nature genetics

SUPPLEMENTARY INFORMATION

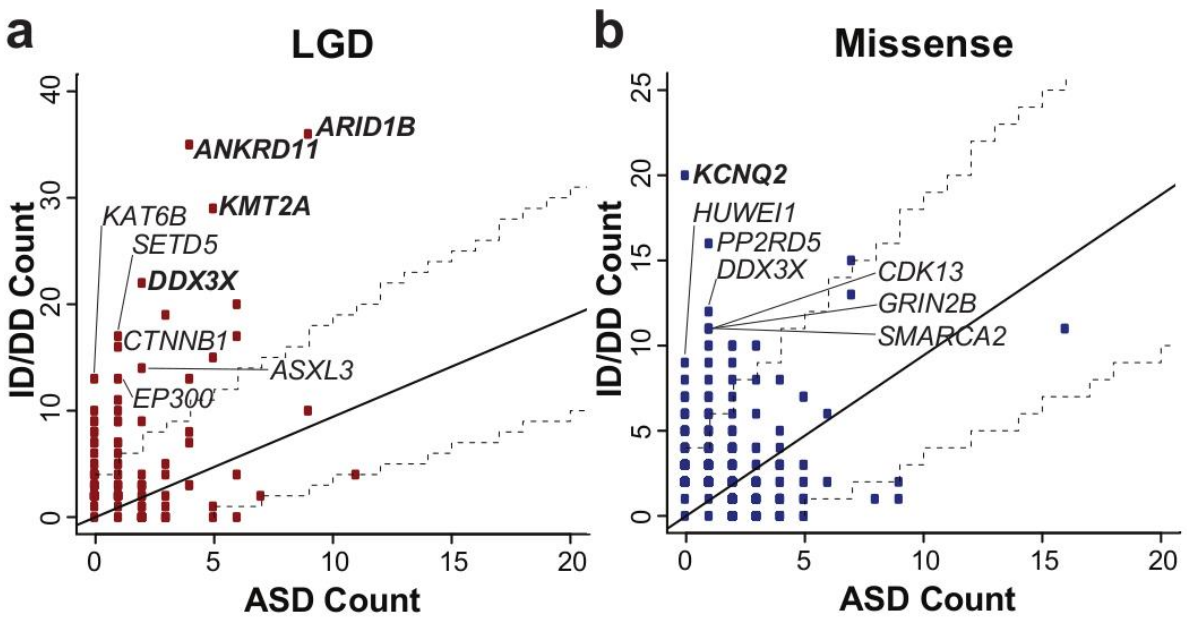
https://doi.org/10.1038/s41588-018-0288-4

In the format provided by the authors and unedited.

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]*

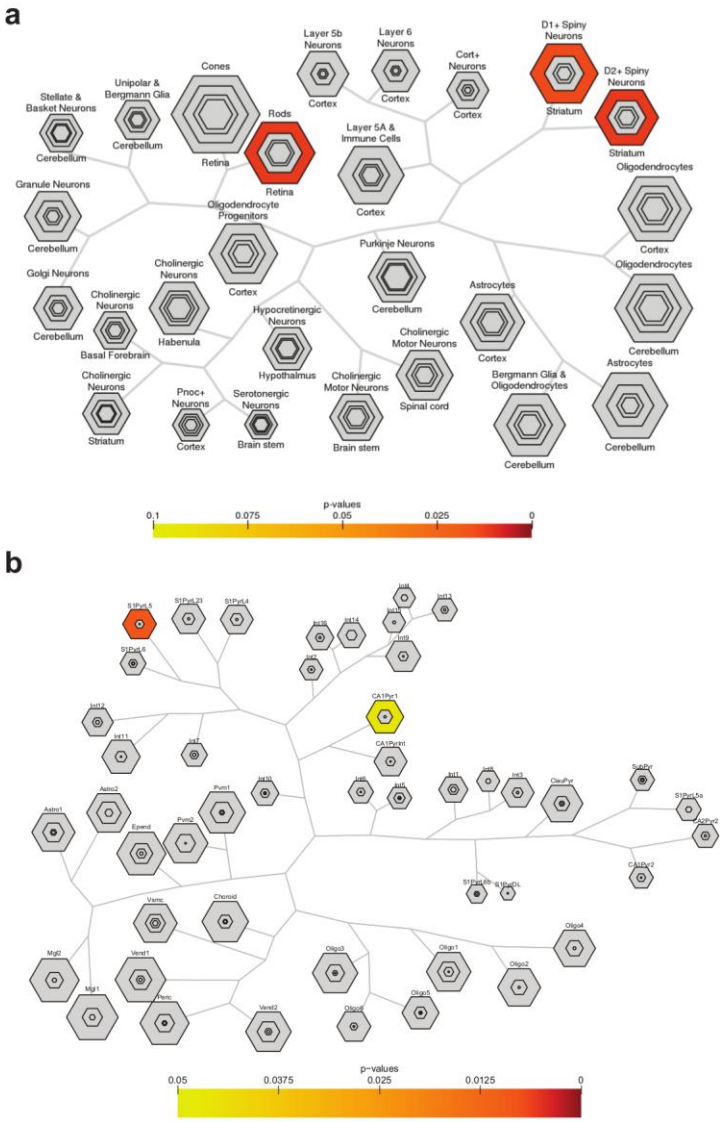
[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [2]Department of Pharmacology, Creighton University Medical School, Omaha, NE, USA. [3]Allen Institute for Brain Science, Seattle, WA, USA. [4]Department of Genetics, Department of Psychiatry, Washington 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



Supplementary Figure 1

Comparison of de novo variation rates in ASD and ID/DD.

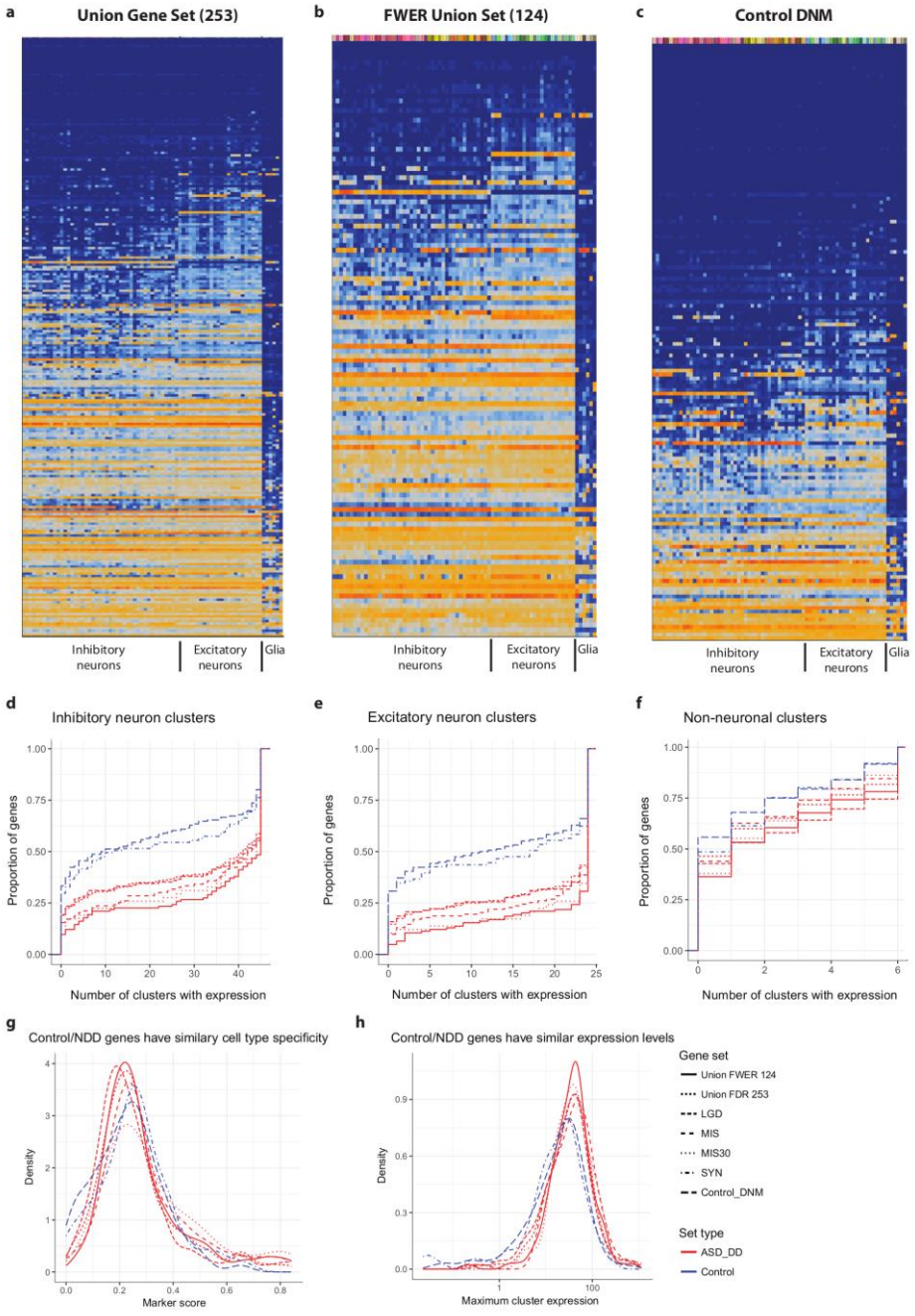
a,**b**, The plots compare DNM rates for genes for patients from ASD (*n* = 5,624 independent samples) and ID/DD (*n* = 5,303 independent samples) studies included in our combined analysis. More than 75% of genes show DNM in both ASD and DD patients. We identify four LGD genes (*ARID1B*, *ANDKRD11*, *KMT2A*, *DDX3X*) (**a**) and one missense gene (*KCNQ2*) (**b**) that are biased for an ID/DD diagnosis at a *q*-value threshold of 0.1 (one-tailed Fisher's exact test). Additional candidates for phenotypic bias at nominal significance (dashed lines at *P* = 0.05, one-tailed Fisher's exact test) were also identified. Larger cohorts will be needed to confirm gene biases, especially with respect to ASD.



Supplementary Figure 2

CSEA identifies bias to specific brain regions.

Cell-specific enrichment analyses (CSEA) of the union set (*n* = 253 independent genes) highlight a strong bias to various developing parts of the brain (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). **a**, We observe enrichment for both classes of striatal medium spiny neurons for our gene set. This tissue has been previously implicated in autism and candidate neurodevelopmental genes (*J. Neurosci* **34**, 1420–1431, 2014), and we now observe cell-specific enrichment among genes with a significant excess of DNM. **b**, Application of CSEA on *n* = 253 independent genes to the additional cell types profiled in Zeisel et al. (*Science* **347**, 1138–1142, 2015) identifies pyramidal neurons in layer 5 of the cortex and hippocampus. 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.



Supplementary Figure 3

Pan-neuronal expression patterns of candidate NDD genes.

a–**c**, Heatmaps demonstrating a broad pattern of inhibitory and excitatory neuronal expression (median $\log_2$ (CPM + 1)) in the NDD gene sets compared to control genes. The FWER union set shows even greater pan-neuronal-enriched expression than the larger union gene set. Rows represent individual genes and are ordered by the number of clusters with expression (median CPM > 1), and columns represent 41 inhibitory neuronal, 24 excitatory neuronal, and 6 glial clusters. **d**–**f**, Genes enriched for DNM are more broadly expressed in inhibitory (**d**) and excitatory (**e**) neurons, while genes enriched for LGD events specifically are enriched in glial expression (**f**). **g**, Comparison of control and test gene lists demonstrates similar maximum average expression (CPM) across cell types. **h**, Cell type specificity as measured by a beta marker score (Methods) is also similar for NDD and control genes.

Supplementary Note

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], Evan E. Eichler[1,7,*]

1) Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA

2) Department of Pharmacology, Creighton University Medical School, Omaha, NE 68178, USA

3) Allen Institute for Brain Science, Seattle, WA 98109, USA

4) Department of Genetics, Department of Psychiatry, Washington University School of Medicine, St. Louis, MO 63110, USA

5) Department of Biochemistry and Molecular Medicine, University of California, Davis, Davis, CA 95817, USA

6) Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA

7) Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

Model comparisons. While both models (CH, denovolyzeR) generally give similar results (Figure 1), there are outliers specific to each. With a few exceptions (Figure 1C-D, Table 1), genes unique to a model are near the boundary of statistical significance. The most substantial outliers are unique to the CH model (LGD outliers: *NONO*, *MEIS2*, *LEO1*, *WDR26*, *CAPRIN1*; missense outliers: *CAPN15*, *SNAPC5*, *DLX3*, *TMEM178A*, *ADAP1*, *SNX5*, *SMARCD1*, *WDR26*, *AGO*) and in all cases represent genes with no variation in the coding sequence between the chimpanzee and human reference sequences used to build the model.

Gene sets and multiple testing correction. The union set of 253 genes was defined based on a target FDR of 5% per model. We also considered the data in the context of the exome-wide Bonferroni family-wise error rate (FWER). Under this threshold, we corrected for five tests (LGD by two models, missense by two models, MIS30 by one model) for 20,000 genes ($p < 5 \times 10^{-7}$). We identify a union of 124 genes (Supplementary Table 2). While this is the traditional threshold applied to largely neutral genetic variation assessed by genome-wide association studies (GWAS)[1], this threshold is particularly strict because the prior probability of disease association for a DNM (especially damaging *de novo* LGD) differs greatly from that of a typical single-nucleotide polymorphism. As evidence, many known "disease" genes fall below this threshold (*RA1*, *FOXP2*, *PAX5*, *PHIP*, etc.) and we anticipate that a large fraction of these second-tier genes will meet a strict GWAS threshold of significance as sample sizes increase (Supplementary Table 2). Nevertheless, we provide to the reader both FDRs as well as p-values so that associated genes with strong suggestive evidence may be distinguished but chose to analyze the properties of the full 253 gene set as a group because these are significantly enriched for pathogenic genes.

Gene constraint and control DNM. A small number of genes are enriched for recurrent DNMs but do not demonstrate conservation by either RVIS or pLI/missense Z scores. For example, there are 11 genes significant for LGD recurrence with a pLI below 0.9 and RVIS percentile above 20 and these would be predicted by some to be enriched for false disease associations[2] (Supplementary Table 2). However, 64% (7/11) of these genes have an established association with ASD or ID/DD, including four X chromosome-linked genes (*MECP2*, *ZC4H2*, *UPF3B*, *HDAC8*)[3-6] and three autosomal genes (*ASXL1*, *PURA*, *PPM1D*)[7-9]. The remaining four genes lack support in the literature (*ENO3*, *HIVEP3*, *KCNS3*, *NFE2L3*) strongly linking them to ID/DD or ASD.
In addition, we should note that a subset of these genes show evidence of DNM in controls. Among the 2,278 controls in denovo-db v.1.5, 82.6% (209/253) of the union genes have no detected LGD or missense DNM in controls. Of the 44 genes with at least one control DNM, 17 have control DNMs matching the primary significance category, including 3 LGD DNMs in *KDM5B*, 4 missense DNMs in CHD4, and 15 genes with a

single control DNM (1 LGD in *KCNS3*, *TCF12*; 1 missense in *AGAP2*, *AGO4*, *CAPN15*, *CHD3*, *DYNC1H1*, *GLRA2*, *KCNH1*, *PBX1*, *SF3B1*, *WDFY4*, *TNPO3*; and 1 MIS30 in *CACNA1E*).

ASD vs. ID independent discovery sets and gene overlap. To eliminate potential ascertainment bias in discovery, we repeated the overlap analysis considering gene discovery independently in each cohort and then comparing its representation in the second (Supplementary Table 2). Because the ASD and DD cohorts treated independently represent substantially fewer individuals than the combined dataset, there is reduced power and, as a result, each disease-specific gene discovery set is necessarily smaller. We identified, for example, 183 genes that reached significance in the ID/DD cohort alone (union of the two models), and 41 genes reached significance in the ASD cohort—of which 19 genes are shared (205 total genes reach significance in the independent analyses). Among the genes significant by a DD-specific analysis, we observe 52.5% (96/183) of these as a DNM in autism cohorts. This overlap rises to 58.4% (80/137) if we restrict the analysis to genes detected only by the intersection of the two models. Conversely, when we examine ASD cohorts alone we observe larger degrees of overlap: 65.9% (27/41) of the union and 82.6% (19/23) of the intersection genes are observed in ID/DD providing strong support to the genetic etiological overlap between these two disorders.

Expression network analyses. We assessed the expression patterns of the union gene set using human brain RNA-seq data sets from single-cell nuclei that had been clustered into 71 transcriptomic clusters (Methods). 41% of NDD risk genes are expressed in all inhibitory neuron types and 57% in all excitatory types, while only 20% of control genes are as broadly expressed in interneurons and 34% in excitatory neurons (Figure 3C-D, Supplementary Figure 3) ($p = 1.1 \times 10\text{-}5$, inhibitory; $p = 1.1 \times 10\text{-}6$, excitatory; Wilcoxon rank-sum test). In contrast, we did not find evidence that NDD risk gene expression is enriched in glial cell types ($p = 0.13$). Focusing on 124 risk genes with LGD mutations, however, shows even more pronounced enrichment in inhibitory ($p = 2.7 \times 10\text{-}9$) and excitatory ($p = 5.1 \times 10\text{-}10$) neuron types as well as a moderate enrichment in glial types ($p = 0.0037$) (Figure 3C-D, Supplementary Figure 3).

Given the cortical development signal from TSEA, but lack of CSEA signal for specific cortical cell types, we also applied the CSEA approach to a more diverse single-cell sequencing data set from mouse cortex and hippocampus to identify cell-specific enrichments. This analysis further identifies two significant pyramidal neuron subtypes at a pSI of 0.05, including neurons in layer 5 (S1PyrL5, BH $p = 0.008$) and the hippocampus (CA1Pyr1, BH $p = 0.046$) (Supplementary Figure 2).

Phenotype analyses of autism DNM carriers. We compared both LGD and missense DNM groups on parent and clinician observation of ASD traits in the Simons Simplex Collection (SSC); namely, SRS Total T score, RBS-R Total score, ADOS Social Affect subscore, ADOS Restricted/Repetitive Behavior subscore. We

detected a nominal enrichment for increased severity of repetitive behavior (RBS-R [t(18) = 3.12, corrected p = 0.048]) among SSC probands carrying missense DNMs in ASD-biased genes suggesting the utility of larger cohorts with detailed quantitative phenotypes. Overall, detailed examination of phenotypes for ASD-biased genes in the SSC cohort detected no strong significant associations after correcting for multiple testing. While large-scale analyses, such as this, will likely require detailed quantitative phenotype data from many samples to truly tease apart genes that drive these phenotypes, future clinical follow-up of these targets on new patients will likely prove the most powerful approach to distinguish the phenotypic spectra associated with each of these targets.

Projected gene discovery. The missense data does not fit any logistic growth model and only poorly fits a linear model as the initial part of the curve demonstrates expected exponential growth. Further parsing the missense signal by examining missense mutations with CADD scores under 30 by a modified CH model also fails to generate a predictable asymptote suggesting that this category will represent the largest yield for novel genes as new exomes are sequenced. Alternatively, the missense signal may be too diffuse or diverse requiring new methods for identifying the functional subset of missense mutations.

REFERENCES

1. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. & Hirschhorn, J.N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33**, 177-82 (2003).
2. Kosmicki, J.A. *et al.* Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat Genet* **49**, 504-510 (2017).
3. Amir, R.E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* **23**, 185-8 (1999).
4. Hirata, H. *et al.* ZC4H2 mutations are associated with arthrogryposis multiplex congenita and intellectual disability through impairment of central and peripheral synaptic plasticity. *Am J Hum Genet* **92**, 681-95 (2013).
5. Jolly, L.A., Homan, C.C., Jacob, R., Barry, S. & Gecz, J. The UPF3B gene, implicated in intellectual disability, autism, ADHD and childhood onset schizophrenia regulates neural progenitor cell behaviour and neuronal outgrowth. *Hum Mol Genet* **22**, 4673-87 (2013).
6. Kaiser, F.J. *et al.* Loss-of-function HDAC8 mutations cause a phenotypic spectrum of Cornelia de Lange syndrome-like features, ocular hypertelorism, large fontanelle and X-linked inheritance. *Hum Mol Genet* **23**, 2888-900 (2014).
7. Hoischen, A. *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat Genet* **43**, 729-31 (2011).
8. Tanaka, A.J. *et al.* De novo mutations in PURA are associated with hypotonia and developmental delay. *Cold Spring Harb Mol Case Stud* **1**, a000356 (2015).
9. 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).