Neurodevelopmental disease genes implicated by *de novo* mutation and CNV morbidity

Bradley P. Coe[1], Holly A.F. Stessman[2], Arvis Sulovari[1], Madeleine Geisheker[1], Fereydoun Hormozdiari[3], Evan E. Eichler[1,4,*]

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

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

Corresponding author & *Lead Contact: Evan E. Eichler, Ph.D.
University of Washington School of Medicine
Howard Hughes Medical Institute
3720 15th Ave NE, Foege S413C
Box 355065
Seattle, WA 98195-5065
Phone: (206) 543-9526
E-mail: eee@gs.washington.edu

ABSTRACT

We combined *de novo* mutation (DNM) data from 10,927 cases of developmental delay and autism to identify 301 candidate neurodevelopmental disease genes showing an excess of missense and/or likely gene-disruptive (LGD) mutations. 164 genes were predicted by two different DNM models, including 116 genes with an excess of LGD mutations. Among the 301 genes, 76% show DNM in both autism and intellectual disability/developmental delay cohorts where they occur in 10.3% and 28.4% of the cases, respectively. Intersecting these results with copy number variation (CNV) morbidity data identifies a significant enrichment for the intersection of our gene set and genomic disorder regions (36/301, LR+ 2.53, p=0.0005). This analysis confirms many recurrent LGD genes and CNV deletion syndromes (e.g., *KANSL1*, *PAFAH1B1*, *RA1*, etc.), consistent with a model of haploinsufficiency. We also identify genes with an excess of missense DNMs overlapping deletion syndromes (e.g., *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. Finally, we also identify pathogenic CNVs overlapping more than one recurrently mutated gene (e.g., Sotos and Kleefstra syndromes) raising the possibility that multiple gene-dosage imbalances may contribute to phenotypic complexity of these disorders. Network analyses of genes showing an excess of DNMs confirm previous well-known enrichments but also highlight new functional networks, including cell-specific enrichments in the D1+ and D2+ spiny neurons of the striatum for both recurrently mutated genes and genes where missense mutations cluster.

INTRODUCTION

The importance of *de novo* mutations (DNMs) underlying neurodevelopmental disorders (NDDs) has been recognized for many years. Some of the strongest genome-wide evidence came from early copy number variation (CNV) studies, which consistently showed an excess of *de novo* as well as large private CNVs in patients with autism, developmental delay and epilepsy[1-4]. Because of the interspersed duplication architecture of the human genome[5], evidence of recurrent mutation in probands was often readily identified among a few hundred families—at least for those CNVs flanked by segmental duplications (SDs)[2] or mapping near telomeres[6]. Significance based on recurrence was more readily achieved from smaller sample sizes because of elevated CNV mutation rates in these specific regions of the genome, although in most cases the individual genes underlying the genomic disorders remained 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 developmental delay and autism[7,8]. The locus heterogeneity of these pediatric disorders, and the more random nature of SNV mutations, however, has made it more difficult for individual studies to provide statistical significance for DNM recurrence[9,10]. Moreover, different statistical models for discovery of genes with potentially pathogenic DNMs have been developed, including those based on chimpanzee–human divergence[11], trinucleotide mutation context[12], and clustering of DNMs[13-15]. Finally, despite extensive CNV analyses of nearly 45,000 patients with autism and developmental delay[16-18], few attempts have been made[18] to integrate the wealth of CNV data with recent exome sequencing results despite the fact that the same model of dosage imbalance might underlie a common genetic etiology for a specific locus (i.e., likely gene-disruptive (LGD) mutations and CNV deletions).

In this study, we perform an integrated meta-analysis of DNMs that have been recently published and compiled based on exome sequencing data of 10,927 individuals with a diagnosis of autism spectrum disorder (ASD) or intellectual disability and/or developmental delay (ID/DD)[19]. We overlay these data with known genomic disorders and a CNV morbidity map generated from 29,085 cases and 19,584 controls and jointly consider data from parent–child trios from ASD and DD cohorts in order to increase statistical power. This treatment of the data is justified because of the significant comorbidity between these two disorders[20,21] and because autism cases with a severe DNM are enriched in DD[22]. We compare two statistical models of recurrence as well as mutational classes because of potentially different functional consequences (e.g., recurrent missense vs. LGD DNMs). 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 database denovo-db v.1.5 release[19]. 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[11,22-37]. We considered all protein-altering and LGD mutations, including

3

frameshifts, splice donor/acceptor mutations, start losses and stop gains. The combined set of 12,172 DNMs included 2,357 LGD and 9,815 missense mutations.

We initially applied two statistical models. The first incorporates locus-specific transition/transversion/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 CH model. The second model, denovolyzeR, estimates mutation rates based using trinucleotide context and incorporates exome depth and divergence adjustments based on macaque–human comparisons over a +/-1 Mbp 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 and missense probabilities, we recently modified the CH model to incorporate CADD scores[38,39] allowing us to also specifically test for enrichment for the most severe 0.1% missense mutations (i.e., 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 to controls[14]. 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). Based on prior assumptions that 50% of DNMs will likely be disease related, we have chosen a q-value cutoff of 0.1 for both models to represent a false discovery rate (FDR) near ~5%[18,40]. We note that a more stringent q-value threshold of 0.05 slightly reduces the number of significant genes. Only 16% (24/147) of the genes which show an excess of de novo missense mutations demonstrate a q-value between 0.05 and 0.1 (four of which are significant by LGD or MIS30 DNM).

Combined, the two models (union set) identified 301 candidate NDD genes with evidence of excess DNM with a q-value < 0.1, and at least two mutations for at least one mutational category (Table 1, Table 2, Supplemental Table 1). This includes 163 genes with excess LGD mutations and 147 genes with excess missense mutations. Among these, 31 genes demonstrate evidence of both LGD and missense mutations (Figure 1, Table 1, and Supplemental Table 1). In general, both models highlight similar genes (Figure 1), particularly for LGD events where 71.2% (116/163) of genes are shared. This stands in contrast to recurrent missense DNMs where only 47.6% (70/147) of the genes overlap between the models. 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.

Since it has been well-established that NDD genes are less tolerant to mutation, we categorized the 301 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 observe a significant increase in pLI scores compared to all other genes, which is almost predictive (p = 2.7e-62, ROC AUC = 0.89) (Figure 1E). We also observed a significant increase in missense constraint (missense Z scores) among genes with enrichment for missense variation (p = 1.3e-51, ROC AUC = 0.84) (Figure 1F). Using RVIS as a constraint metric demonstrates a similar trend. We observe a significant RVIS depletion for all categories where at least

two genes were identified (Figure 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 = 2.4e-3, missense Z score = 4.47).

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 14 genes significant for LGD recurrence with a pLI below 0.9 and RVIS percentile above 20 and these would be predicted to be enriched for false disease associations[43]. However, half (7/14) of these genes have an established association with ASD or ID/DD, including four X chromosome-linked genes (*MECP2*, *ZC4H2*, *UPF3B*, *HDAC8*)[44-47] and three autosomal genes (*ASXL1*, *PURA*, *PPM1D*)[48-50]. The remaining seven genes lack support in the literature (*ENO3*, *HIVEP3*, *C9orf142*, *KCNS3*, *NFE2L3*, *SALL3*, *GPRIN1*) strongly linking them to ID/DD or ASD. In addition we should note that a subset of genes show evidence of DNM in controls. Among the 2,278 controls in denovo-db v1.5 82.7% (249/301) of the union genes have no detected LGD or Missense DNM in controls. Among the 52 genes with at least 1control DNM, 14 genes have control DNM matching the primary significance category including 3 LGD in *KDM5B* and 13 genes with a single control DNM (1 LGD in *KCNS3*, *TCF12*; 1 Missense in *AGAP2*, *AGO4*, *CAPN15*, *GLRA2*, *GRIN1*, *PBX1*, *SF3B1*, *WDFY4*, *TNPO3*; and 1 MIS30 in *CACNA1E*, *MEF2D*). The detection of control events may represent incomplete penetrance, variable expressivity, undiagnosed/subclinical controls, or benign variation (primarily in the case of missense variation). None of the recurrent control DNM genes reach exome wide FDR significance. While some of these are plausible candidates, disease significance should be considered with caution until additional functional and clinical data establish their role.

ASD versus DD genes: Because the number of samples ascertained for ASD and DD are approximately equal, we examined the distribution of LGD and missense mutations between the two populations. The majority of the genes—76% (229/301) of the union and 84% (137/164) of the intersection—show evidence of DNM in both ASD and ID/DD cohorts. A small number of genes are specific or enriched. For example, we directly compared the variant distributions among the genes with DNM and identified four genes (*ANKRD11*, *KMT2A*, *ARID1B*, *DDX3X*) with recurrent LGD and one gene (*KCNQ2*) with recurrent missense mutations as being strongly biased toward a diagnosis of ID/DD (q < 0.1). Although no genes are statistically enriched for ASD (Figure 2), several genes (e.g., *WDFY3*, nominal p = 0.019; *DSCAM*, nominal p = 0.036; *CHD8*, nominal p = 0.074) tend toward ASD diagnosis. While larger sample sizes are still required, these results highlight the significant comorbidity of ID/DD and ASD but also suggest that there will be genes that associate preferentially with ID or ASD. Considering all 301 genes and the full set of NDD patients, we calculate that 19.1% (2,085/10,927) of 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 (28.4% or 1,507/5,303 patients) when compared to ASD (10.3% or 578/5,624 patients) (OR = 3.47, p = 1.22e-131).

Network enrichment: Examination of the 301 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 (1,283 edges vs. 699 expected) enrichment in interactions among the 301 union genes (p < 2.2e-16). Given this high level of interconnectivity, we applied MAGI[51], a gene network discovery tool, to identify potential gene clusters, functional enrichments, and additional candidate interactions. Here, we present the top four networks from this analysis and their associated PANTHER functional enrichments (Figure 3A-D). Module 1 (20 genes)

highlights "regulation of transcription from RNA polymerase II promoter" (p = 0.0269) (Figure 3A) and contains 15 significant genes in addition to three additional 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) and synaptic signaling (p = 8.91e-5) (Figure 3B) and contains 11 significant genes in addition to eight genes that do not reach significance (*DLG2*, *HTT*, *AP2A2*, *KCNJ4*, *KCNB1*, *STX1A*, *GRIN2A*, *CAMK2A*) and one gene with no DNM (*PRKCB*). Module 3 highlights the "transmembrane receptor protein serine/threonine kinase signaling pathway" (p = 0.002) (Figure 3C) and contains nine significant genes in addition to 20 genes that do not reach significance (*SMURF2*, *SMURF1*, *CDC73*, *RNPS1*, *RBBP4*, *UBE3A*, *CUL1*, *FBXW11*, *VCP*, *VPS4A*, *PPP5C*, *PRPF38A*, *SKIL*, *HSPA4*, *PSMD3*, *UIMC1*, *GAPVD1*, *NLGN2*, *GTF3C1*, *NRXN1*), and six genes with no DMN (*ING2*, *SRSF4*, *FAF1*, *UBC*, *HSP90AB1*, *YWHAB*). Finally, Module 4 highlights c-Jun N-terminal kinase (JNK) (p = 4.65e-5) and Mitogen-activated protein kinase (MAPK) (p = 3.13e-6) cascades (Figure 3D) 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 (e.g., *NRXN1*, *GRIN2A*, *CAMK2A*)[52-54] suggesting this group as a potential target for future screening and disease gene discovery.

In addition to enrichment in the PPI context, we also note strong functional cell-specific and tissue-specific enrichment analyses (CSEA and TSEA)[55,56]. As expected, the 301 gene set is enriched for the brain expression with a bias toward early to mid-fetal gene expression in the cortex, striatum and amygdala (Figure 3E). Among these the greatest specificity is observed for early to early mid-fetal cortical 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 BH corrected p = 0.043 and p = 0.02) at a pSI (specificity index p-value) threshold of 0.05. Additionally, we observe nominal significance for D1+ and D2+ spiny neurons (uncorrected p = 0.018 and p = 0.015) at a pSI of 0.01 (Figure 3F). This cell-specific enrichment appears to be driven by genes with an excess of LGD DNMs (n = 163) (BH p = 0.029 for D1+ and D2+ neurons at pSI = 0.05, nominal p = 0.009 and 0.008 at pSI = 0.01), but not among genes enriched for recurrent missense mutations.

Clustered missense: As a final form of candidate gene annotation, we examined the distribution of missense DNMs for potential clustering within the predicted protein as previously described[14]. We identified a set of 183 nominally significant genes (using ExAC as a reference). Interestingly, 78% of this gene set does not overlap genes that reach *de novo* significance for missense recurrence (Supplemental Table 1). Examination of this clustered missense set by CSEA once again highlights the D1+ and D2+ spiny neurons of the striatum (nominal p = 0.066 and p = 0.005, respectively, at pSI = 0.05). Several D1+ genes were identified as significant for recurrent missense and clustered mutations (*PPP2R5D*, *SMARCD1*, *GRIN2B*). Among those showing evidence of clustered DNM exclusively, we identify genes enriched for D1+ (*PPP2R5D ARID1A*, *CEP131*, *INF2*, *RYR3*, *PHACTR1*, *KDM3A*, *GRIN2B*, *SLITRK5*) and D2+ (*ARID1A*, *CEP131*, *INF2*, *RYR3*, *AFF3*, *MAPKBP1*, *PHACTR1*, *KDM3A*, *MYO18A*, *SMAD4*, *SLITRK5*) spiny neuron expression. Mutations in these genes should be considered as candidates for further follow-up since not all genes reach exome-wide statistical significance. However, 99 genes remain significant after multiple testing corrections at a q-value threshold of 0.05. This set is significant for gene

enrichment only in D2+ (*PRKD1*, *ARID1A*, *INF2*, *RYR3*, *AFF3*, *MAPKBP1*, *SLITRK5*, *KDM3A*, *MYO18A*) (BH p = 0.002) but not in D1+ (BH p = 0.112) spiny neurons.

Projected rates of gene discovery: Based on the number of genes that reach significance for DNM in our cohort of 10,927 cases, we estimated 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 of similar phenotypic composition (i.e., an NDD cohort composed of a nearly equal number of patients with autism and ID/DD). 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 213 genes (95% CI 203-225) (ΔBIC linear model – Weibull model = 402) (Figure 4). Similarly, for genes with an excess of MIS30 DNMs the model predicts an asymptote of only 100 genes (95% CI 95-105) (ΔBIC linear model – Weibull model = 235) (Figure 4). By contrast, genes with an excess of recurrent missense mutations cannot yet be projected. The 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.

CNV intersection: In order to identify potentially dosage-sensitive genes underlying pathogenic CNVs, we intersected the 301 candidate gene set with a list of 58 genomic disorders based on previous CNV morbidity maps and the DECIPHER database (Table 3, Figure 5). Considering all genes with a *de novo* variant (n = 6,886), we find that 36 of 301 significant genes intersect a genomic disorder region. This represents a significant (p = 0.0005) enrichment (LR+ 2.53 [95% CI 1.43 – 1.91]) compared to expectations supporting the notion that CNVs and DNMs converge on a common genetic etiology of gene-dosage imbalance.

Many genomic disorders intersect with a single DNM-enriched gene, confirming a known CNV gene association, including *KANSL1* (Koolen-de Vries)[57], *PAFAH1B1* aka *LIS1* (Miller-Dieker)[58], *SHANK3* (Phelan-McDermid)[59], *RAI1* (Smith-Magenis)[60], *NSD1* (Sotos)[61], *WHSC1* (Wolf-Hirschhorn)[62], *BCL11A* (2p15-16.1 microdeletion)[63], and *CREBBP* (Rubinstein-Taybi)[64] (Table 3). In addition, this analysis also highlights genes that have been implicated as candidates by case reports, functional studies, or smaller CNVs (Table 3). Among these, we identify an excess of recurrent missense mutation in *MAPK3* mapping to the 16p11.2 microdeletion/microduplication region associated with autism and ID[65]. Recurrent LGD mutations in *PHF21A*, a gene previously implicated by translocations and focal CNVs[66,67], map to the Potocki-Shaffer deletion region, while an excess of missense mutations in *KIF1A* correspond to the 2q37 deletion syndrome region[68,69]. Recurrent LGD DNMs in *SIN3A*, a REST and MECP2 interactor, map to the 15q24 deletion region[44,70,71]. *CLTC* has been linked to multiple malformations and DD and is located in the 17q23 deletion region[72]. Genes enriched for recurrent missense DNM, *YWHAG* and *GABRB3*, co-localize to the Williams-Beuren and Prader-Willi deletion/duplication regions, respectively[73-75].

Finally, we also considered as part of this analysis the 14 regions identified as significant for CNV burden[18] and identified five candidate intersections (Table 3). These include *SATB2* in the 2q33.1 region [18,76]; *MEF2C*, which demonstrated focal deletions, functions at cortical synapses and has been

7

independently linked to hyperkinesis and epilepsy[77,78]; *CHD4* in the 12p13 duplication region, which has been linked by both a genome-wide association study and DNM to an ID syndrome[79,80]; *WDFY4* in the 10q11.23 duplication region, which appears to be a novel finding at this time; and, finally, *PAX5* in the 9p13 region, which was previously highlighted in O'Roak et al.[81].

A second category of CNVs are those that intersect more than a single gene (Table 3). The large 9q34 deletion region, for example, contains two genes with recurrent LGD mutations: the expected Kleefstra syndrome gene *EHMT1*[82] and a relatively unknown gene, *C9orf142*, implicated in DNA repair (see above)[83]. The region also harbors *GRIN1*, where recurrent missense mutations have been identified in DD[84]. 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[26,85]; *SATB2*, which has been independently identified by focal CNVs[18,76]; *ABI2*, which is a candidate for autosomal recessive ID[86]; and *SF3B1*, which interacts directly with the ID gene *PQBP1*[87]. In the Sotos syndrome region we identify not only the critical gene *NSD1*[61] but also significance in *GPRIN1* by LGD enrichment (denovolyzeR model only). Similarly, in the 2p15-p16.1 deletion region we identify both the primary gene *BCL11A*[63] as well as a second candidate gene in the minimal critical region *PAPOLG*[88], which is enriched for severe missense DNM. In the 17q23.1q23.2 deletion we identify enrichment in *PPM1D*, which has been directly linked to ID[50], and *MED13*, which is also part of the minimal region but only marginally significant for *de novo* enrichment. Finally, in the 2q11.2q13 deletion region we identify both *POU3F3*, which has been linked to ID and dysmorphic features by focal deletions[89], 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. While CNV maps historically offered sensitivity by detecting submicroscopic structural variants below the level of karyotype and high rates of recurrence based on the genome duplication architecture, exome sequencing added a new level of specificity allowing the identification of single genes by detection of disruptive *de novo* variants in the entire coding sequence of the human genome. Here, we expand and integrate the data by performing a meta-analysis of DNMs in 10,927 children with autism and NDDs with CNV morbidity data. The use of two DNM models (CH model and denovolyzeR) identifies a high-confidence intersection (n = 164 genes) and a comprehensive union (301 genes), which reach significance by one or both models. The combined analysis leverages the unique strengths of each model, such as the coding sequence chimpanzee–human divergence metrics in the CH model, and high-quality triplet context-based mutation rates in denovolyzeR. Examination of this gene set in the context of general population conservation statistics based on the ExAC population (pLI and missense Z scores from 45,376 individuals excluding neuropsychiatric cases, RVIS from the full 60,706 cases) confirms that this set list is enriched for genes that are constrained in the general population (LGD pLI 2.7e-62, Missense Z p=1.3e-51, RVIS p < 1e-7 to 5e-5) (Figure 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, while several are poorly constrained by both metrics (e.g. *MECP2* and Rett syndrome, RVIS = 32.4, pLI = 0.66). For example, we identify 22 genes that are intolerant to mutation by

pLI but not RVIS (RVIS > 20). Some of these are well-established genes (e.g., *KANSL1* and the Koolen-de Vries syndrome)[57] 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 scores are derived)[90]. This suggests challenges in the interpretation of disorders that may be driven by both *de novo* and somatic variation in the context of control exomes with almost no phenotypic detail. The presence of several X-chromosome linked genes in this set strongly suggests that gender will also have to be considered in population controls to account for sex-chromosome-specific differences, and potential female protective effects.

Application of subsampling and forward projection estimates indicates under the current DNM enrichment models that gene discovery based on recurrent LGD or severe missense mutations (MIS30) will soon plateau (Figure 4). Simply increasing sample size in this context may be in itself insufficient as the data suggest we are approaching an "n of 1" problem with respect to new gene discovery. Partitioning patients based on additional phenotypic criteria, sub-selecting genes based on functional pathway enrichment[51], integration of inherited variation[91], or targeting a small number of genes in much larger cohorts[10,11] are all strategies for increasing sensitivity for such classes of mutation. For example, analysis of this set of DNMs in the context of MAGI modules further identified key functional categories, including neurotransmitter/synaptic signaling and JNK/MAPK cascades. This analysis identifies 46 genes with DNMs among the four modules that do not yet reach significance but likely represent functionally important targets in 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 where 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]. It is interesting that many of the nominally significant genes are distinct from those identified by severe recurrent DNM identifying new sets of patients and new candidate genes for clinical investigation and potential therapeutic targets. We propose that this class of mutation (less severe or clustered missense DNMs) represents the most promising reservoir for future gene discovery. 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[19] as well as the development of more sophisticated missense clustering algorithms[15].

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 301 genes originates from ID/DD cases (>3:1). Importantly, DNM-enriched genes significantly overlap known pathogenic CNV regions (p = 0.0005, LR+ 2.53 [95% CI 1.43 – 1.91]) supporting a common genetic etiology. These specific targets have offered both independent confirmation of existing single-gene associations (e.g., *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 recent findings suggesting a role for multiple noncoding mutations in putative regulatory sequence in ASD[92,93], a feature that cannot be assayed by exome datasets. Among the genes with recurrent mutation and CNV intersection, the *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 protein-protein interactions[94-96]. Our analysis builds on these studies by providing evidence of recurrent missense mutation enrichment in human NDDs.

Finally, it is interesting that the 301 genes we highlight in this meta-analysis are enriched for expression in the D1+ and D2+ medium spiny neurons of the striatum (Figure 4). Previously, Dougherty and colleagues highlighted this particular brain region based on a survey of genes reported as autism candidate risk genes[56]. 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[93]. 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[97-106]. The striatum is particularly compelling as it has been linked to repetitive behaviors[97] core to the autism phenotype and also to genes known to be involved in DD, including *CHD8*, *SHANK3*, *FOXP2*, and *KCNA4*[98,99,102,105,106]. 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[101] 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 301 genes that show evidence for recurrent DNM represent a starting point for further functional and phenotypic investigations. These genes 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. In addition, we identify new candidates by clustered missense mutations, many of which are enriched in synaptic function but appear distinct from those genes where recurrent DNM are observed. 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[20,21], our results support a common genetic etiology among broad neurodevelopmental phenotypes. These genes are candidates for a genotype-first paradigm[107], where 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[108] and additional support for their role in NDD.

ACKNOWLEDGEMENTS

METHODS

Data set: We selected *de novo* variation from 10,927 cases with neurodevelopmental diagnoses of ASD or ID/DD compiled in the denovo-db v.1.5 release[19]. 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.

Statistical 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. In addition, we ran a recently published modified version[39] that separately tests for enrichment of variants with CADD v.1.3 scores[38] of 30 and higher, which are predicted to be the most damaging of missense variation. Similarly denovolyzeR[12] 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 < 0.1 and a DNM count of 2 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, CLUMP). Detection of excess missense clustering in genes was performed using the CLUMP tool as previously described[13,14]. Briefly, we implemented the permutation (-z 1000) and minimum mutation options (-m 2) and calculated a p-value based on the null distribution of case and control CLUMP score differences. To enable exact transcript comparisons between cases and controls, all variants were re-annotated with CRAVAT[109] resulting in annotation of 9,772 *de novo* missense variants in cases. The control set included private missense mutations in individuals from ExAC without neuropsychiatric disorders (n = 45,376; 1,466,439 mutations). All genes with at least two mutations in cases and controls were tested (n = 1930). For each gene we initially performed 1,000 simulations and in cases where nominal significance was reached we performed 1,000,000 simulations and calculated an empirical p-value[110]. Functional enrichment was examined using the CSEA and TSEA tools[56]. Identification of clustered gene modules was performed using the MAGI (merging affected genes into integrated networks)[51] enrichment tool with default settings and incorporating co-expression and physical interaction data from geneMANIA[111]. 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 ≤ 0.1 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 by BIC. Model fits and confidence bounds were performed using the base stats and propagate (https://CRAN.R-project.org/package=propagate) packages in the R statistical language.

FIGURE LEGENDS

Figure 1: *de novo* enriched genes and their characteristics. Shown are the results of applying both the chimpanzee–human (CH) divergence model and denovolyzeR to *de novo* variation in 10,927 individuals with ASD/ID/DD. The two models show considerable gene overlap **(A,B)** with correlated significance values (LGD $r^2 = 0.94$, 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 outliers include *CAPN15*, *SNAPC5*, *DLX3*, *TMEM178A*, *ADAP1*, *SNX5*, *SMARCD1*, *WDR26*, and *AGO4*, and denovolyzeR missense outliers include *ITPR1*, *RAC1*, *SETD1B*, *WDFY4*, and *UNC80* **(D)**. Recurrent mutated LGD genes (true) are highly enriched for genes intolerant to mutation as defined by ExAC pLI score **(E)**. Genes significantly enriched for missense DNMs are outliers by the ExAC missense depletion Z scores **(F)**. Similarly, all subcategories of significant genes are intolerant to mutation (RVIS percentile) when compared to non-significant genes (p-values are corrected for all

possible group comparisons). The LGD and MIS30 group is not shown as it only represents one gene (*TCF7L2*, RVIS percentile = 10.96) **(G)**.

Figure 2: Comparison of *de novo* variation rates in ASD and ID/DD. The plots compare DNM rates for genes for patients from ASD and ID/DD 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. Additional candidates for phenotypic bias at nominal significance (dashed lines at p = 0.05) were also identified. Larger cohorts will be needed to confirm gene biases, especially with respect to ASD.

Figure 3: Gene expression and protein-interaction networks. (A-D) MAGI[51] highlights four modules of co-expression and protein-protein interaction, including genes significant for DNM enrichment (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 patients with DNM within this cohort. Edges depict protein-protein interactions (pink arcs) and co-expression (green arcs) scaled by their scores from geneMANIA[111]. **(E-F)** Cell and tissue-specific enrichment analyses (CSEA and TSEA) highlight a strong bias to various developing parts of the brain with the strongest signal early to mid-fetal development (color corresponds to p-values, shaded regions closer to the center of each hexagon indicate increasing tissue specificity). 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[56] and we now observe cell-specific enrichment among genes with a significant excess of DNMs.

Figure 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 missense (CADD score >30) DNM 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, likely represents the most important reservoir for new gene discovery as sequence data are generated from additional ASD and DD cohorts.

Figure 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 (red, turquoise and blue points). The analysis confirms known associations such as *PAFAH1B1* (LIS1), *RAI1*, and *KANSL1* and candidate associations for *MAPK3* and *SIN3A*. Recurrent severe missense of *GABRB3* has 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 dataset) are known to cause the core phenotype of Prader-Willi and Angelman syndromes, respectively, but do not reach significance in this analysis.

Table 1: Recurrent DNM gene summary and model comparison.

Variant Category	CH Count	denovolyzeR Count	Union Count	Intersection Count	ID/DD Only Union Count	ASD Only Union Count
LGD & Missense_30 & Missense	14	NA	14	NA	4	0
LGD & Missense_30	1	NA	1	NA	0	0
LGD & Missense	14	24	16	22	22	4
LGD	104	122	132	94	103	35
Missense_30 & Missense	32	NA	34	NA	11	0
Missense_30	23	NA	21	NA	0	0
Missense	58	75	83	48	80	6
total	246	221	301	164	220	45

Table 2: Genes enriched for *de novo* variation in 10,927 ASD/ID/DD patients in denovo-db v1.5

Significance Category	Genes
LGD	*ARID1B*[†], *ANKRD11*[†], *KMT2A*[†], *ADNP*[†], *SYNGAP1*[†], *CTNNB1*[†], *ASXL3*[†], *SETD5*[†], *CHD8*[†], *KAT6B*[†], *EP300*[†], *GATAD2B*[†], *WDR45*[†], *AHDC1*[†], *TCF20*[†], *WAC*[†], *PPM1D*[†], *KAT6A*[†], *BCL11A*[†], *SMC1A*[†], *EHMT1*[†], *ASH1L*[†], *SUV420H1*[†], *SHANK3*[†], *ANP32A*[†], *ENO3*[†], *NONO*, *MEIS2*, *KANSL1*[†], *LEO1*, *SRCAP*[†], *SET*[†], *CAPRIN1*, *NSD1*[†], *SOX5*[†], *CHAMP1*[†], *CUL3*[†], *TRIP12*[†], *CNKSR2*[†], *KIAA2022*[†], *USP9X*[†], *KDM5B*[†], *NAA15*[†], *SYNCRIP*[†], *QRICH1*[†], *MSL3*[†], *MBD5*[†], *WDFY3*[†], *CHD7*[†], *SPAST*[†], *BRPF1*[†], *WHSC1*[†], *PHF12*[†], *CLTC*[†], *ZC4H2*[†], *DLG4*[†], *TBR1*[†], *DSCAM*[†], *KMT2C*[†], *SETD2*[†], *SETBP1*[†] ,*EBF3*[†], *KIF11*[†], *TCF12*[†], *KDM6A*[†], *UPF3B*[†], *SPRY2*[†], *DVL3*, *HNRNPD*, *HIVEP3*[†], *AUTS2*[†], *TRA2B*, *CASZ1*[†], *FOXP2*[†], *SIN3A*[†], *PDHA1*[†], *ASXL1*[†], *CDC42BPB*, *ZBTB7A*[†], *LARP4B*[†], *SRSF11*, *SON*[†], *PHF21A*, *PHIP*[†], *RPL26*[†], *ANK2*[†], *VEZF1*[†], *HIST1H1E*[†], *PHF3*[†], *PRPF40A*, *SRRM2*[†], *FOSL2*[†], *HNRNPK*, *MED13*[†], *SLC2A1*[†], *PHF8*, *FIGN*[†], *C9orf142*[†], *SPRED2*[†], *TNRC6B*[†], *ZBTB10*, *BAZ2B*, *PAX5*, *CELF2*, *CDKL5*, *HIVEP2*, *GIGYF1*, *ODC1*, *FAM200B*, *KCNS3*, *SPEN*, *KDM6B*, *SKIDA1*, *ZBTB18*, *NFIB*, *ZNF292*, *ARID2*, *FAM200A*, *PRR12*, *TAB2*, *RAI1*, *POU3F3*, *ZFHX3*, *NFE2L3*, *PTCHD1*, *STARD9*, *WDR87*, *IRF2BPL*, *SALL3*, *GPRIN1*, *MN1*, *YTHDF3*
LGD & Missense	*FOXP1*[†], *MEF2C*[†], *CHD2*[†], *TLK2*[†], *FBXO11*[†], *MYT1L*[†], *DYRK1A*[†], *CNOT3*[†], *POGZ*[†], *NFIX*[†], *HNRNPU*[†], *WDR26*
LGD & MIS30	*TCF7L2*[†]
LGD & Missense & MIS30	*PPP2R5D*[†], *SCN2A*[†], *DDX3X*[†], *SLC6A1*[†], *TBL1XR1*[†], *STXBP1*[†], *HDAC8*[†], *MECP2*[†], *TCF4*[†], *SATB2*[†], *CASK*[†], *PUF60*[†], *MED13L*[†], *IQSEC2*[†]
Missense	*GNAI1*[†], *PPP1CB*[†], *ITPR1*, *PTPN11*[†], *RAC1*, *GABRB3*[†], *SETD1B*, *WDFY4*, *UNC80*, *PLK5*, *BTF3*[†], *RFX8*, *SHISA6*, *GABRB2*[†], *SMAD4*[†], *SUSD4*, *AGAP2*[†], *PRKD1*[†], *TRRAP*[†], *DNM1*, *HMGXB3*, *SMC3*[†], *CHD4*, *PSMG4*, *VAMP2*[†], *MYO1E*[†], *PRKCA*[†], *SYT1*, *TMEM42*[†], *RYR2*, *PLAC8L1*, *TRAF7*[†], *PBX1*[†], *KCNJ6*, *RRP8*[†], *SSC5D*, *CBL*[†], *LIPT2*, *MUC12*, *TNPO3*[†], *TMEM232*[†], *CHMP1A* ,*GPRIN2*, *TET2*, *PLCG2*, *C2orf42*, *SOX4*, *CABP7*, *AQP10*, *EZR*, *CEP85*, *SF3B1*, *DEAF1*, *HUWE1*, *PRPF18*, *GRIN1*, *YWHAG*, *SLC35A2*, *SNAPC5*, *ACHE*, *CLASP1*, *GLRA2*, *INSM2*, *TAOK1*, *ABI2*, *MAPK3*, *G3BP2*, *AGO1*, *AP2M1*, *MTF2*, *EGLN2*, *DDX50*, *KCNC1*, *ANKRD17*, *SNX5*, *TMEM178A*, *DLX3*, *ADAP1*, *STXBP5L*, *SMARCD1*, *AGO4*, *CAPN15*, *EIF4G2*
MIS30	*CACNA1A*, *CACNA1E*, *KIF5C*, *ATP1A3*, *PDK2*, *GABRG2*, *ACTC1*, *PACS2*, *TAF1*, *MYO1G*, *PAPOLG*, *FAM104A*, *HIST1H2AC*, *STC1*, *AGO3*, *PAFAH1B1*, *SEPT10*, *DDX23*, *UBTF*, *KIAA0195*, *MEF2D*
Missense & MIS30	*KCNQ2*[†], *CSNK2A1*[†], *EEF1A2*[†], *NAA10*[†], *CDK13*[†], *DYNC1H1*[†], *GNAO1*[†], *SMARCA2*[†], *PACS1*[†], *GRIN2B*[†], *SMARCA4*[†], *COL4A3BP*[†], *DNMT3A*[†], *SCN8A*[†], *CTCF*[†], *KIF1A*[†], *BRAF*[†], *HECW2*[†], *CHD3*[†], *TRIO*[†], *ZMYND11*[†], *KCNQ3*[†], *MAP2K1*[†], *KCNH1*, *PIK3CA*[†], *EFTUD2*[†], *NR2F1*[†], *TNPO2*, *RAB11A*[†], *SOX11*, *PPP2R1A*, *U2AF2*, *NR4A2*, *KCND3*

[†]Gene also in the intersection set.

Table 3: Intersection between pathogenic CNVs and recurrently mutated genes.

	Gene Symbol*	LGD de novo Variants All (n=10,927)	ASD (n=5,624)	ID/DD (n=5,303)	Missense de novo Variants All (n=10,927)	ASD (n=5,624)	ID/DD (n=5,303)	Missense CADD ≥ 30 de novo Variants All (n=10,927)	ASD (n=5,624)	ID/DD (n=5,303)	Union CH and denovolyzeR Significance (q ≤ 0.1, count > 1)	Deletion Syndrome	Duplication Syndrome	CNV Significance Type[18]	Decipher† 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	MIS30 & 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
	GRIN1	0	0	0	5	1	4	2	1	1	MIS	9q34-deletion	9q34-duplication	DEL AND DUP	Category 1
	C9orf142	2	0	2	0	0	0	0	0	0	LGD	9q34-deletion	9q34-duplication	DEL AND DUP	Category 1
	PAFAH1B1	0	0	0	2	0	2	2	0	2	MIS30	Miller-Dieker syndrome (MDS)	17p13.3-duplication-(both-YWHAE-and-PAFAH1B1)	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
	GPRIN1	2	1	1	0	0	0	0	0	0	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		
	MED13	3	3	0	3	1	2	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 & MIS	2q33.1	None		Category 1
	SATB2‡	9	0	9	6	0	6	5	0	5	LGD & MIS30 & MIS	2q33.1	None		Category 1

Gene										Type	Region	Region	Category
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 intellectual disability	
HUWE1	0	0	0	9	0	9	3	0	3	MIS	None	Xp11.22-linked intellectual disability	
WDR45	8	0	8	2	1	1	0	0	0	LGD	None	Xp11.22-p11.23 Microduplication	
SLC35A2	1	0	1	3	0	3	0	0	0	MIS	None	Xp11.22-p11.23 Microduplication	
MECP2	11	4	7	7	0	7	4	0	4	LGD & MIS30 & MIS	None	Xq28 (MECP2) duplication	
CREBBP	3	0	3	13	3	10	1	1	0	LGD & 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	
14 Significant Regions[18]													
SATB2‡	9	0	9	6	0	6	5	0	5	LGD & MIS30 & MIS	2q33.1 (SATB2) Deletion		
MEF2C	4	0	4	5	1	4	0	0	0	LGD & 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	
PAX5	2	2	0	0	0	0	0	0	0	LGD		9p13 Duplication	

* Underlined genes represent established causal associations.

† https://decipher.sanger.ac.uk/disorders#syndromes/overview

‡ *SATB2* is significant both by focal deletions and as part of the 2q33.1 region.
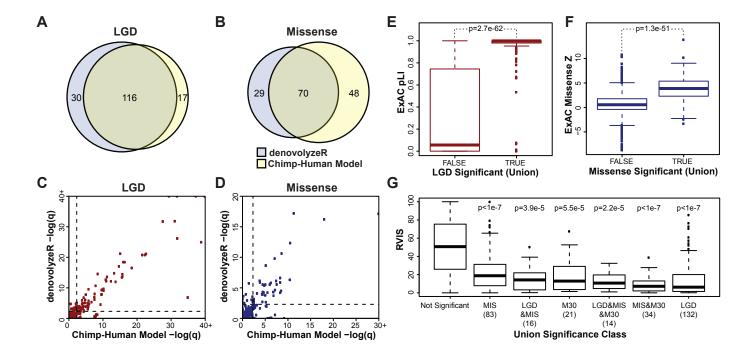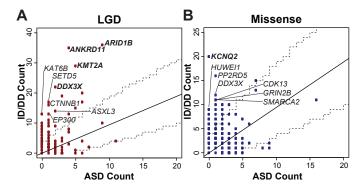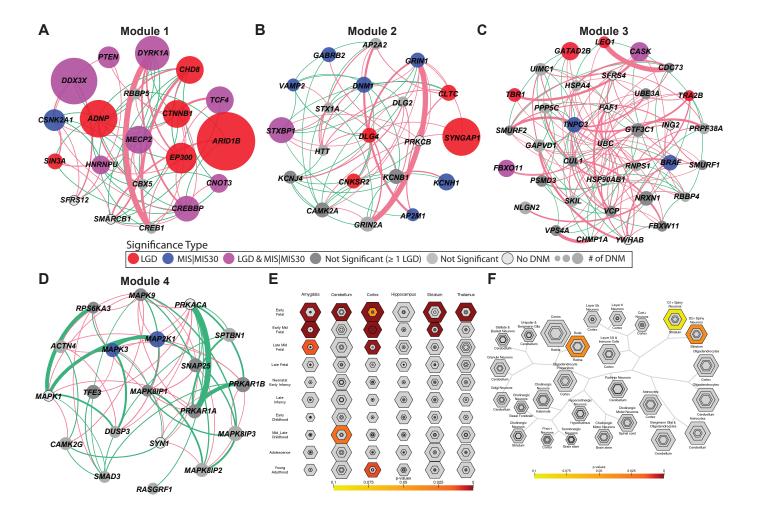
Figure 1

Figure 2

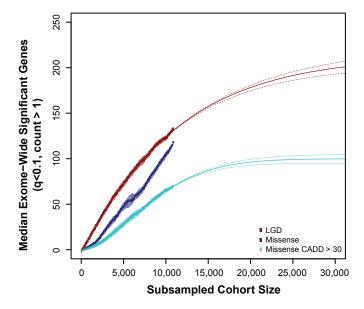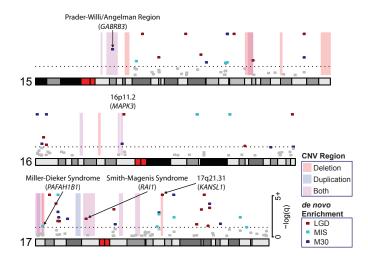# Figure 3

Figure 4

Figure 5

REFERENCES

1.   Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* 316, 445-9 (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, 727-32 (2005).

4.   de Vries, B.B. *et al.* Diagnostic genome profiling in mental retardation. *Am J Hum Genet* 77, 606-16 (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, 1005-17 (2001).

6.   de Vries, B.B. *et al.* Clinical studies on submicroscopic subtelomeric rearrangements: a checklist. *J Med Genet* 38, 145-50 (2001).

7.   Firth, H.V. & Wright, C.F. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol* 53, 702-3 (2011).

8.   O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* 43, 585-9 (2011).

9.   Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 519, 223-8 (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, 1619-22 (2012).

12.   Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* 46, 944-50 (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, 838-46 (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, 777-84 (2011).

18.   Coe, B.P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* 46, 1063-71 (2014).

19.   Turner, T.N. *et al.* denovo-db: a compendium of human de novo variants. *Nucleic Acids Res* 45, D804-D811 (2017).

20.   Matson, J.L. & Shoemaker, M. Intellectual disability and its relationship to autism spectrum disorders. *Res Dev Disabil* 30, 1107-14 (2009).

21.   Diagnostic and statistical manual of mental disorders. 5. *American Psychiatric Association* (2013).

22. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216-21 (2014).
23. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 542, 433-438 (2017).
24. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* 367, 1921-9 (2012).
25. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515, 209-15 (2014).
26. Halvardson, J. *et al.* Mutations in HECW2 are associated with intellectual disability and epilepsy. *J Med Genet* 53, 697-704 (2016).
27. Hashimoto, R. *et al.* Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. *J Hum Genet* 61, 199-206 (2016).
28. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat Genet* 47, 582-8 (2015).
29. 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, 3481-9 (2014).
30. Lelieveld, S.H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat Neurosci* 19, 1194-6 (2016).
31. Michaelson, J.J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* 151, 1431-42 (2012).
32. 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).
33. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 380, 1674-82 (2012).
34. RK, C.Y. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat Neurosci* 20, 602-611 (2017).
35. Tavassoli, T. *et al.* De novo SCN2A splice site mutation in a boy with Autism spectrum disorder. *BMC Med Genet* 15, 35 (2014).
36. 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).
37. Yuen, R.K. *et al.* Genome-wide characteristics of de novo mutations in autism. *NPJ Genom Med* 1, 160271-1602710 (2016).
38. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46, 310-5 (2014).
39. Wang, T. *et al.* De novo genic mutations among a Chinese autism spectrum disorder cohort. *Nat Commun* 7, 13316 (2016).
40. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* 74, 285-99 (2012).
41. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285-91 (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, 22-9 (2010).
43. 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).

44.   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).

45.   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).

46.   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).

47.   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).

48.   Hoischen, A. *et al.* De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat Genet* 43, 729-31 (2011).

49.   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).

50.   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).

51.   Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E.E. The discovery of integrated gene networks for autism and related disorders. *Genome Res* 25, 142-54 (2015).

52.   Endele, S. *et al.* Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat Genet* 42, 1021-6 (2010).

53.   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, 937-47 (2010).

54.   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).

55.   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, 4218-30 (2010).

56.   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, 1420-31 (2014).

57.   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, 652-9 (2016).

58.   Takahashi, S. *et al.* Characterization of intragenic tandem duplication in the PAFAH1B1 gene leading to isolated lissencephaly sequence. *Mol Cytogenet* 8, 84 (2015).

59.   Phelan, K. & Rogers, R.C. Phelan-McDermid Syndrome. in *GeneReviews(R)* (eds. Adam, M.P. *et al.*) (Seattle (WA), 1993).

60.   Bi, W. *et al.* Mutations of RAI1, a PHD-containing protein, in nondeletion patients with Smith-Magenis syndrome. *Hum Genet* 115, 515-24 (2004).

61.   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).

62.   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, 1028-32 (2010).

63.   Shimbo, H. *et al.* Haploinsufficiency of BCL11A associated with cerebellar abnormalities in 2p15p16.1 deletion syndrome. *Mol Genet Genomic Med* 5, 429-437 (2017).

64.   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).

65. Kumar, R.A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* 17, 628-38 (2008).

66. 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, 3011-8 (2015).

67. 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).

68. 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, 602-12 (2013).

69. 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, 645-9 (2012).

70. Halder, D. *et al.* Suppression of Sin3A activity promotes differentiation of pluripotent cells into functional neurons. *Sci Rep* 7, 44818 (2017).

71. 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, 877-87 (2016).

72. DeMari, J. *et al.* CLTC as a clinically novel gene associated with multiple malformations and developmental delay. *Am J Med Genet A* 170A, 958-66 (2016).

73. 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).

74. Buxbaum, J.D. *et al.* Association between a GABRB3 polymorphism and autism. *Mol Psychiatry* 7, 311-6 (2002).

75. Guella, I. *et al.* De Novo Mutations in YWHAG Cause Early-Onset Epilepsy. *Am J Hum Genet* 101, 300-310 (2017).

76. Asadollahi, R. *et al.* The clinical significance of small copy number variants in neurodevelopmental disorders. *J Med Genet* 51, 677-88 (2014).

77. Harrington, A.J. *et al.* MEF2C regulates cortical inhibitory and excitatory synapses and behaviors relevant to neurodevelopmental disorders. *Elife* 5(2016).

78. 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).

79. Kohannim, O. *et al.* Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression. *Front Neurosci* 6, 115 (2012).

80. 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).

81. O'Roak, B.J. *et al.* Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nat Commun* 5, 5595 (2014).

82. 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).

83. Craxton, A. *et al.* XLS (c9orf142) is a new component of mammalian DNA double-stranded break repair. *Cell Death Differ* 22, 890-7 (2015).

84. Zehavi, Y. *et al.* De novo GRIN1 mutations: An emerging cause of severe early infantile encephalopathy. *Eur J Med Genet* 60, 317-320 (2017).

85. 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).

86. 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* (2017).

87. 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, 615-26 (2013).

88. Levy, J. *et al.* Molecular and clinical delineation of 2p15p16.1 microdeletion syndrome. *Am J Med Genet A* 173, 2081-2087 (2017).

89. 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, 32-5 (2014).

90. 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).

91. 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).

92. Werling DM *et al.* Limited contribution of rare, noncoding variation to autism spectrum disorder from sequencing of 2,076 genomes in quartet families. *BioRxiv* (2017).

93. Turner TN *et al.* Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* (2017).

94. 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).

95. 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, 3190-200 (2015).

96. 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 U S A* 113, 8520-5 (2016).

97. Langen, M. *et al.* Changes in the development of striatum are involved in repetitive behavior in autism. *Biol Psychiatry* 76, 405-11 (2014).

98. Platt, R.J. *et al.* Chd8 Mutation Leads to Autistic-like Behaviors and Impaired Striatal Circuits. *Cell Rep* 19, 335-350 (2017).

99. 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* 10, 26 (2017).

100. Balsters, J.H., Mantini, D. & Wenderoth, N. Connectivity-based parcellation reveals distinct cortico-striatal connectivity fingerprints in Autism Spectrum Disorder. *Neuroimage* (2017).

101. 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).

102. Kaya, N. *et al.* KCNA4 deficiency leads to a syndrome of abnormal striatum, congenital cataract and intellectual disability. *J Med Genet* (2016).

103. Flanigan, M. & LeClair, K. Shared Motivational Functions of Ventral Striatum D1 and D2 Medium Spiny Neurons. *J Neurosci* 37, 6177-6179 (2017).

104. Sanders, S.J. First glimpses of the neurobiology of autism spectrum disorder. *Curr Opin Genet Dev* 33, 80-92 (2015).

105. Schreiweis, C. *et al.* Humanized Foxp2 accelerates learning by enhancing transitions from declarative to procedural performance. *Proc Natl Acad Sci U S A* 111, 14253-8 (2014).

106.   Chen, Y.C. *et al.* Foxp2 controls synaptic wiring of corticostriatal circuits and vocal communication by opposing Mef2c. *Nat Neurosci* 19, 1513-1522 (2016).
107.   Stessman, H.A., Bernier, R. & Eichler, E.E. A genotype-first approach to defining the subtypes of a complex disease. *Cell* 156, 872-7 (2014).
108.   Bernier, R. *et al.* Disruptive CHD8 mutations define a subtype of autism early in development. *Cell* 158, 263-276 (2014).
109.   Douville, C. *et al.* CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics* 29, 647-8 (2013).
110.   Phipson, B. & Smyth, G.K. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol* 9, Article39 (2010).
111.   Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 38, W214-20 (2010).

Table S1: De Novo Variation in 10,927 ASD/ID/DD Patients in denovodb v1.5

The table contains the following columns for each panel: Gene, LGD, MIS, MIS30, and model-indicator columns (CH Model / denovolyzeR under LGD, MIS, MIS30, and CLUMP) shown as colored cells. Only the numeric values are transcribed below.

Gene	LGD	MIS	MIS30
ARID1B	45	4	0
ANKRD11	39	1	0
ADNP	26	2	0
KMT2A	34	6	1
DDX3X	24	13	4
SYNGAP1	23	5	1
ASXL3	22	0	0
DYRK1A	20	5	1
POGZ	17	5	0
HNRNPU	7	2	0
SCN2A	19	22	6
CTNNB1	17	1	0
SETD5	18	5	0
MED13L	16	8	4
CHD8	15	3	0
MECP2	11	7	4
AHDC1	12	1	0
FOXP1	11	6	1
EP300	14	7	1
KAT6B	13	3	1
TCF4	11	7	4
PURA	7	7	0
GATAD2B	10	0	0
WDR45	8	2	0
TCF20	10	1	0
SHANK3	10	1	0
SATB2	9	6	5
CHD2	12	8	3
KAT6A	9	1	0
WAC	8	1	1
PPM1D	8	0	0
KANSL1	8	1	0
BCL11A	6	3	0
SMC1A	8	2	1
EHMT1	9	3	2
ASH1L	9	1	0
SUV420H1	7	3	1
STXBP1	7	9	7
CHAMP1	4	1	0
TBL1XR1	5	8	2
ANP32A	2	0	0
ENO3	2	0	0
NONO	2	0	0
MEIS2	2	1	0
SRCAP	8	5	0
KIAA2022	6	1	0
WDR26	2	2	0
LEO1	4	0	0
SET	5	0	0
CAPRIN1	2	1	0
NSD1	8	5	1
SLC6A1	5	12	9
NFIX	7	3	1
USP9X	7	3	1
CASK	5	6	3
MYT1L	5	6	1
SOX5	5	2	2
CUL3	5	2	0
PTEN	4	8	0
TRIP12	6	4	2
CNKSR2	5	0	0
MEF2C	4	5	0
UPF3B	5	1	0
KDM5B	7	4	2
IQSEC2	4	4	3
MSL3	4	0	0
HIVEP3	4	1	0
ZBTB18	4	2	1
MBD5	5	0	0
NAA15	4	1	0
SYNCRIP	3	0	0
QRICH1	4	0	0
FOXG1	3	9	1
PUF60	4	5	2

Gene	LGD	MIS	MIS30
DLG4	4	1	1
SETBP1	4	2	0
WDFY3	6	6	1
DSCAM	5	1	0
BRPF1	4	1	0
CASZ1	4	0	0
SETD2	5	2	1
CHD7	5	5	2
WHSC1	4	2	1
SPAST	3	1	0
CNOT3	4	5	2
FBXO11	4	5	0
ZC4H2	3	0	0
KMT2C	7	4	0
TLK2	3	5	1
ANK2	6	5	1
ASXL1	4	1	0
HDAC8	3	6	3
PHF12	3	2	0
CLTC	4	2	1
FAM200B	2	0	0
TBR1	3	2	0
SON	4	2	0
AUTS2	4	1	0
YTHDF3	3	1	0
HIST1H1E	2	0	0
SMARCD1	1	2	0
EBF3	3	3	1
PDHA1	3	1	1
KDM6A	4	3	2
KIF11	3	1	0
TCF12	3	1	0
LARP4B	3	0	0
SPRY2	2	0	0
DVL3	2	0	0
SRRM2	4	5	0
PPP2R5D	2	17	12
HNRNPD	2	0	0
SPEN	4	4	1
STARD9	3	3	0
TRA2B	2	0	0
RAI1	3	3	0
CDKL5	3	4	1
TCF7L2	3	3	3
TAB2	3	1	0
PRR12	3	3	0
POU3F3	2	2	2
FOXP2	3	1	1
FAM200A	2	0	0
KDM6B	3	3	1
RPL26	2	0	0
PHIP	4	6	2
SIN3A	3	1	0
CDC42BPB	3	3	0
ZBTB7A	2	0	0
KCNS3	2	0	0
FOSL2	2	0	0
SRSF11	2	0	0
WDR87	2	1	0
PHF21A	2	0	0
HIVEP2	3	2	2
ARID2	3	1	1
ODC1	2	0	0
TNRC6B	3	2	0
IRF2BPL	2	0	0
VEZF1	2	2	0
ZNF292	3	3	0
SKIDA1	2	2	0
NFE2L3	2	0	0
PHF3	3	2	1
PTCHD1	2	0	0
CREBBP	3	13	1
PRPF40A	3	1	0
PAX5	2	0	0

Gene	LGD	MIS	MIS30
HNRNPK	2	1	1
MED13	3	3	0
SALL3	2	0	0
GIGYF1	3	1	0
GPRIN1	2	0	0
CELF2	2	0	0
SLC2A1	2	1	1
ZFHX3	3	5	0
PHF8	2	0	0
FIGN	2	2	0
NFIB	2	0	0
SPRED2	2	1	0
C9orf142	2	0	0
MN1	2	0	0
ZBTB10	2	0	0
BAZ2B	3	2	0
EFTUD2	2	6	4
MED12	3	5	0
GRIN2B	2	12	5
DNMT3A	2	9	6
KCNQ2	2	20	5
TNPO3	2	5	1
CDK13	2	12	6
KDM5C	2	2	0
ARID1A	2	3	0
SOX11	1	4	2
CACNA1A	2	6	3
CHD3	2	10	7
SLC35A2	1	3	0
CSNK2A1	0	12	6
CAPN15	0	3	1
EEF1A2	0	11	6
CTCF	1	8	6
GNAO1	1	8	5
PACS1	0	10	9
COL4A3BP	0	8	5
SNAPC5	0	2	0
BRAF	0	7	2
DLX3	0	2	0
TMEM178A	0	2	1
ADAP1	0	2	0
SNX5	0	2	0
NAA10	0	7	4
DYNC1H1	0	20	13
SMARCA2	0	12	7
AGO4	0	2	0
SCN8A	1	11	3
SMARCA4	0	13	10
GNAI1	0	7	2
PIK3CA	0	6	2
MAP2K1	0	5	2
GABRB2	0	5	1
PPP1CB	0	6	1
PTPN11	0	7	2
ZMYND11	1	6	3
KIF1A	0	11	9
ITPR1	0	12	4
RAC1	0	5	2
GABRB3	1	6	0
SETD1B	0	5	0
WDFY4	0	5	1
UNC80	1	6	2
HECW2	1	9	7
TRIO	1	12	5
PLK5	0	3	0
BTF3	0	4	0
KCNQ3	0	7	6
RFX8	0	3	0
NR2F1	0	5	3
SMAD4	0	5	1
KCNH1	0	7	5
AGAP2	0	2	0
VAMP2	0	3	1
PRPF18	0	3	0

Gene	LGD	MIS	MIS30
TMEM42	0	3	1
ABI2	0	3	1
PPP2R1A	0	4	3
CBL	1	5	0
KCND3	0	4	2
SHISA6	0	3	1
SUSD4	0	5	0
TRRAP	1	12	3
PRKD1	0	6	3
SF3B1	0	5	1
CABP7	0	3	0
CLASP1	0	5	2
DNM1	0	6	4
PBX1	1	4	0
HMGXB3	0	3	1
NR4A2	0	4	2
C2orf42	0	4	0
SMC3	1	6	1
CHD4	1	8	2
PSMG4	0	2	0
MYO1E	1	6	1
ACHE	0	4	0
AQP10	0	3	0
PRKCA	0	5	1
SYT1	0	4	2
MTF2	0	3	0
RAB11A	0	3	2
EGLN2	0	3	0
GLRA2	0	3	0
TRAF7	1	5	2
YWHAG	0	3	1
DEAF1	0	4	1
AGO1	0	4	0
KCNC1	0	3	0
RRP8	0	4	1
TAOK1	0	4	2
MAPK3	0	3	1
RYR2	0	11	3
PLAC8L1	0	3	0
HUWE1	0	9	3
KCNJ6	0	4	0
SSC5D	0	3	1
CEP85	0	2	0
GRIN1	0	5	2
EIF4G2	1	4	1
EZR	1	4	0
INSM2	0	4	0
SOX4	0	4	0
LIPT2	0	2	0
MUC12	1	4	0
TNPO2	0	5	4
U2AF2	0	4	3
ANKRD17	0	5	1
TMEM232	0	2	0
CHMP1A	0	3	2
GPRIN2	0	4	0
TET2	0	5	2
DDX50	0	3	1
STXBP5L	1	3	0
PLCG2	0	6	1
G3BP2	1	3	0
AP2M1	0	3	0
KCNK3	0	3	0
MAP2K6	0	2	0
SEPT10	0	2	2
GRIA1	0	4	2
ACTC1	0	3	3
KRTAP6-3	0	2	0
ZFP69	0	3	0
KIF5C	0	4	2
CACNA1E	2	7	4
FAM104A	0	2	0
ATP1A3	0	5	3
SPANXD	0	2	0

Gene	LGD	MIS	MIS30
PDK2	0	3	3
PAPOLG	0	3	3
ZRANB1	0	3	2
CD59	0	2	0
RFX2	0	4	0
ZNF215	0	3	0
RAD18	0	3	0
GABRG2	0	3	3
AGT	0	3	0
PACS2	0	4	4
GRIN2A	0	5	0
PRKAR1B	1	3	2
METTL16	0	3	1
AGO3	0	3	3
HIST1H2AC	0	2	2
PHACTR1	0	2	1
NUDCD2	0	2	0
CSNK1E	0	3	0
DNAH1	0	8	0
GSTA1	0	2	0
TAF1	0	5	5
PRB4	0	2	0
MYO1G	0	4	3
CIDEC	0	2	0
CBLL1	0	2	0
SH3BP2	0	3	0
SCN3A	0	4	2
OR5H14	0	2	0
C2orf70	0	2	0
STC1	0	2	2
SOX15	0	2	0
TNK2	0	4	0
KLHL14	0	3	0
CLCA2	0	3	0
DCDC1	0	2	0
HMGCL	0	2	0
LONP1	0	4	1
MAGEA6	0	2	0
HIBCH	0	2	0
ABI3BP	0	3	1
PAFAH1B1	0	2	2
MEF2D	0	2	2
RBL2	0	3	0
SHOC2	0	2	0
EZH2	0	2	1
FZD10	0	3	0
ABLIM1	0	3	0
RBM12	0	3	0
AKT3	1	2	2
ABR	0	3	2
GPN1	0	2	0
PCGF2	0	2	0
LRFN2	1	3	0
GALNTL5	0	2	0
SLC35C2	0	2	1
NDRG4	0	2	1
STXBP5	1	3	0
EPHA5	0	3	0
PFKP	0	3	0
DDX23	0	3	3
GLRA3	0	2	0
STIL	0	3	0
OGFOD3	0	2	2
UBTF	1	2	2
KIAA0195	0	3	3
CTR9	1	3	2
AGO2	0	3	2
CDH18	0	2	2
ATP6V0A1	0	2	0
NHLRC1	0	2	1
PCNX	0	3	0
UBQLN3	0	2	1
ANK3	0	6	1

Gene	LGD	MIS	MIS30
ACACB	1	4	2
CEP131	0	5	0
KDM3A	1	3	0
FANCM	1	3	0
TRIOBP	0	4	2
POLE	0	4	0
CEP350	0	4	0
LAMA1	0	4	0
UACA	1	2	0
RBM19	1	2	0
RTL1	1	2	0
ARHGEF1	0	3	1
RGS12	0	3	1
SPG11	0	3	1
RYR3	0	3	0
AFF3	0	3	0
SIN3B	0	3	0
UBR2	0	3	0
ZNF142	0	3	0
SLX4	0	3	0
PRUNE2	0	3	0
MUC17	0	3	0
ROCK1	1	2	0
RUNDC1	1	2	0
SLC9A3	1	2	0
ANKHD1-EIF4EBP3	1	1	0
MYO18A	0	2	2
INF2	0	2	1
DSP	0	2	1
COL5A2	0	2	1
KLHL17	0	2	0
DENND4B	0	2	0
SLITRK5	0	2	0
NPR2	0	2	0
COPG1	0	2	0
TTLL5	0	2	0
XPO5	0	2	0
FGA	0	2	0
MYO9A	0	2	0
MAPKBP1	0	2	0
MANBA	0	2	0
AP5Z1	0	2	0
FHOD3	0	2	0
IGSF9	0	2	0
IGHMBP2	0	2	0
EPHA10	0	2	0
MEFV	0	2	0
ZFYVE16	0	2	0
IQCE	0	2	0
MYBBP1A	0	2	0
CR1	0	2	0
STK39	0	2	0
TSPYL1	0	2	0
GALR3	0	2	0
STAG1	0	2	0
P2RX1	0	2	0
ZNF233	0	2	0
HCN3	0	2	0
NPBWR2	0	2	0
ABCG2	0	2	0
ATP1A2	0	2	1
ATIC	0	2	1
ZFP36	0	2	0
PATL1	0	2	0
SIGIRR	0	2	0
ZNF7	0	2	0
SELO	0	2	0
DPF2	0	2	1
ZNF420	0	2	0
CHST2	0	2	0
RGL1	0	2	1
STIM2	0	2	0
MIDN	0	2	0