs and recurrently mutated genes

		denovo-db v.1.5 counts									Union CH model and denovolyzeR	Genomic disorders			
		LGD de novo variants			Missense de novo variants			MIS30 de novo variants							
	Gene symbol*	All (*n* = 10,927)	ASD (*n* = 5,624)	ID/DD (*n* = 5,303)	All (*n* = 10,927)	ASD (*n* = 5,624)	ID/DD (*n* = 5,303)	All (*n* = 10,927)	ASD (*n* = 5,624)	ID/DD (*n* = 5,303)	Significance (FDR ≤ 5%, count > 1)	Deletion ayndrome	Duplication ayndrome	CNV significance type[18]	Decipher[a] classification
Significant in Morbidity Map[18] (37 disorders)	***MAPK3***	**0**	0	0	**3**	2	1	**1**	1	0	MIS	16p11.2-deletion	16p11.2-duplication	DEL and DUP	
	KANSL1	**8**	0	8	**1**	1	0	**0**	0	0	LGD	17q21.31-deletion	17q21.31-duplication	DEL	Category 1
	KIF1A	**0**	0	0	**11**	1	10	**9**	0	9	MIS	2q37-deletion	None	DEL	Category 1
	EHMT1	**9**	0	9	**3**	0	3	**2**	0	2	LGD	9q34-deletion	9q34-duplication	DEL and DUP	Category 1
	SHANK3	**10**	6	4	**1**	1	0	**0**	0	0	LGD	Phelan-McDermid-syndrome-deletion	None	DEL and DUP	Category 1
	PHF21A	**2**	1	1	**0**	0	0	**0**	0	0	LGD	Potocki–Shaffer-syndrome	None	DEL	Category 1
	GABRB3	**1**	1	0	**6**	2	4	**0**	0	0	MIS	Prader–Willi/Angelman	PWS-duplication	DEL and DUP	Category 1
	RAI1	**3**	1	2	**3**	2	1	**0**	0	0	LGD	Smith–Magenis-syndrome-deletion	Potocki–Lupski-syndrome-duplication	DEL and DUP	Category 1
	NSD1	**8**	1	7	**5**	2	3	**1**	0	1	LGD	Sotos-syndrome-deletion	None	DEL	Category 1
	WHSC1	**4**	1	3	**2**	1	1	**1**	0	1	LGD	Wolf–Hirschhorn-deletion	None	DEL and DUP	Category 1
21 Additional genomic disorders	***SIN3A***	**3**	0	3	**3**	2	1	**1**	0	1	LGD	15q24 deletion (A to E Inclusive)	None	Other SD pairs	
	CLTC	**4**	0	4	**2**	0	2	**1**	0	1	LGD	17q23 deletion	None		
	PPM1D	**8**	1	7	**0**	0	0	**0**	0	0	LGD	17q23.1q23.2 deletion	None		
	BCL11A	**6**	2	4	**3**	0	3	**0**	0	0	LGD	2p15–16.1 microdeletion syndrome	None		
	PAPOLG	**0**	0	0	**3**	3	0	**3**	3	0	MIS30	2p15–16.1 microdeletion syndrome	None		
	POU3F3	**2**	0	2	**2**	0	2	**2**	0	2	LGD	2q11.2q13 deletion	None		
	RFX8	**0**	0	0	**3**	2	1	**0**	0	0	MIS	2q11.2q13 deletion	None		
	HECW2	**1**	0	1	**9**	2	7	**7**	2	5	MIS30 and MIS	2q33.1	None		Category 1
	SATB2b	**9**	0	9	**6**	0	6	**5**	0	5	LGD, MIS30 and MIS	2q33.1	None		Category 1
	ABI2	**0**	0	0	**3**	2	1	**1**	1	0	MIS	2q33.1	None		Category 1
	SF3B1	**0**	0	0	**5**	3	2	**1**	0	1	MIS	2q33.1	None		Category 1
	CAPN15	**0**	0	0	**3**	0	3	**1**	0	1	MIS	ATR-16	None		Category 1
	SMC1A	**8**	0	8	**2**	0	2	**1**	0	1	LGD	None	Xp11.22-linked ID		
	HUWE1	**0**	0	0	**9**	0	9	**3**	0	3	MIS	None	Xp11.22-linked ID		
	WDR45	**8**	0	8	**2**	1	1	**0**	0	0	LGD	None	Xp11.22-p11.23 microduplication		
	MECP2	**11**	4	7	**7**	0	7	**4**	0	4	LGD, MIS30 and MIS	None	Xq28 (MECP2) duplication		
	CREBBP	**3**	0	3	**13**	3	10	**1**	1	0	LGD and MIS	Rubinstein–Taybi syndrome	None		Category 1
	SUV420H1	**7**	4	3	**3**	3	0	**1**	1	0	LGD	SHANK2 FGFs deletion	None		
	YWHAG	**0**	0	0	**3**	1	2	**1**	0	1	MIS	Wms-distal deletion	Wms-distal duplication		
	AUTS2	**4**	0	4	**1**	1	0	**0**	0	0	LGD	Wms-prox deletion	Wms-prox duplication		

Continued

Table 3 | Intersection between pathogenic CNVs and recurrently mutated genes (continued)

		denovo-db v.1.5 counts									Union CH model and denovolyzeR	Genomic disorders			
		LGD de novo variants			Missense de novo variants			MIS30 de novo variants							
	Gene symbol*	All (n = 10,927)	ASD (n = 5,624)	ID/DD (n = 5,303)	All (n = 10,927)	ASD (n = 5,624)	ID/DD (n = 5,303)	All (n = 10,927)	ASD (n = 5,624)	ID/DD (n = 5,303)	Significance (FDR ≤ 5%, count > 1)	Deletion ayndrome	Duplication ayndrome	CNV significance type[18]	Decipher[a] classification
14 Significant regions[18]	***SATB2*** [b]	**9**	0	9	**6**	0	6	**5**	0	5	LGD, MIS30 and MIS	2q33.1 (SATB2) deletion			
	MEF2C	**4**	0	4	**5**	1	4	**0**	0	0	LGD and MIS	5q14 (MEF2C) deletion			
	CHD4	**1**	0	1	**8**	1	7	**2**	0	2	MIS		12p13 (SCNN1A to PIANP)		
	WDFY4	**0**	0	0	**5**	4	1	**1**	0	1	MIS		1q11.23 duplication		

[a] DECIPHER syndrome classification system; see URLs. [b] *SATB2* is significant both by focal deletions and as part of the 2q33.1 region. DNM counts and significance categories from n = 10,927 independent samples are shown for genes from the union significance set (n = 253 genes denovolyzeR or CH model FDR < 5%) that intersect a previously established genomic disorder region. *Bolded gene symbols represent confirmation of known CNV associations. SD, segmental duplication; DEL, deletion; DUP, duplication.

(RVIS > 20). Some of these are well-established genes (for example, *KANSL1* and the Koolen–de Vries syndrome)[50], and the basis for this discrepancy is unknown but may relate to the fact that part of the gene is duplicated, complicating genome-wide analyses of intolerance. Among targets poorly constrained by both metrics, the Bohring–Opitz syndrome gene, *ASXL1*, was recently highlighted for the presence of somatic mosaic variants in the ExAC population (from which both the RVIS and pLI score are derived)[81].

Our projection estimates indicate that gene discovery based on recurrent LGD or severe missense mutations (MIS30) will soon plateau (Fig. 3). Clinical interpretation of patients with the only DNM in a previously unobserved gene will remain a challenge. Partitioning patients based on additional phenotypic criteria, subselecting genes based on functional pathway enrichment[43], integration of inherited variation (for example, TADA)[82] and targeting a small number of genes in much larger cohorts[10,11] are all potential strategies for associating specific genes with a phenotype under these conditions. For example, analysis of this set of DNMs in the context of MAGI modules further identified key functional categories, including neurotransmitter and synaptic signaling and JNK and MAPK cascades. This analysis identifies 46 genes with DNMs among the four modules that do not yet reach significance but are likely to represent functionally important targets for future screens.

In contrast to LGD and MIS30 DNM, the number of genes that will be identified by missense DNM generally has not yet begun to approach an asymptote. Samples sizes are just now beginning to reach the level at which signatures of recurrence and missense clustering are being detected for a relatively modest number of genes, most of which are only nominally significant[14,15]. We propose that this class of mutation (less severe or clustered missense DNMs) represents the most promising reservoir for future gene discovery but will require much larger whole-exome and whole-genome data sets to tease apart. As the number of exomes grows for both ASD and DD, the maintenance and curation of de novo databases will be especially important in this regard[20].

Previous studies have implicated larger CNVs and increased mutation burden with more-severe phenotypic outcomes, and here we observe that the majority of DNM in the 253 genes originates from ID/DD (>3:1, Supplementary Table 2). Importantly, DNM-enriched genes significantly overlap known pathogenic CNV regions (P = 0.0017, two-tailed Fisher's exact test, LR+1.85 (95% CI 1.38–2.43)), thus supporting a common genetic etiology. These specific targets have offered both independent confirmation of existing single-gene associations (for example, *KANSL1* in the 17q21.31 region), additional support for candidate genes (*MAPK3* in the 16p11.2 region), and further support for potentially oligogenic effects with multiple compelling candidate genes (*HECW2*, *SATB2*, *ABI2* and *SF3B1* in the 2q33.1 deletion region). This is consistent with findings suggesting a role for multiple hits in ASD[83,84]. Among the genes with recurrent mutation and CNV intersection, *MAPK3* is particularly interesting with respect to the chromosome 16p11.2 microduplication. Several functional studies on 16p11.2 deletion and duplication mice as well as *Drosophila* models have suggested that *MAPK3* is a key regulator of the syndrome being downstream of other ASD target genes, involved in axon targeting and regulation of cortical cytoarchitecture, and being the most topologically important gene in the region by PPIs[85–87]. Our analysis builds on these studies by providing evidence of recurrent missense mutation enrichment in human NDDs.

Finally, it is interesting that the 253 genes we highlight in this meta-analysis demonstrate a pan-neuronal expression pattern with the majority of genes being expressed in all GABAergic (inhibitory) and glutamatergic (excitatory) neuron types (Fig. 3), suggesting these genes have the potential to alter many paths in the adult cortical circuit. While the majority of genes are broadly expressed across neuronal cell types, a subset demonstrates evidence of specific expression. More specifically, we observed enrichment for genes specifically expressed in the D1$^+$ (19 genes) and D2$^+$ (18 genes) medium spiny neurons of the striatum (13 genes are shared in the D1$^+$ and D2$^+$ lists) in mouse brain (Supplementary Fig. 2a). Previously, Dougherty and colleagues highlighted this particular brain region based on a survey of genes reported as autism candidate risk genes[48]. Using a CSEA, we now extend this observation to NDD genes enriched for recurrent DNM. Remarkably, a similar enrichment was recently reported in autistic individuals with multiple DNMs in coding and putative noncoding regulatory DNA[84]. We also observe a similar signature for genes where nominal significance has been observed for clustered DNMs. While many of the genes enriched for D1$^+$ and D2$^+$ expression are not exclusive to the striatum and are more broadly expressed (as demonstrated by the enrichment signal at the lowest specificity threshold), the striatum has been implicated in ID and autism pathology by numerous studies[88–97]. The striatum is particularly compelling, as it has been linked to repetitive behaviors[88] that are core to the autism phenotype and to genes known to be involved in DD, including *CHD8*, *SHANK3*, *FOXP2*, and *KCNA4* (refs. [89,90,93,96,97]). While the striatum is most strongly linked to autism core phenotypes, our observation of enrichment in a more general DD cohort suggests that, while the general bias of cortex genes to ID and striatum genes to ASD[92] still holds, the diverse expression patterns of genes across the brain at complex developmental time points may have substantial functional overlap among subtypes of NDDs that will require deep phenotyping and imaging to tease apart.

In conclusion, the 253 genes that show evidence for recurrent DNM represent a starting point for further functional and phenotypic investigations. Of these genes, 124 reach a strict GWAS threshold of significance strongly arguing that DNM in these genes contributes significantly to disease. Overall, the genes we highlight demonstrate strong conservation, refine pathogenic CNVs, define distinct functional pathways, and support the role of striatal networks in the pathogenicity of both ASD and ID/DD. Strikingly, the majority of the genes identified in this study present with DNMs in both ASD and ID/DD. While we expect a degree of diagnostic overlap[21,22], our results support a common genetic etiology among broad neurodevelopmental phenotypes. These genes are candidates for a genotype-first paradigm[98], in which downstream follow-up of patients with the same de novo−disrupted gene is likely to provide additional insight into unique phenotypic features associated with these different genetic subtypes[99] and additional support for their role in NDD.

URLs. denovo-db, http://denovo-db.gs.washington.edu/; denovo-db v.1.5 documentation, http://denovo-db.gs.washington.edu/denovo-db.v.1.5.pdf; DECIPHER, https://decipher.sanger.ac.uk; Online Mendelian Inheritance in Man (OMIM), https://www.omim.org; SFARI Gene, https://gene.sfari.org/; ID Gene Database Project, http://gfuncpathdb.ucdenver.edu/iddrc/iddrc/home.php; Linnarsson Lab Single-cell analysis of mouse cortex, http://linnarssonlab.org/cortex; ea-utils fastqMCF program, https://expressionanalysis.github.io/ea-utils/; R package 'propagate', https://CRAN.R-project.org/package=propagate; DECIPHER Syndrome Overview https://decipher.sanger.ac.uk/disorders#syndromes/overview; Human MTG single nucleus RNA-seq data, http://celltypes.brain-map.org/download.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0288-4.

Received: 4 December 2017; Accepted: 23 October 2018;
Published online: 17 December 2018

References

1. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
2. Sharp, A. J. et al. Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
3. Tuzun, E. et al. Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
4. de Vries, B. B. et al. Diagnostic genome profiling in mental retardation. *Am. J. Hum. Genet.* **77**, 606–616 (2005).
5. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
6. de Vries, B. B. et al. Clinical studies on submicroscopic subtelomeric rearrangements: a checklist. *J. Med. Genet.* **38**, 145–150 (2001).
7. Firth, H. V. & Wright, C. F. The Deciphering Developmental Disorders (DDD) study. *Dev. Med. Child Neurol.* **53**, 702–703 (2011).
8. O'Roak, B. J. et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585–589 (2011).
9. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
10. Stessman, H. A. et al. Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat. Genet.* **49**, 515–526 (2017).
11. O'Roak, B. J. et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
12. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
13. Turner, T. N. et al. Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns. *Hum. Mol. Genet.* **24**, 5995–6002 (2015).
14. Geisheker, M. R. et al. Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nat. Neurosci.* **20**, 1043–1051 (2017).
15. Lelieveld, S. H. et al. Spatial clustering of de novo missense mutations identifies candidate neurodevelopmental disorder-associated genes. *Am. J. Hum. Genet.* **101**, 478–484 (2017).
16. Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
17. Kaminsky, E. B. et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet. Med.* **13**, 777–784 (2011).
18. Coe, B. P. et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014).
19. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
20. Turner, T. N. et al. denovo-db: a compendium of human de novo variants. *Nucleic Acids Res.* **45**, D804–D811 (2017).
21. Matson, J. L. & Shoemaker, M. Intellectual disability and its relationship to autism spectrum disorders. *Res. Dev. Disabil.* **30**, 1107–1114 (2009).
22. *American Psychiatric Association* Diagnostic and statistical manual of mental disorders, 5th edition: (DSM−5) (APA Publishing, Arlington, 2013).
23. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
24. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
25. de Ligt, J. et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
26. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
27. Halvardson, J. et al. Mutations in HECW2 are associated with intellectual disability and epilepsy. *J. Med. Genet.* **53**, 697–704 (2016).
28. Hashimoto, R. et al. Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. *J. Hum. Genet.* **61**, 199–206 (2016).
29. Krumm, N. et al. Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
30. Lee, H., Lin, M. C., Kornblum, H. I., Papazian, D. M. & Nelson, S. F. Exome sequencing identifies de novo gain of function missense mutation in KCND2 in identical twins with autism and seizures that slows potassium channel inactivation. *Hum. Mol. Genet.* **23**, 3481–3489 (2014).
31. Lelieveld, S. H. et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19**, 1194–1196 (2016).
32. Michaelson, J. J. et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431–1442 (2012).
33. Moreno-Ramos, O. A., Olivares, A. M., Haider, N. B., de Autismo, L. C. & Lattig, M. C. Whole-exome sequencing in a South American cohort links ALDH1A3, FOXN1 and retinoic acid regulation pathways to autism spectrum disorders. *PLoS ONE.* **10**, e0135927 (2015).
34. Rauch, A. et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
35. RK, C. Y. et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20**, 602–611 (2017).
36. Tavassoli, T. et al. De novo SCN2A splice site mutation in a boy with Autism spectrum disorder. *BMC Med. Genet.* **15**, 35 (2014).
37. Turner, T. N. et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am. J. Hum. Genet.* **98**, 58–74 (2016).
38. Yuen, R. K. et al. Genome-wide characteristics of de novo mutations in autism. *NPJ Genom. Med.* **1**, 160271–1602710 (2016).
39. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
40. Wang, T. et al. De novo genic mutations among a Chinese autism spectrum disorder cohort. *Nat. Commun.* **7**, 13316 (2016).
41. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
42. Le Meur, N. et al. MEF2C haploinsufficiency caused by either microdeletion of the 5q14.3 region or mutation is responsible for severe mental retardation with stereotypic movements, epilepsy and/or cerebral malformations. *J. Med. Genet.* **47**, 22–29 (2010).
43. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. *Genome Res.* **25**, 142–154 (2015).

44. Endele, S. et al. Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat. Genet.* **42**, 1021–1026 (2010).
45. Ching, M. S. et al. Deletions of NRXN1 (neurexin-1) predispose to a wide spectrum of developmental disorders. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153B**, 937–947 (2010).
46. Stephenson, J. R. et al. A novel human CAMK2A mutation disrupts dendritic morphology and synaptic transmission, and causes ASD-related behaviors. *J. Neurosci.* **37**, 2216–2233 (2017).
47. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Heintz, N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res.* **38**, 4218–4230 (2010).
48. Xu, X., Wells, A. B., O'Brien, D. R., Nehorai, A. & Dougherty, J. D. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *J. Neurosci.* **34**, 1420–1431 (2014).
49. Deshpande, A. & Weiss, L. A. Recurrent reciprocal copy number variants: Roles and rules in neurodevelopmental disorders. *Dev. Neurobiol.* **78**, 519–530 (2018).
50. Koolen, D. A. et al. The Koolen-de Vries syndrome: a phenotypic comparison of patients with a 17q21.31 microdeletion versus a KANSL1 sequence variant. *Eur. J. Hum. Genet.* **24**, 652–659 (2016).
51. Phelan, K. & Rogers, R. C. Phelan-McDermid Syndrome. in *GeneReviews(R)* (eds. Adam, M. P. et al.) (Seattle (WA), 1993).
52. Bi, W. et al. Mutations of RAI1, a PHD-containing protein, in nondeletion patients with Smith-Magenis syndrome. *Hum. Genet.* **115**, 515–524 (2004).
53. Han, J. Y. et al. Identification of a novel de novo nonsense mutation of the NSD1 gene in monozygotic twins discordant for Sotos syndrome. *Clin. Chim. Acta* **470**, 31–35 (2017).
54. Izumi, K. et al. Interstitial microdeletion of 4p16.3: contribution of WHSC1 haploinsufficiency to the pathogenesis of developmental delay in Wolf-Hirschhorn syndrome. *Am. J. Med. Genet. A* **152A**, 1028–1032 (2010).
55. Shimbo, H. et al. Haploinsufficiency of BCL11A associated with cerebellar abnormalities in 2p15p16.1 deletion syndrome. *Mol. Genet. Genomic Med.* **5**, 429–437 (2017).
56. Kleefstra, T. et al. Further clinical and molecular delineation of the 9q subtelomeric deletion syndrome supports a major contribution of EHMT1 haploinsufficiency to the core phenotype. *J. Med. Genet.* **46**, 598–606 (2009).
57. Fergelot, P. et al. Phenotype and genotype in 52 patients with Rubinstein-Taybi syndrome caused by EP300 mutations. *Am. J. Med. Genet. A.* **170**, 3069–3082 (2016).
58. Kumar, R. A. et al. Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17**, 628–638 (2008).
59. Labonne, J. D. et al. A microdeletion encompassing PHF21A in an individual with global developmental delay and craniofacial anomalies. *Am. J. Med. Genet. A* **167A**, 3011–3018 (2015).
60. McCool, C., Spinks-Franklin, A., Noroski, L. M. & Potocki, L. Potocki-Shaffer syndrome in a child without intellectual disability-The role of PHF21A in cognitive function. *Am. J. Med. Genet. A* **173**, 716–720 (2017).
61. Leroy, C. et al. The 2q37-deletion syndrome: an update of the clinical spectrum including overweight, brachydactyly and behavioural features in 14 new patients. *Eur. J. Hum. Genet.* **21**, 602–612 (2013).
62. Klebe, S. et al. KIF1A missense mutations in SPG30, an autosomal recessive spastic paraplegia: distinct phenotypes according to the nature of the mutations. *Eur. J. Hum. Genet.* **20**, 645–649 (2012).
63. Halder, D. et al. Suppression of Sin3A activity promotes differentiation of pluripotent cells into functional neurons. *Sci. Rep.* **7**, 44818 (2017).
64. Witteveen, J. S. et al. Haploinsufficiency of MeCP2-interacting transcriptional co-repressor SIN3A causes mild intellectual disability by affecting the development of cortical integrity. *Nat. Genet.* **48**, 877–887 (2016).
65. Amir, R. E. et al. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23**, 185–188 (1999).
66. Jansen, S. et al. De novo truncating mutations in the last and penultimate exons of PPM1D cause an intellectual disability syndrome. *Am. J. Hum. Genet.* **100**, 650–658 (2017).
67. DeMari, J. et al. CLTC as a clinically novel gene associated with multiple malformations and developmental delay. *Am. J. Med. Genet. A* **170A**, 958–966 (2016).
68. Fusco, C. et al. Smaller and larger deletions of the Williams Beuren syndrome region implicate genes involved in mild facial phenotype, epilepsy and autistic traits. *Eur. J. Hum. Genet.* **22**, 64–70 (2014).
69. Buxbaum, J. D. et al. Association between a GABRB3 polymorphism and autism. *Mol. Psychiatry* **7**, 311–316 (2002).
70. Guella, I. et al. De novo mutations in YWHAG cause early-onset epilepsy. *Am. J. Hum. Genet.* **101**, 300–310 (2017).
71. Asadollahi, R. et al. The clinical significance of small copy number variants in neurodevelopmental disorders. *J. Med. Genet.* **51**, 677–688 (2014).
72. Harrington, A. J. et al. MEF2C regulates cortical inhibitory and excitatory synapses and behaviors relevant to neurodevelopmental disorders. *eLife* **5**, e20059 (2016).
73. Paciorkowski, A. R. et al. MEF2C haploinsufficiency features consistent hyperkinesis, variable epilepsy, and has a role in dorsal and ventral neuronal developmental pathways. *Neurogenetics* **14**, 99–111 (2013).
74. Kohannim, O. et al. Discovery and replication of gene influences on brain structure using LASSO regression. *Front. Neurosci.* **6**, 115 (2012).
75. Weiss, K. et al. De novo mutations in CHD4, an ATP-dependent chromatin remodeler gene, cause an intellectual disability syndrome with distinctive dysmorphisms. *Am. J. Hum. Genet.* **99**, 934–941 (2016).
76. Berko, E. R. et al. De novo missense variants in HECW2 are associated with neurodevelopmental delay and hypotonia. *J. Med. Genet.* **54**, 84–86 (2017).
77. Harripaul, R. et al. Mapping autosomal recessive intellectual disability: combined microarray and exome sequencing identifies 26 novel candidate genes in 192 consanguineous families. *Mol. Psychiatry* **23**, 973–984 (2018).
78. Wang, Q., Moore, M. J., Adelmant, G., Marto, J. A. & Silver, P. A. PQBP1, a factor linked to intellectual disability, affects alternative splicing associated with neurite outgrowth. *Genes Dev.* **27**, 615–626 (2013).
79. Levy, J. et al. Molecular and clinical delineation of 2p15p16.1 microdeletion syndrome. *Am. J. Med. Genet. A* **173**, 2081–2087 (2017).
80. Dheedene, A., Maes, M., Vergult, S. & Menten, B. A de novo POU3F3 deletion in a boy with intellectual disability and dysmorphic features. *Mol. Syndromol.* **5**, 32–35 (2014).
81. Carlston, C. M. et al. Pathogenic ASXL1 somatic variants in reference databases complicate germline variant interpretation for Bohring-Opitz Syndrome. *Hum. Mutat.* **38**, 517–523 (2017).
82. He, X. et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* **9**, e1003671 (2013).
83. Werling D. M. et al. Limited contribution of rare, noncoding variation to autism spectrum disorder from sequencing of 2,076 genomes in quartet families. *BioRxiv* https://dx.doi.org/10.1101/127043 (2017).
84. Turner T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722.e12 (2017).
85. Park, S. M., Park, H. R. & Lee, J. H. MAPK3 at the autism-linked human 16p11.2 locus influences precise synaptic target selection at drosophila larval neuromuscular junctions. *Mol. Cells* **40**, 151–161 (2017).
86. Pucilowska, J. et al. 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 (2015).
87. Blizinsky, K. D. et al. Reversal of dendritic phenotypes in 16p11.2 microduplication mouse model neurons by pharmacological targeting of a network hub. *Proc. Natl Acad. Sci. USA* **113**, 8520–8525 (2016).
88. Langen, M. et al. Changes in the development of striatum are involved in repetitive behavior in autism. *Biol. Psychiatry* **76**, 405–411 (2014).
89. Platt, R. J. et al. Chd8 mutation leads to autistic-like behaviors and impaired striatal circuits. *Cell Rep.* **19**, 335–350 (2017).
90. Reim, D. et al. Proteomic analysis of post-synaptic density fractions from Shank3 Mutant mice reveals brain region specific changes relevant to autism spectrum disorder. *Front. Mol. Neurosci.* https://doi.org/10.3389/fnmol.2017.00026 (2017).
91. Balsters, J. H., Mantini, D. & Wenderoth, N. Connectivity-based parcellation reveals distinct cortico-striatal connectivity fingerprints in autism spectrum disorder. *Neuroimage* **170**, 412–423 (2018).
92. Shohat, S., Ben-David, E. & Shifman, S. Varying intolerance of gene pathways to mutational classes explain genetic convergence across neuropsychiatric disorders. *Cell Rep.* **18**, 2217–2227 (2017).
93. Kaya, N. et al. KCNA4 deficiency leads to a syndrome of abnormal striatum, congenital cataract and intellectual disability. *J. Med. Genet.* **53**, 786–792 (2016).
94. Flanigan, M. & LeClair, K. Shared motivational functions of ventral striatum D1 and D2 medium spiny neurons. *J. Neurosci.* **37**, 6177–6179 (2017).
95. Sanders, S. J. First glimpses of the neurobiology of autism spectrum disorder. *Curr. Opin. Genet. Dev.* **33**, 80–92 (2015).
96. Schreiweis, C. et al. Humanized Foxp2 accelerates learning by enhancing transitions from declarative to procedural performance. *Proc. Natl Acad. Sci. USA* **111**, 14253–14258 (2014).
97. Chen, Y. C. et al. Foxp2 controls synaptic wiring of corticostriatal circuits and vocal communication by opposing Mef2c. *Nat. Neurosci.* **19**, 1513–1522 (2016).
98. Stessman, H. A., Bernier, R. & Eichler, E. E. A genotype-first approach to defining the subtypes of a complex disease. *Cell* **156**, 872–877 (2014).
99. Bernier, R. et al. Disruptive CHD8 mutations define a subtype of autism early in development. *Cell* **158**, 263–276 (2014).
100. Warde-Farley, D. et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214–W220 (2010).

Acknowledgements

We wish to thank T. Turner and J. Shendure for helpful discussion and T. Brown for edits. This research was supported, in part, by the following: the Simons Foundation Autism Research Initiative (SFARI 303241) and US National Institutes of Health (NIH R01MH101221) to E.E.E. The J.D.D. laboratory is supported by a NARSAD Independent Investigator Award from the Brain and Behavior Research Foundation and NIH grant 5R01MH107515-03. We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). E.E.E. is supported by the Howard Hughes Medical Institute.

Author contributions

B.P.C. and E.E.E. designed the study. B.P.C. performed the primary statistical data analysis. B.P.C., H.A.F.S., and M.R.G. curated DNMs and performed enrichment analyses. A.S. assisted with statistical analyses and interpretation. R.A.B. performed phenotype analysis. T.E.B. and E.S.L. performed the human expression analysis. A.M.L. and J.D.D. performed CSEA on cortex and assisted with additional CSEA and TSEA. F.H. performed the gene network analysis. B.P.C. and E.E.E. wrote the manuscript. All authors have read and approved the final version of the manuscript.

Competing interests

E.E.E. is on the scientific advisory board of DNAnexus, Inc.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0288-4.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to E.E.E.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Data set. We analyzed de novo single-nucleotide and indel variants from whole-exome sequencing (WES) data generated for 10,927 cases with neurodevelopmental diagnoses of ASD or ID/DD compiled in the denovo-db v.1.5 release[20] (Supplementary Table 1). The subset of denovo-db v.1.5 cohorts used was specifically chosen to avoid potential sample overlap, as described in Geisheker et al[14]. and in the release documentation for denovo-db v.1.5 (see URLs). Briefly, we first assumed minimal overlap between studies based in Europe and America and studies with exclusion criteria, including participation in another study (for example, SSC and the Autism Sequencing Consortium or ASC). This was supported by screening for individuals with shared mutational sites and second events[14]. We excluded studies from The Autism Simplex Collection or TASC due to known sample overlaps; additionally, we utilized only the more recent MSSNG data set[35] to avoid redundant annotations. All variants were annotated to RefSeq transcripts using SnpEff and collapsing to the most severe variant across isoforms. Variants were further binned into LGD (stop loss/gain, splice and frameshift), missense or synonymous categories for analysis. While these data are derived from a diverse set of WES platforms with differing sensitivities, our DNM-based analysis assumed samples to have perfect sensitivity. The combined set of 12,172 DNMs includes 2,357 LGD and 9,815 missense mutations, representing the largest such analysis to date. Each source study reports validated sites or validation rates ranging from 88.2% to 100% (with the exception of the DDD study, which opted for a high-sensitivity approach) (Supplementary Table 1). Among these events, 1,106 LGD and 2,594 missense DNMs were validated and confirmed in the original studies, and the remainders have unknown validation status (invalidated sites are not included in this analysis). CNV region data were obtained from a previous study of 29,085 children with developmental disorders and 19,584 population controls[18], as well as a curated list of genomic disorders maintained by DECIPHER (v.9.18) (see URLs).

Statistics. All standard statistical tests not reported by a described application (Methods) were performed using the R statistical language (v3.2.4), and nonparametric tests were used whenever possible. Wilcoxon rank-sum tests were used for distribution comparisons, with the exception of the Tukey Honest Significant Difference test for RVIS testing across all categories and *t* test for IQ comparisons. Fisher's exact test was used for all count comparisons. Likelihood ratio tests for goodness of fit were performed using one million multinomial sampling simulations. Multiple testing correction was applied when appropriate using either the Benjamini–Hochberg FDR or Bonferroni FWER as described in the relevant sections.

Recurrent variant analysis. Enrichment of de novo LGD and missense variation per gene was calculated using two statistical models. The CH model, as previously described[11], was run using the default setting and assuming a baseline rate of 1.8 de novo variants per individual (Supplementary Table 1). We note that ID/DD cohorts tend to have higher DNM rates than ASD cohorts, potentially relating to a combination of both technical (chosen sensitivity, platform differences) and biological biases, with the DDD cohort[9] demonstrating the highest DNM rate among large cohorts[24]. We anticipated that false positives would be randomly distributed and not enriched within specific genes. Observed coding DNM (LGD or missense) rates in the exome studies ranged from 0.87 to 1.36 among the large cohorts (Supplementary Table 1). While these variances in DNM rate probably influence our results, all rates are below the 1.8 DNM per individual rate used in the CH model; thus, overall, our statistics are conservative. Additionally, we ran a recently published modified version[40] that separately tests for enrichment of variants with CADD v.1.3 scores[39] of 30 and higher, which are predicted to be the most damaging of missense variation. Similarly, denovolyzeR[12] (R package version v0.2.0) was run using default settings. Each test (LGD, missense, MIS30) was individually adjusted to a *q* value by the Benjamini–Hochberg procedure based on the number of genes in the model (exome wide) and genes with *q* value < 0.05 and a DNM count of two or more were considered for the union set. Wherever necessary, gene symbols were adjusted to match those used in the individual models (CH model, denovolyzeR, CSEA). In cases in which no model was generated for a gene of interest, 'no model' is indicated in the significance column of Supplementary Table 2. Each analysis was corrected genome-wide for the number of genes present in the corresponding models (18,946 for CH model and 19,618 for denovolyzeR).

Although the *q*-value threshold should control the FDR within a single list, the combination of two models has the potential to increase the upper bound of the FDR for both the union and intersection. Although we assume that the FDR should remain at 5% on average, the upper limits to the merged FDRs can be defined as follows:

$$FDR_{A\cup B} \leq \begin{cases} \frac{FDR_A \times |A| + FDR_B \times |B|}{|A\cup B|}, if \ |B \setminus A| \geq FDR_B \\ \quad \times |B| \ \ AND \ \ |A \setminus B| \geq FDR_A \times |A| \\ \frac{|B \setminus A| + FDR_A \times |A|}{|A\cup B|}, if \ |B \setminus A| < FDR \ _B \\ \quad \times |B| \ \ AND \ (|A \setminus B| \geq FDR \ _A \times |A| \ \ OR \ \ |B \setminus A| < |A \setminus B|) \\ \frac{|A \setminus B| + FDR_B \times |B|}{|A\cup B|}, if \ |A \setminus B| < FDR_A \\ \quad \times |A| \ \ AND \ (|B \setminus A| \geq FDR_B \times |B| \ \ OR \ \ |A \setminus B| \leq |B \setminus A|) \end{cases}$$

$$FDR_{A\cap B} \leq \left\{ \begin{array}{l} \frac{FDR_A{}^{*}|A|}{|A\cap B|}, if \ |A| \leq |B| \ \ AND \ |A \bigcap B| \geq FDR_A{}^{*}|A| \\ \frac{FDR_B{}^{*}|B|}{|A\cap B|}, if \ |B| \leq |A| \ \ AND \ |A \bigcap B| \geq FDR_B{}^{*}|B| \end{array} \right\}$$

Application of these upper bounds estimates the maximal error rates in the unions as LGD FDR < 8.59%, missense FDR < 7.56%. As MIS30 is only examined by one model, its estimated FDR remains at 5%. Similarly, we estimate the maximal error rates in the intersections as LGD FDR < 5.7%, missense FDR < 6.75%. We wish to stress that we anticipate these to be upper bounds. In addition to the described treatment, to increase the stringency of our union set, we excluded any gene with a single DNM from further consideration. As a result, we discarded *q* values for 1,183 LGD, 4,246 missense, and 1,083 MIS30 genes, of which 113 reach corrected significance but are primarily small genes. This represents elimination of ~5–15% of the genes further reducing the overall FDR. Despite their likely enrichment for false positives, 11 of 113 single-hit genes have been previously implicated in an NDD (Supplementary Table 2) and are thus potentially of interest for future studies with additional samples.

Identification of novel genes. To address novelty, we considered statistically significant genes as defined by five recent publications involving large-scale exome, whole-genome, or targeted sequencing (DeRubeis et al.[26], Sanders et al. [19], an in-press revised version of Stessman et al.[10], Yuen et al.[35], DDD 2017 (ref.[24])) and three well-curated databases (OMIM, SFARI Gene and ID Gene Database Project; see URLs; all database queries were made on 02/19/2018). We undertook the following approach to 'novelty' when comparing these data sets to our list of 253 significant genes. Any gene listed as significant (regardless of thresholds or alternative methods of significance (CNV integration, private variation, inherited variation)) in any of these data sets was not considered 'novel' in our data set. Additionally, we considered SFARI genes of any score category as known. For the 80/253 genes not 'known' by this initial screen, we followed up with literature searches for evidence of any link to case reports or human studies of ID, autism or mental retardation or DD. For 49/80 genes, no evidence was found through the PubMed search; this represents our most conservative set of 'novel' genes (Supplementary Table 3). While this pool is the most likely to contain false positive findings, 10 of these 49 genes were also significant under an FWER correction (*SMARCD1*, *SNX5*, *TNPO2*, *ADAP1*, *CAPN15*, *CHD3*, *TMEM178A*, *AGO4*, *SNAPC5* and *ANP32A*). Clinical evidence, even single case reports, adds credence to our gene list as high-confidence candidates for NDD pathology.

Network analysis. Identification of clustered gene modules was performed using the MAGI (merging affected genes into integrated networks)[43] enrichment tool with default settings, followed by visualization incorporating co-expression and physical interaction data from geneMANIA[100]. Significance for MAGI modules was performed by permuting DNM across genes according to their mutation rates in the CH model 100 times and enumerating the number of random modules with scores greater than or equal to the module of interest.

Expression analysis in mouse. Functional enrichment was examined using the CSEA and TSEA tools[48]. CSEA[48] was additionally applied to candidate genes using single-cell transcriptomic profiling data from mouse cortex and hippocampus[101] using custom R scripts. Raw mRNA count data were downloaded from the Linnarson lab (see URLs), and only genes with at least 25 total molecules and at least 1 molecule in ≥ 100 cells were retained. Molecule counts were then incremented by a pseudo-count of 0.125 and reads per kilobase of transcript, per million mapped reads (RPKM) normalized. RPKMs were averaged over each of the 47 cell subclasses identified by the authors using the BackSPIN clustering algorithm. Significantly enriched transcripts in each subclass were identified using the pSI package[47], with a minimum expression value of 3 RPKM and default settings otherwise.

Expression analysis in humans. 15,928 single nuclei were isolated from the middle temporal gyrus of adult post-mortem brains of three human donors and profiled with RNA-sequencing. Raw read (fastq) files were aligned to the GRCh38 human genome sequence (Genome Reference Consortium, 2011) with the RefSeq transcriptome version GRCh38.p2 (current as of 4/13/2015). For alignment, Illumina sequencing adapters were clipped from the reads using the fastqMCF program (ea-utils, see URLs). After clipping, the paired-end reads were mapped using Spliced Transcripts Alignment to a Reference (STAR)[102] with default settings. Unsupervised clustering identified 71 distinct transcriptomic clusters, including 41 GABAergic (inhibitory) neuronal, 24 glutamatergic (excitatory) neuronal, and 6 non-neuronal cell types (unpublished). For each gene, the expression pattern was characterized as the number of cell types with appreciable expression (median counts per million (CPM) > 1) in three broad classes: inhibitory and excitatory neurons and glia. Heatmaps were constructed of log-normalized expression ($\log_2$CPM + 1) of NDD risk genes and control genes across cell types. The number of inhibitory and excitatory neuronal and glial types that expressed NDD risk genes and control genes were quantified and visualized as empirical cumulative distributions. Distributions were compared with two-sided Wilcoxon rank-sum tests for each broad class of cell types, and *P* values were Bonferroni corrected for multiple testing. A cell type specificity or marker score (beta) was defined for all genes to measure how binary expression was among clusters, independent of the number of clusters labeled. First, the proportion (x) of samples in each cluster that expressed a gene above background level (CPM > 1) was calculated. Then, scores were defined as the squared differences in proportions between all pairs (i,j) of n clusters normalized by the sum of absolute differences plus a small constant (ε) to avoid division by zero. Scores ranged from 0 to 1, and a perfectly binary marker had a score equal to 1.

$$\beta = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i - x_j\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right| + \varepsilon}$$

Shapiro−Wilk tests rejected ($P < 0.05$) the null hypothesis that distributions were normally distributed of cell type counts for each broad class, maximum average expression, and marker scores. Therefore, distributions were compared with two-sided Wilcoxon rank-sum tests, and *P* values were Bonferroni-corrected for multiple testing.

Projected rates of gene discovery. Prediction of future LGD and missense variation discovery rates was determined by sampling (with replacement) populations of 100 to 10,900 cases 10,000 times each and calculating DNM statistics using the CH model. The number of genes with two or more mutations and a q value below 0.05 were then enumerated for each simulation, and linear as well as logistic growth models were fit to each curve, with the best model being chosen via BIC. Model fits and confidence bounds were performed using the base stats and propagate (see URLs) packages in the R statistical language.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All variant data in this study are available to download from denovo-db v.1.5 (http://denovo-db.gs.washington.edu/). Human MTG single-nucleus RNA-seq data and clusters can be downloaded from the Allen Institute for Brain Science website at http://celltypes.brain-map.org/download.

References

101. Zeisel, A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
102. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

nature research

Corresponding author(s): Eichler Evan E

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
□	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	All samples were obtained through the denovodb v. 1.5 web server (denovo-db.gs.washington.edu/)
Data analysis	We utilized two published algorithms (CH, and denovolyzer R package version 0.2.0) for detecting enrichment of de novo variation. Both have been previously described in cited manuscripts. All additional statistics were calculated using standard functions in the R statistical environment (v3.2.4).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All variant data in this study is available to download from denovo-db v.1.5 (see URLs). Human MTG single-nucleus RNA-seq data and clusters can be downloaded from the Allen Institute for Brain Science web site: http://celltypes.brain-map.org/download.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

The sample size represents all published samples with de novo variation in neurodevelopmental disorders that were available through denovo-db at the time of the study (v1.5). Sample size was chosen to be as large as possible to maximize power.

Data exclusions

We selected samples with diagnoses of autism or intellectual disability or developmental delay and associated unaffected individuals. Samples from epilepsy-only studies were not included as they did not match our target phenotypes of autism or intellectual disability or developmental delay

Replication

This study is based on a meta-analysis of genetic data and no experimental (biological) assays were used.

Randomization

This study represents a meta-analysis of published de novo variation in neurodevelopmental cases. All cases were assigned to either a bulk group of ASD or ID/DD based on the primary dataset's description.

Blinding

As this meta-analysis primarily represents a single group enrichment, test blinding was neither possible nor performed.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Unique biological materials
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology
☒	☐	Animals and other organisms
☐	☒	Human research participants

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

All studies present in this meta-analysis are described in their source manuscripts. Briefly, samples represent childhood genetic profiling of individuals with broad neurodevelopmental diagnoses of ASD (n=5,624), ID, or DD (ID/DD n= 5,303). Additional

	phenotypes and covariates were not consistently availible and thus not examined.
Recruitment	As this study is a meta-analysis no patient recruitment was performed.