

Identification of novel driver risk genes in CNV loci associated with neurodevelopmental disorders

Sara Azidane,^{1,2} Xavier Gallego,¹ Lynn Durham,¹ Mario Cáceres,^{2,3,4} Emre Guney,¹ and Laura Pérez-Cano^{1,5,*}

Summary

Copy-number variants (CNVs) are genome-wide structural variations involving the duplication or deletion of large nucleotide sequences. While these types of variations can be commonly found in humans, large and rare CNVs are known to contribute to the development of various neurodevelopmental disorders (NDDs), including autism spectrum disorder (ASD). Nevertheless, given that these NDD-risk CNVs cover broad regions of the genome, it is particularly challenging to pinpoint the critical gene(s) responsible for the manifestation of the phenotype. In this study, we performed a meta-analysis of CNV data from 11,614 affected individuals with NDDs and 4,031 control individuals from SFARI database to identify 41 NDD-risk CNV loci, including 24 novel regions. We also found evidence for dosage-sensitive genes within these regions being significantly enriched for known NDD-risk genes and pathways. In addition, a significant proportion of these genes was found to (1) converge in protein-protein interaction networks, (2) be among most expressed genes in the brain across all developmental stages, and (3) be hit by deletions that are significantly over-transmitted to individuals with ASD within multiplex ASD families from the iHART cohort. Finally, we conducted a burden analysis using 4,281 NDD cases from Decipher and iHART cohorts, and 2,504 neurotypical control individuals from 1000 Genomes and iHART, which resulted in the validation of the association of 162 dosage-sensitive genes driving risk for NDDs, including 22 novel NDD-risk genes. Importantly, most NDD-risk CNV loci entail multiple NDD-risk genes in agreement with a polygenic model associated with the majority of NDD cases.

Introduction

Copy-number variants (CNVs) are genome-wide structural changes involving the duplication or deletion of large nucleotide sequences, which can range from one kilobase (kb) to several megabases (Mb). This type of variation is commonly found in humans, and, even though it has been shown that most large CNVs are subject to high purifying selection,¹ previous reports have suggested that between 5% and 10% of the human genome is affected by CNVs.²

These variants could have significant functional consequences and they have been identified as an important mechanism of evolutionary adaptation. Rare genetic material gains or losses are also known to contribute substantially to the development of various neurodevelopmental disorders (NDDs), including autism spectrum disorder (ASD). Rare chromosomal rearrangements have been found in up to 10% of individuals with ASD,^{3,4} and it has additionally been observed that a larger fraction of the genome of these individuals is affected by CNVs compared with neurotypical individuals.^{5,6}

In particular, both rare *de novo* and inherited CNVs have been identified as penetrant variants that confer a major risk in the development of NDDs.³ Identification of such CNVs requires exceptionally large cohorts^{1,7} in genome-wide analysis, which are not often available, to

reach sufficient statistical power to detect the association of loci with NDD susceptibility. Despite such low frequency of rare CNVs, certain NDD-risk loci have been recurrently identified,^{8–10} most notably 1q21.1, 3q29, 7q11.23, 15q11-q13, 15q13.3, 16p11.2, 5q35, 17q12, and 22q11.2. These risk loci can usually be affected by copy-number gains or losses (Table S2), most of which have been independently associated with the manifestation of various NDDs. For instance, both 16p11.2 deletion and duplications confer 10-fold increase for ASD-risk,¹⁰ while 16p11.2 duplications also show >10-fold increase of risk of schizophrenia.⁹

There are significant limitations that condition the study of the impact of these structural variants in NDDs. First, considering the nature of rare CNVs, it is expected that they affect broad regions of the genome, often spanning several genes. Indeed, this fact makes it particularly challenging to perform fine mapping and to pinpoint the critical genes responsible for the manifestation of the phenotype. Furthermore, the current paradigm of CNV pathogenicity is primarily oriented toward finding a single gene responsible for causing the disease. These pathogenicity models can explain diseases with minimal phenotypic variability (i.e., similar or nearly identical clinical features and symptoms among affected individuals) or with highly penetrant variants, as in the case of *RAI1* in Smith-Magenis syndrome, for example. But in the case of

¹STALICLA Discovery and Data Science Unit, World Trade Center, Moll de Barcelona, Edif Este, 08039 Barcelona, Spain; ²Institut de Biotecnologia i de Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain; ³ICREA, 08010 Barcelona, Spain; ⁴Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute, Barcelona, Spain

⁵Lead contact

*Correspondence: laura.perez-cano@stalicla.com

<https://doi.org/10.1016/j.xhgg.2024.100316>.

© 2024 This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



complex diseases such as most neurodevelopmental conditions, these monogenic models have proven unsuccessful.

Due to the inherent degree of complexity in the study of such risk variants, not only are the causative genes often difficult to identify, but also their role in the pathophysiology underlying the manifestation of the disorder is difficult to characterize. Prior literature shows that pathogenic CNVs can give rise to a certain clinical phenotype through different events, either by affecting regulatory regions of the genome, by disrupting genes, or by affecting gene dosage.¹¹ However, there is growing evidence that the main mechanism whereby pathogenic CNVs trigger the phenotype is by means of dose sensitivity. To function properly, genes in risk regions that are sensitive to these dosage perturbations (i.e., haploinsufficient [HI] and/or triplosensitive [TS] genes) might require to be in stoichiometric equilibrium with other genes to ensure proper protein assembly, to form aggregates at high concentration levels, or to comply with concentration thresholds.¹² Therefore, when a copy-number gain or loss overlaps with any of these HI and/or TS genes and changes their dosage, the function of such gene is disrupted, resulting in the manifestation of the disease. Besides, previous studies have shown that the reciprocal genomic disorders, i.e., when the same region is associated to an NDD by both deletion or duplication, often show "mirror phenotypes." In this way, duplication of the region causes the opposite phenotype of what would be observed by the deletion, suggesting that one or more genes in each region might indeed be causing the phenotype through a dosage sensitivity mechanism.

In this study, we performed a meta-analysis of CNVs called from 11,614 affected individuals with NDDs and 4,031 neurotypical control individuals from the SFARI gene database to identify NDD-risk regions and the causative dosage-sensitive genes within these broad regions.

Subjects and methods

Meta-analysis for the identification of NDD-risk loci

We carried out a statistical meta-analysis of CNV calls covering 2,273 genomic loci in 11,614 affected individuals with NDDs and 4,031 control individuals from 18 studies (14 case studies and 4 additional control studies) gathered under the SFARI gene website. The data were standardized to avoid duplicated samples and to include only data derived from array analysis (see [supplemental methods](#)).

The statistical procedure of this case-control study entailed a two-sided Fisher's exact test in which a 2×2 contingency table is created for each of the 2,273 genomic regions analyzed, based on the number of cases and control individuals with and without CNVs in the given area of the genome. In this way, the degree of association between the presence of a CNV in that region and the expression of the phenotype is evaluated. To correct for multiple hypothesis testing, the p values were adjusted using Bonferroni correction and associations called statistically significant if $p < 2.19 \times 10^{-5}$ (i.e., below adjusted $p = 0.05$).

Dosage sensitivity

To identify which dosage-sensitive genes in the NDD-risk loci lead the manifestation of the phenotypes, we evaluated all genes whose coding sequence fully overlapped with the pathogenic risk region and annotated them with scores for probability of haploinsufficiency (pHI) and the probability of triplosensitivity (pTS). These scores were recently described by Collins et al.⁸ and provide a more precise metric of genes intolerant to dosage alterations than those used for short variants. The validity of these scores, which range from 0 to 1, has been investigated by assessing their accuracy in already known dosage-sensitive genes ($p < 10^{-100}$, two-sided Kolmogorov-Smirnov test of pTS and pHI scores).⁸ Cut-offs to define haploinsufficiency and triplosensitivity in these genes were established by Collins et al.⁸ at $pHI \geq 0.84$ and $pTS \geq 0.993$, taking into consideration that the effects of deleted regions are typically stronger than those of duplications.

We therefore generated a catalog of qualifying dosage-sensitive risk genes (qDSGs), by including dosage-sensitive genes located within our previously defined NDD-risk CNV loci. We categorized these qDSGs into three different tiers: HI genes (with $pHI \geq 0.840$ and $pTS \leq 0.993$), TS genes (with $pTS \geq 0.993$ and $pHI \leq 0.840$), and genes that are both haploinsufficient and triplosensitive (bHITS) (with $pTS \geq 0.993$ and $pHI \geq 0.84$).

To assess whether our list of qDSGs was enriched with genes previously identified as NDD-risk genes, we employed the GeneTrek catalog of NDD-risk genes. In this gene catalog high-confidence genes are extracted from SPARK/SFARI category 1, SysID primary, G2P DD "confirmed," and "Brain[Cognition]," as well as DBD tier1 and AR categories. On the other hand, candidate genes are derived from SPARK/SFARI categories 2, 3, and S, the SysID candidates, the DBD tier 2 to 4, the G2P DD others than categories "confirmed" or "Brain[Cognition]."¹³

Gene ontology enrichment analysis

To assess the biological implications at the level of function and disease mechanism, we performed a gene set enrichment analysis of the dosage-sensitive genes found in the NDD-risk CNV regions. For this, we used g:Gost from g:Profiler (version e109_eg56_p17_1d3191d), a suitable tool for checking enrichment of unranked gene lists.¹⁴ Gene ontology (GO) (release 2023-03-06) was used as the reference gene set database, which provides thousands of standardized terms classified in biological processes, cellular components, and molecular functions.^{15,16} As a result of this analysis, we obtained lists of enriched (i.e., over-represented) terms associated with our three sets of dosage-sensitive genes, namely TS genes, HI genes, or bHITS genes. Bonferroni correction was used to adjust multiple hypothesis testing and terms below an adjusted p value of 0.05 were deemed significantly associated to the genes provided as input.

TDT analysis

To evaluate how rare deletions affecting genes in our catalog are transmitted, we analyzed multiplex family CNV data from the iHART cohort, extracted from Ruzzo et al.,¹⁷ to perform a transmission disequilibrium test (TDT). The CNV callers utilized to detect variants in the iHART cohort (SMuFin, GenomeSTRiP, BreakDancer, LUMPY) exhibited lower sensitivity in identifying duplications and only rare deletions were used in this analysis.¹⁷ We performed a one-sided binomial test that evaluates if the transmission rate of inherited deletions in HI genes within qDSGs deviate significantly from the expected Mendelian transmission

rate (i.e., ~50%), considering all the siblings with ASD within each family. This approach yields a statistical p value that determines whether there is a significant over transmission of CNVs in our risk genes from parents to children with an ASD diagnosis.

Identification of dosage-sensitive genes associated with NDD-risk

A two-sided Fisher's test was used to infer the association of each qDSGs with risk for NDDs. In particular, we compared the frequency of affected individuals with NDDs from iHART¹⁷ and Decipher¹⁸ carrying rare deletions and duplications hitting the gene, against the frequency of neurotypical control individuals from the 1000 Genomes from the DGV¹⁹ database carrying such variation within the gene. Rare variants were defined as those not overlapping more than 50% with common variants ($AF \geq 0.001$) in the gnomAD database.²⁰ It is important to note that, when evaluating deletions, we compared the frequency of rare deletions in both iHART and Decipher affected individuals with that in the 1000 Genomes control individuals. However, for duplications, our analysis exclusively utilized data from Decipher affected individuals, due to the lack of sensitivity of the calling methods used to identify duplications in the iHART cohort. A Bonferroni-based multiple test correction was applied, with significance established at 1.93×10^{-4} for HI genes, 8.33×10^{-3} for TS genes, and 1.66×10^{-3} .

Brain-specific gene expression analysis

To better grasp how our qDSGs behave in the developing brain over time, we analyzed data from the Brain Span atlas database, which contains gene expression data from typical brain development. We made the RNA-seq counts more comparable by utilizing a variance stabilizing transformation,²¹ which helps to correct for differences in gene expression levels. We then identified genes that were among the top 5% in terms of expression during various developmental stages (fetal, perinatal, infancy, childhood, adolescence, and adulthood) and in different brain regions (such as neocortex, amygdala, and cerebellum). We then conducted several two-sided Fisher's exact tests to assess whether there was enrichment of our qDSGs in the top 5% expressed genes at different developmental stages and in different areas of the brain.

Protein-protein interaction network

The broad spectrum of NDDs involves a lattice of complex interactions of the proteins encoded by the risk genes. To understand the functional modules that can be established between the genes in the previously defined NDD-risk regions, we built a protein-protein interaction (PPI) network using the STRING platform, providing a protein-protein connection in the network for interactions with an interaction score value ≥ 0.400 , considering only high-confidence PPIs as defined by experiments and/or databases.

To investigate the biological significance of the connectedness of nodes in the PPI network, the expected and observed numbers of edges between specific groups of genes were evaluated. The expected number of edges is the number of connections that would be observed if the selection of nodes were random, while the observed is the actual number of connections within the network. If the observed number of edges is significantly higher than the expected number of connections, it indicates a higher interconnectivity between the proteins in the network, pointing to a non-random organization of the genes of interest.

Association of clinical phenotypes with dosage-sensitive genes

We conducted a statistical analysis to infer significant gene-phenotype relationships by analyzing clinical data from 3,708 affected individuals with NDDs in the Decipher¹⁸ public database. We calculated the incidence frequencies of each phenotype per gene (number of affected individuals with a certain clinical phenotype and a CNV overlapping a specific gene divided by the total of affected individuals with a CNV affecting such a gene) and compared them with the phenotype frequencies for the Decipher population of affected individuals with NDDs. We then selected those phenotypes whose frequency of incidence in the gene was higher than the frequency in the general population and performed a two-sided Fisher's test to infer whether each gene-phenotype association was statistically significant. By using Bonferroni correction, significance association threshold was established a p value below 1.22×10^{-6} .

Results

Identification of 23 novel NDD-risk CNV loci

Out of the 2,273 genomic loci (also known as regions) with CNV calls analyzed from the SFARI gene database in our meta-analysis, 180 were statistically significantly overrepresented in individuals with NDDs ($p < 0.05$) with an odds ratio higher than 1, and 41 of those passed Bonferroni threshold to strictly correct for multiple testing (Table S1). We hence determined these 41 regions as high-confidence NDD-risk loci. To validate and assess which of these CNV loci were novel, significantly associated regions from our analysis were compared with the ones previously reported by widely recognized reference studies (Sanders et al. 2015¹⁰, Collins et al.,⁸ and regions recognized as relevant in SFARI database). These studies have identified a set of 50 NDD-risk (Table S2) regions after analyzing CNV data from thousands of affected individuals with NDDs and control individuals.

Out of these 41 regions, 24 were novel. These included 19 not previously reported genomic loci as well as 5 loci expanding and/or overlapping with previously known NDD-risk regions (Table S1). For instance, the 8p23.1-p23.3 genomic locus reported in our analysis as a novel region associated to NDDs comprises a large region that expands the 8p23.1 region included in the SFARI NDD-risk CNV catalog, as well as 8p23.2 and 8p23.3 regions associated with NDDs in the Collins et al.⁸ study.

Nevertheless, Bonferroni multiple test correction approach is strict in the identification of region-disorder associations. Considering the signal from regions that do not reach the Bonferroni threshold but pass nominal significance (with $p < 0.05$ and odds ratio >1), we identified 11 additional previously reported risk regions (please see Table S2). Importantly, if we examine a subset of 12 high-confidence NDD-risk regions that were previously reported in at least two studies (i.e., 3q29, 7q11.23, 15q11.2-q13.1, 15q13.2-q13.3, 16p11.2, 22q11.21, 1q21.1, 2p16.3, 15q13.3, 16p12.2, 16p13.3, and 17q12)

Table 1. Number of dosage-sensitive genes located within all/novel NDD-risk regions that are also included in the SFARI catalog of ASD-risk genes

SFARI gene score	No. of qDSGs in SFARI catalog out of total qDSGs	No. of qDSGs in SFARI catalog out of total qDSGs in novel regions
1	25	18
2	14	10
3	14	9
S	10	8
Total	63/294 total qDSGs (57 HI/0 TS/6 TS + HI (258 HI/6 TS/30 TS + HI)	45/197 total qDSGs (40 HI/0 TS/5 TS + HI) (172 HI/4 TS/21 TS + HI)

Score 1 defines high confidence genes, while score 2 labels defines genes with strong association to the phenotype. Score 3 genes present little evidence of association and S genes are considered syndromic.

we identified all of them in our meta-analysis, 11 surpassing the Bonferroni correction threshold, and 1 with nominal significance ($p < 0.05$).

NDD-risk CNV loci are significantly enriched for dosage-sensitive risk genes

We then evaluated the haploinsufficiency and triplosensitivity of genes located in the 41 high-confidence NDD-risk regions identified. Of the 2,097 genes enclosed in these regions, we found a total of 294 qDSGs, from which 258 were HI ($pHI > 0.84$), 6 were TS ($pTS > 0.993$), and 30 were both TS and HI ($pTS > 0.993$ and $pHI > 0.840$). When limiting the scope of the evaluation to the 1,256 genes located in the 24 novel regions, we could identify 172 HI genes ($pHI > 0.840$), 4 showing triplosensitivity ($pTS > 0.993$) and 21 that were both TS and HI ($pTS > 0.993$ and $pHI > 0.840$), totaling 197 qDSGs.

Interestingly, a statistically significant overrepresentation of both HI ($p = 7.32 \times 10^{-3}$) and TS ($p = 1.33 \times 10^{-6}$) genes was found in the 41 NDD-risk regions, although no enrichment was observed for genes showing both haploinsufficiency and triplosensitivity. Similarly, we detect enrichment of HI genes ($p = 4.25 \times 10^{-2}$) and TS genes ($p = 7.31 \times 10^{-5}$) in the novel NDD-risk regions, but no enrichment for bHITS genes was observed either ($p = 2.48 \times 10^{-1}$).

Among the 258 genes that were identified as HI within the NDD-risk regions, 57 HI genes also appear in the SFARI catalog of ASD-risk genes. From the group of 172 HI genes in novel regions, 40 HI genes are also reported as ASD-risk in SFARI (Table 1; Figure 1). This overlap implies a statistically significant overrepresentation of ASD-risk SFARI genes among HI genes in both novel ($p = 2.74 \times 10^{-4}$) and all NDD-risk regions ($p = 4.6 \times 10^{-5}$) identified in this study. We also performed an enrichment analysis with genes reported in recent large-scale ASD-risk SNP association studies to evaluate, from an orthogonal perspective, the enrichment of known ASD-risk genes in our list of qDSGs. We employed the genes reported in two different studies, Satterstrom et al.²² and Trost et al.,²³ and also detected an enrichment of ASD-risk genes in the list of qDSGs ($p = 1.72 \times 10^{-4}$) and also in the subset of those genes that are in novel CNV regions reported in this study ($p = 3.68 \times 10^{-3}$).

On the other hand, we sought to ascertain whether a broader list of known NDD-risk genes was significantly enriched for our qDSGs. Using a recent extensive catalog of genes associated with NDDs (GeneTrek) we could observe that out of the 294 qDSGs, 259 appear in this catalog as candidate NDD-risk genes ($p = 1.25 \times 10^{-64}$), and from the subset of 197 qDSGs found in novel regions, 173 also overlap with this catalog ($p = 5.59 \times 10^{-43}$; Table 2).

Validation of 22 novel NDD-risk qDSGs

To provide further validation of our qDSG list in terms of their association with NDD-risk, we examined whether these genes were also significantly affected by CNVs in affected individuals with NDDs from independent cohorts. To this end, we first conducted a TDT and observed that deletions fully overlapping with HI genes within qDSGs were transmitted to offspring with autism more than expected by chance in 493 multiplex ASD families from the iHART dataset, including both maternally ($p = 4.07 \times 10^{-5}$, probability of success = 0.726 [95% CI, 0.618–0.818]) and paternally ($p = 6.3 \times 10^{-3}$, probability of success = 0.666 [95% CI, 0.545–0.773]) inherited deletions.

We then conducted a burden analysis by performing a Fisher's test using 4,281 affected individuals with NDDs from iHART and Decipher and 2,504 control individuals from the 1000 Genomes cohort, where only rare deletions and duplications from these subjects (with an allelic frequency $< 0.1\%$) were considered. This analysis provided evidence for the association of 162 genes out of 294 qDSGs with risk for NDDs (Bonferroni correction and odds ratio > 1), of which 140 qDSGs were previously described as high confidence or candidate NDD-risk genes and 22 are novel risk genes (Figure 2).

On the other hand, a total of 56 out of the 132 qDSGs for which we did not find association with NDD-risk (i.e., not surpassing multiple test correction threshold) exhibited an association trend with a nominal p value below 0.05 (Table S3).

Implication of qDSGs in biological processes related to neuronal development and function

To assess whether qDSGs may be involved in biological processes essential for brain development and/or function,

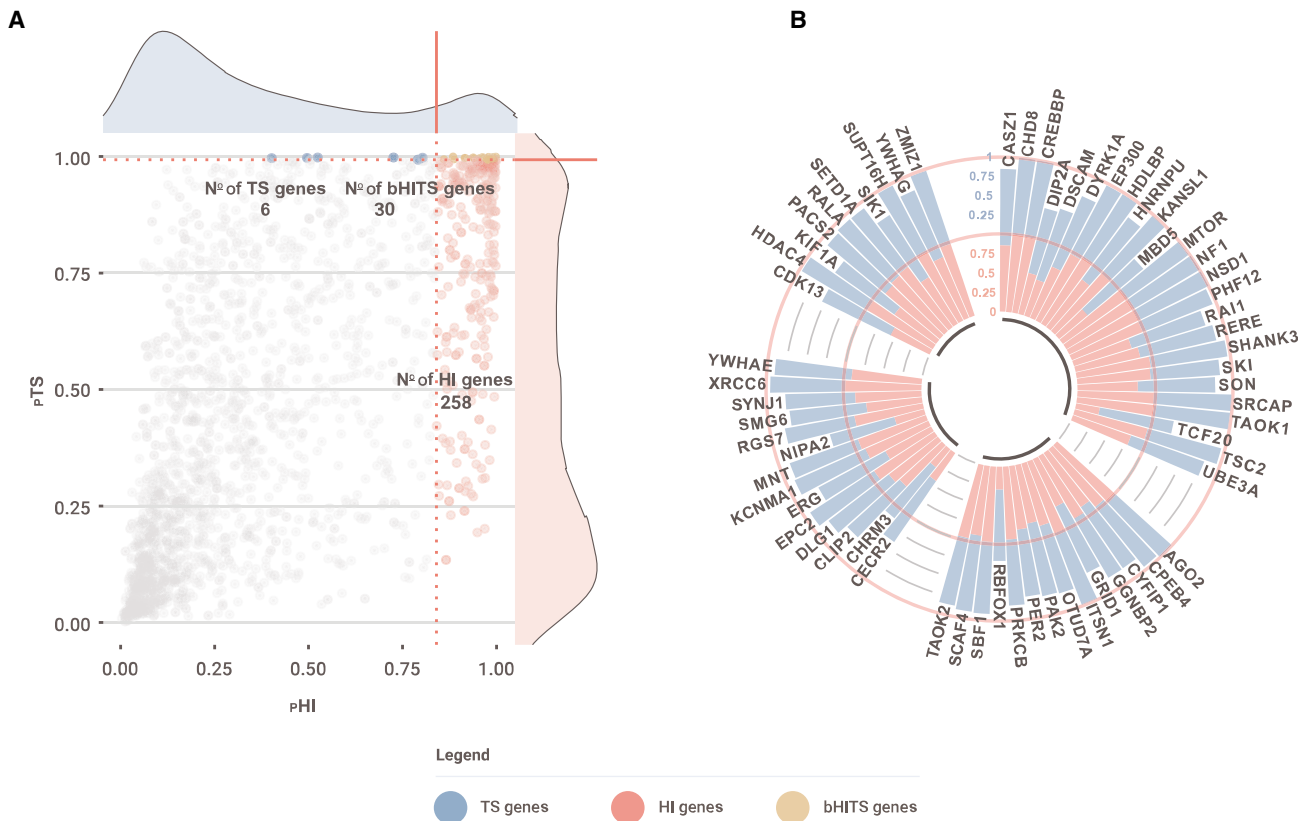


Figure 1. Dosage-sensitive genes in NDD-risk regions overlap with ASD-risk genes

Scatterplot A shows the dose sensitivity score values (pHI and pTS) for genes in regions identified as NDD-risk. Gray dots in the scatterplot stand for genes that do not meet the threshold to be considered HI genes, TS genes, or bHITS genes. Depicted in red are the genes that meet the threshold for haploinsufficiency, HI genes, and in blue the ones that meet the threshold for triplosensitivity, TS genes. Genes that surpass the threshold for both haploinsufficiency and triplosensitivity, bHITS genes, are depicted in yellow in the upper right quadrant of the grid. The circular bar plot (B) represents the 63 dosage-sensitive genes that are present in both our list of qDSGs and the SFARI curated catalog of ASD-risk genes, according to the degree of evidence of association.

we obtained functional information from the enrichment of GO terms, i.e., over-represented terms for each of the three gene lists according to its dosage-sensitive type.

We identified a total of 384 unique GO terms significantly (p value adjusted after Bonferroni correction < 0.05) enriched in qDSGs (Figure 3; Table S4). Taking into account genes in novel NDD-risk regions, we found 297 significantly enriched GO terms in qDSGs enclosed in those regions (Table S5). In both cases strong statistical enrichment is observed for terms such as neuronal morphology, synapses, signaling pathways, and control of gene expression (Figures 3A–3C). This association is observed at the level of cellular components, molecular functions, and biological processes. For example, of the 391 terms related to qDSGs in all NDD-risk regions, 39 GO terms refer to synaptic processes associated with 64 qDSGs, of which 2 are novel NDD-risk candidate genes, while 19 are known high-confidence genes, i.e., with an SFARI score 1 or present in the high confidence curated list of NDD-risk genes mentioned above. We also found 29 GO terms associated with morphology or development of neuronal elements such as dendrites and axons involving 57 qDSGs, of which 54 had been previously associated with NDDs by reference cat-

alogs. The identification of enriched terms associated with chromatin regulatory processes is in agreement with previous literature reporting the crucial role of chromatin remodelers during corticogenesis, cell migration, and cell-type specification.^{24,25} In particular, we have found 44 GO terms associated with processes related to histone and chromatin modification, as well as DNA binding, associated with 80 genes from our dosage-sensitive gene list, of which 9 are novel NDD-risk candidate genes not described before.

Importantly, we performed the same GO enrichment analysis using all 1,031 genes from the curated SFARI catalog and found a significant overlap between terms enriched for SFARI genes and those identified from our qDSG list, including a significant overlap for terms related to HI genes ($p = 8.14 \times 10^{-292}$; Figure 3D), for TS genes ($p = 1.26 \times 10^{-8}$; Figure 3E) and for bHITS genes ($p = 6.56 \times 10^{-46}$; Figure 3F).

qDSGs are among the highest-expressed genes in brain along all developmental stages

Previous studies in affected individuals with NDD have detected aberrant expression of genes involved in pathways enriched in the GO analysis (synaptic formation, dendrite

Table 2. NDD-risk regions identified through our meta-analysis and the 294 corresponding qDSGs encompassed in them according to their dosage sensitivity type

CNV	Both TS and HI	HI	TS	CNV type	Chromosome
1p36.33-p36.22 (novel)	<i>CAMTA1**+</i> , <i>MTOR**+</i>	<i>SPSB1+</i> , <i>TARDBP**+</i> , <i>SKI**+</i> , <i>LRRC47+</i> , <i>TP73**+</i> , <i>ICMT+</i> , <i>PARK7**+</i> , <i>TMEM201**+</i> , <i>CLSTN1**+</i> , <i>PIK3CD**+</i> , <i>UBE4B**+</i> , <i>KIF1B**+</i> , <i>PGD**+</i> , <i>MFN2**+</i> , <i>GNB1**+</i> , <i>AJAPI**+</i> , <i>CHD5**+</i> , <i>DNAJC11**+</i> , <i>RERE**+</i> , <i>CASZ1**+</i> , <i>PRDM16**+</i>	N/A	deletion	1
1p36.33-p36.32 (novel)	N/A	<i>SKI**+</i> , <i>LRRC47+</i> , <i>TP73**+</i> , <i>GNB1**+</i> , <i>AJAPI**+</i> , <i>PRDM16**+</i>	N/A	deletion	1
1q21.1	N/A	<i>TXNIP</i> , <i>PIAS3*</i>	N/A	deletion-duplication	1
1q43-q44	<i>AKT3**+</i>	<i>HEATR1*</i> , <i>DES12+</i> , <i>HNRNPU**+</i> , <i>AHCTF1**+</i> , <i>ZNF496**+</i> , <i>ACTN2**+</i> , <i>RYR2</i> , <i>CHRM3*</i> , <i>ZBTB18**+</i> , <i>KIF26B*</i> , <i>FMN2**</i> , <i>RGS7*</i>	N/A	duplication	1
2q23.1 (novel)	N/A	<i>MBD5**</i> , <i>EPC2*</i>	N/A	deletion-duplication	2
2q37.3 (novel)	<i>HDLBP**+</i>	<i>COPS8+</i> , <i>COL6A3**+</i> , <i>ILKAP**+</i> , <i>PER2**+</i> , <i>SNED1**+</i> , <i>KIF1A**+</i> , <i>HDAC4**+</i>	N/A	deletion-duplication	2
3q22.1 (novel)	<i>PLXND1*</i>	<i>TMCC1</i> , <i>PIK3R4</i> , <i>CDV3</i> , <i>TOPBP1*</i> , <i>RAB6B*</i> , <i>ATP2C1*</i> , <i>DNAJC13*</i> , <i>TMEM108*</i>	N/A	deletion-duplication	3
3q29	N/A	<i>HES1+</i> , <i>ATP13A3*</i> , <i>UBXN7**+</i> , <i>PAK2**+</i> , <i>RPL35A*</i> , <i>OPA1*</i> , <i>ACAP2**+</i> , <i>SENP5**+</i> , <i>DLG1**+</i>	N/A	deletion-duplication	3
4p16.3 (novel)	<i>CTBPI**+</i>	<i>PCGF3**+</i> , <i>FAM193A**+</i> , <i>ADD1**+</i> , <i>HTT**+</i>	N/A	deletion-duplication	4
5q35.2-q35.3 (novel due to novel region boundaries)	<i>HNRNPH1**+</i> , <i>CANX**+</i> , <i>UNC5A**+</i>	<i>NSD1**</i> , <i>DBN1*</i> , <i>FAF2*</i> , <i>HNRNPAB*</i> , <i>MAML1*</i> , <i>MGAT1*</i> , <i>TRIM41*</i> , <i>CPEB4*</i> , <i>FLT4*</i> , <i>ADAMTS2*</i>	N/A	duplication	5
6p25.3 (novel)	N/A	<i>FOXF2+</i> , <i>IRF4**+</i> , <i>GMD5*</i>	N/A	deletion-duplication	6
6p25.3-p25.1 (novel)	N/A	<i>NRN1**+</i> , <i>FOXF2+</i> , <i>CDYL*</i> , <i>IRF4**+</i> , <i>PRPF4B**+</i> , <i>SLC22A23</i> , <i>GMD5*</i>	N/A	deletion	6
7p14.1 (novel)	N/A	<i>GLI3**</i> , <i>RALA**</i> , <i>INHBA*</i> , <i>PSMA2*</i> , <i>CDK13**</i>	N/A	deletion-duplication	7
7q11.23	N/A	<i>BAZ1B**+</i> , <i>LIMK1**+</i> , <i>HIP1*</i> , <i>YWHAG**</i> , <i>CLIP2**+</i> , <i>RSBN1L</i> , <i>PTPN12*</i>	N/A	deletion-duplication	7
8p23.1	N/A	<i>TNKS**+</i> , <i>XKR6**+</i> , <i>LONRF1+</i> , <i>ANGPT2+</i>	N/A	deletion-duplication	8
8p23.3-p23.1 (novel due to novel region boundaries)	N/A	<i>TNKS**+</i> , <i>XKR6**+</i> , <i>LONRF1+</i> , <i>ANGPT2+</i>	N/A	deletion	8
8q24.3 (novel)	<i>AGO2**+</i> , <i>PUF60**+</i>	<i>PTK2*</i> , <i>FAM83H</i> , <i>RPL8*</i> , <i>SCRIB*</i>	<i>CYHR1+</i>	deletion-duplication	8
10q22.3-q23.2 (novel)	N/A	<i>KCNMA1**</i> , <i>RPS24*</i> , <i>TSPAN14*</i> , <i>CCSER2</i> , <i>DLG5*</i> , <i>ZCCHC24*</i> , <i>ZMIZ1**</i> , <i>BMPRIA</i> , <i>GRID1*</i>	N/A	deletion-duplication	10
14q11.2 (novel)	<i>ZNF219**+</i> , <i>SUPT16H**+</i> , <i>CHD8**+</i> , <i>ACIN1*</i>	<i>PRMT5*</i> , <i>MMP14*</i> , <i>TOX4*</i>	<i>JPH4*</i>	deletion-duplication	14
14q32.33	<i>MTA1**+</i>	<i>INF2**+</i> , <i>AKT1**+</i> , <i>CEP170B**+</i> , <i>JAG2**+</i> , <i>KLC1**+</i> , <i>PACS2**+</i> , <i>PPP1R13B**+</i>	N/A	deletion-duplication	14
15q11.2 (novel due to novel region boundaries)	N/A	<i>UBE3A**+</i> , <i>CYFIP1**+</i> , <i>NIPA2**+</i>	N/A	deletion-duplication	15
15q11.2-q13.1	N/A	<i>UBE3A**+</i> , <i>CYFIP1**+</i> , <i>NIPA2**+</i> , <i>TJP1**+</i>	N/A	duplication	15
15q13.2-q13.3	N/A	<i>OTUD7A**+</i>	N/A	deletion-duplication	15

(Continued on next page)

Table 2. Continued

CNV	Both TS and HI	HI	TS	CNV type	Chromosome
15q13.3	N/A	<i>OTUD7A</i> *+	N/A	deletion-duplication	15
16p11.2	<i>FBXL19</i> *	<i>FBR</i> *, <i>FUS</i> *, <i>SH2B1</i> *+, <i>TAOK2</i> *+, <i>TBC1D10B</i> *, <i>SRCAP</i> **+, <i>SETD1A</i> **+, <i>ATXN2L</i> *+	<i>MAZ</i> *+, <i>STX1B</i> **	deletion-duplication	16
16p12.1 (novel)	N/A	<i>RBBP6</i> *, <i>ARHGAP17</i> , <i>IL21R</i> *, <i>TNRC6A</i> *, <i>GTF3C1</i> *, <i>XPO6</i> *	N/A	deletion-duplication	16
16p12.2-p11.2 (novel due to novel region boundaries)	<i>FBXL19</i> *	<i>PLK1</i> *, <i>FBR</i> *, <i>FUS</i> *, <i>RBBP6</i> *, <i>ARHGAP17</i> , <i>IL21R</i> *, <i>SH2B1</i> *+, <i>TAOK2</i> *+, <i>TBC1D10B</i> *, <i>SRCAP</i> **+, <i>SETD1A</i> **+, <i>TNRC6A</i> *, <i>GTF3C1</i> *, <i>XPO6</i> *, <i>ATXN2L</i> *+, <i>PRKCB</i> *, <i>USP31</i>	<i>MAZ</i> *+, <i>STX1B</i> **	duplication	16
16p13.11	N/A	<i>MYH11</i> +	N/A	deletion-duplication	16
16p13.3	<i>MAPK8IP3</i> **+	<i>RPS2</i> *, <i>CAPN15</i> **+, <i>TSC2</i> **+, <i>CASKIN1</i> *, <i>TFAP4</i> *, <i>GLYR1</i> *, <i>UBN1</i> *, <i>ADCY9</i> *, <i>CREBBP</i> **+, <i>RFXO1</i> *+	<i>UBE2I</i> *+	deletion-duplication	16
17p11.2	N/A	<i>MPRIP</i> *+, <i>COPS3</i> *+, <i>RAI1</i> **+, <i>ALKBH5</i> *+, <i>EPN2</i> +, <i>USP22</i> *+	<i>SREBF1</i> +	deletion-duplication	17
17p13.3 (novel)	<i>PRPF8</i> *+, <i>CLUH</i> *+	<i>NXN</i> *+, <i>ABR</i> *+, <i>YWHA</i> E*+, <i>PITPNA</i> *+, <i>MNT</i> *+, <i>PAFAH1B1</i> **+, <i>CRK</i> *+, <i>RTN4RL1</i> *+, <i>METTL16</i> +, <i>RAP1GAP2</i> *+, <i>SMG6</i> *+	N/A	deletion-duplication	17
17q11.2	<i>NLK</i> *, <i>GIT1</i> *, <i>SUPT6H</i> *, <i>NFI</i> **+	<i>TRAF4</i> *, <i>SSH2</i> *, <i>PHF12</i> **+, <i>NUFIP2</i> *, <i>ATAD5</i> *, <i>RHOT1</i> , <i>PSMD11</i> *, <i>MYO18A</i> *, <i>TAOK1</i> **	N/A	deletion-duplication	17
17q12	<i>ACACA</i> *+, <i>PIP4K2B</i> *	<i>AP2B1</i> *, <i>GGNBP2</i> *+, <i>MLLT6</i> *, <i>PCGF2</i> **+, <i>SYNRG</i> *+, <i>SRCIN1</i> *, <i>MED1</i> *, <i>CDK12</i> *, <i>SOC7</i> *, <i>FBXL20</i> *	N/A	deletion-duplication	17
17q21.31 (novel)	<i>UBTF</i> **+, <i>HDAC5</i> *, <i>EFTUD2</i> **	<i>RPL27</i> *, <i>FMNL1</i> *, <i>GPATCH8</i> *, <i>KANSL1</i> **+, <i>NMT1</i> *	N/A	duplication	17
21q11.2-q22.3 (novel)	N/A	<i>U2AF1</i> *+, <i>NRIP1</i> *, <i>BTG3</i> *, <i>GABPA</i> *, <i>ADAMTS1</i> *, <i>CCT8</i> *, <i>BACH1</i> , <i>TIAM1</i> *+, <i>CHAF1B</i> *, <i>MORC3</i> *, <i>ETS2</i> *+, <i>BRWD1</i> *+, <i>ZBTB21</i> *+, <i>SIK1</i> **+, <i>ADARB1</i> **+, <i>COL6A1</i> *+, <i>USP25</i> *, <i>APP</i> *, <i>LTN1</i> *, <i>SCAF4</i> **+, <i>PAXBPI</i> *, <i>SYNJ1</i> **+, <i>ITSN1</i> **+, <i>SLC5A3</i> *, <i>RUNX1</i> +, <i>DYRK1A</i> **+, <i>TTC3</i> , <i>ERG</i> *+, <i>DSCAM</i> **+, <i>DIP2A</i> **+, <i>PKNOX1</i> *+, <i>TRAPPC10</i> *+, <i>SON</i> **	N/A	duplication	21
22q11.21	N/A	<i>CECR2</i> *+, <i>MICAL3</i> *+, <i>DGCR8</i> *+, <i>SCARF2</i> *+, <i>UBE2L3</i> +, <i>HIRA</i> *+, <i>MED15</i> *+	N/A	deletion-duplication	22
22q11.21-q11.23	N/A	<i>CECR2</i> *+, <i>MICAL3</i> *+, <i>DGCR8</i> *+, <i>SCARF2</i> *+, <i>UBE2L3</i> +, <i>MAPK1</i> **+, <i>SMARCB1</i> **+, <i>HIRA</i> *+, <i>MED15</i> *+, <i>BCR</i> *+, <i>KIAA1671</i> +	N/A	deletion	22
22q13.2-q13.33 (novel due to novel region boundaries)	<i>PLXNB2</i> *+	<i>PIM3</i> *+, <i>SHANK3</i> **+, <i>ST13</i> +, <i>RBX1</i> *+, <i>ZC3H7B</i> *+, <i>ACO2</i> **+, <i>TOB2</i> +, <i>XRCC6</i> *+, <i>SREBF2</i> +, <i>FBLN1</i> *+, <i>GRAMD4</i> *+, <i>MAPK8IP2</i> *+, <i>EP300</i> **+, <i>TCF20</i> **+, <i>SCUBE1</i> *+, <i>PHF21B</i> *+, <i>CELSR1</i> *+, <i>BRD1</i> *+, <i>SBF1</i> *+	N/A	deletion-duplication	22

*Candidate and high confidence. **NDD-risk genes according to the GeneTrek catalog. +Genes validated through our burden analysis surpassing Bonferroni correction. N/A, not applicable, no dosage sensitive genes have been found in this region.

and axon morphology, transcriptional regulation, or chromatin remodeling), so we assessed longitudinal and regional gene expression in neurotypical brains. Using RNA-seq data from Brain Span brain samples, and after a

chordate standardization of the counts, we obtained a distribution of gene expression across all developmental stages in neocortex, amygdala, and cerebellum. Remarkably, we found that the top 5% most expressed genes of

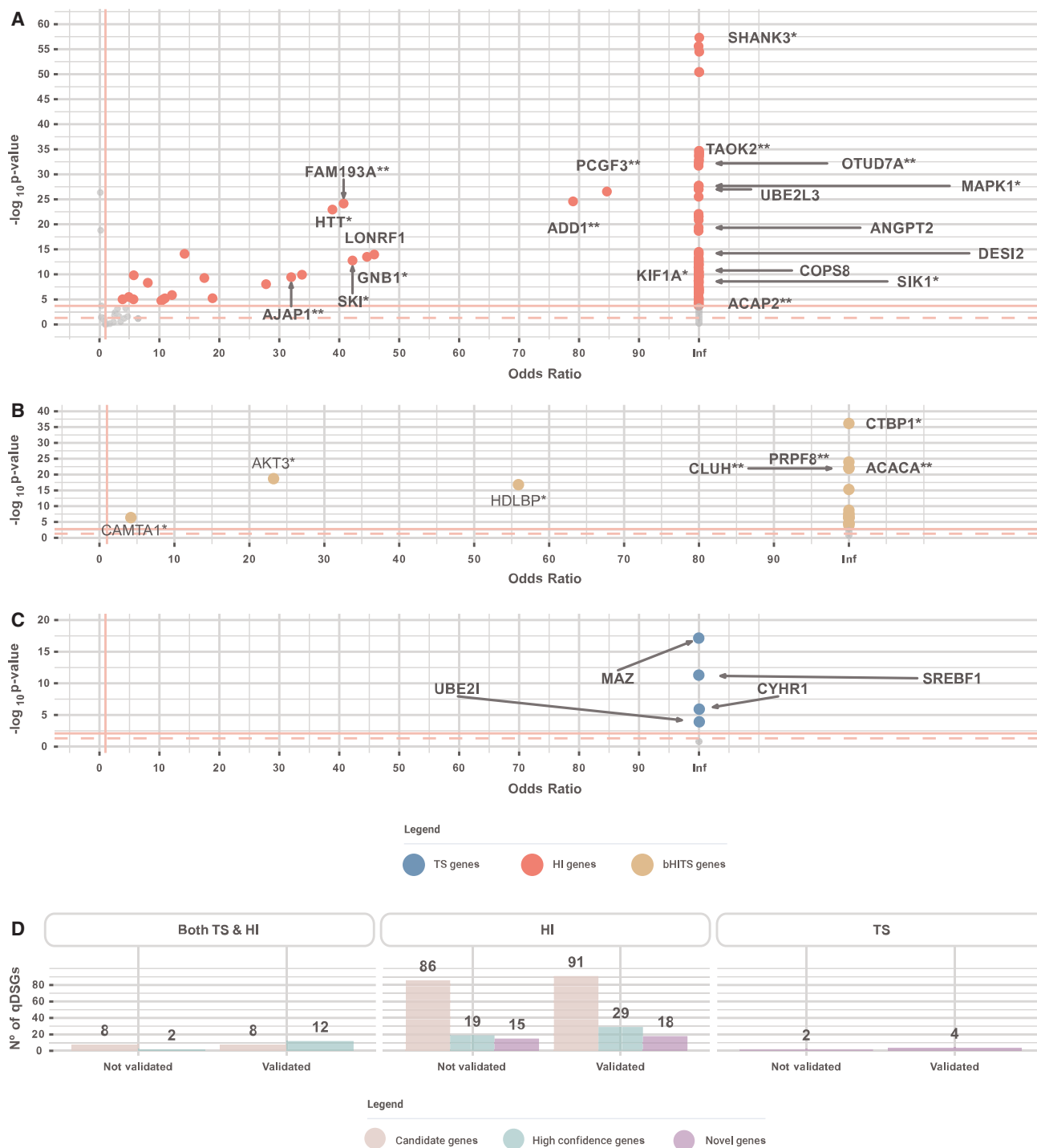


Figure 2. Validation of NDD-risk for 162 qDSGs, 22 of them novel genes never associated before with risk for NDD

Validation of qDSG in NDD-risk regions found associated to autism in affected individuals from iHART and Decipher compared with control individuals from 1000 Genomes.

(A–C) Volcano plot depicting the p values and odds ratios for qDSGs genes in NDD-risk regions: (A) HI genes, (B) bHITS genes, and (C) TS genes. Vertical red line delimits odds ratio 1, whereas horizontal red line stands for the multiple test correction threshold. Dashed red line represents $p = 0.05$. (D) Bar plot describing the number of qDSG validated according to type of novelty. *Candidate and high confidence, **NDD-risk genes according to the GeneTrek catalog.

this distribution were enriched for qDSGs including HI genes, TS genes, and bHITS genes (Figures 4, S1, and S2; Table S6). Furthermore, this enrichment could be observed throughout all stages of development up to adulthood in

key brain regions such as the neocortex regions, the amygdala, and the cerebellum.

Interestingly, we observed that, among the 87 qDSGs in our list that are highly expressed in any of the tissues

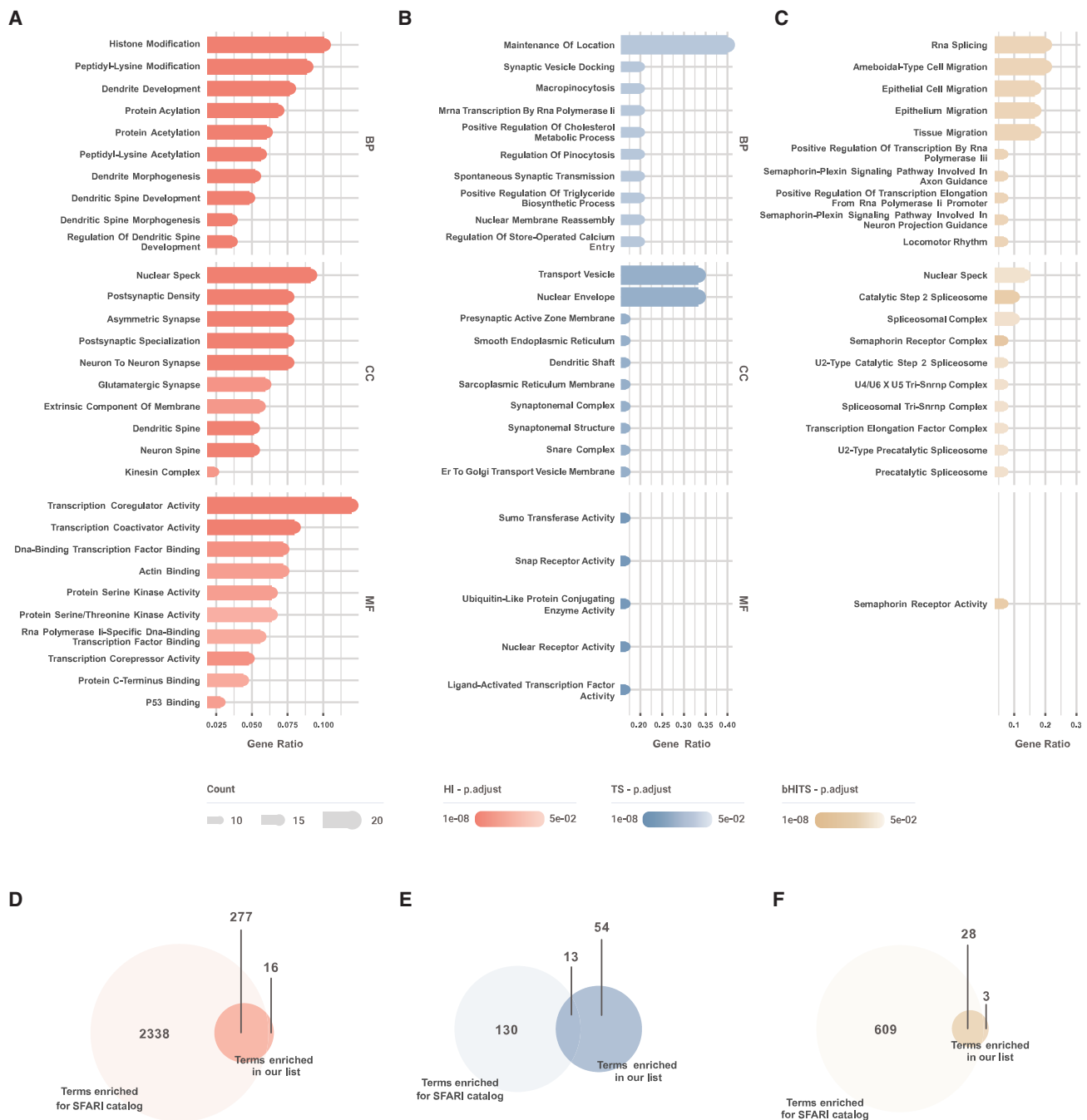


Figure 3. Pathways where qDSGs are involved, and their overlap with known NDD-risk pathways

Subset of top GO terms enriched for HI genes (A), for TS genes (B), and for bHITS genes (C). In the lower half is depicted the overlap between the GO terms enriched for our qDSGs and the GO terms enriched for the genes in the SFARI catalog of ASD-risk genes, also divided according to their dosage type, for HI genes (D), for TS genes (E), and for bHITS genes (F).

assessed (top 5% percentile), 36 of them remain in this percentile category from fetal age to adulthood, of which 8 have already been reported to be associated with autism and other NDDs (*CLIP2*, *HDLBP*, *HNRNPU*, *KIF1A*, *SBF1*, *YWHAH*, *YWHAQ*, *PRKCB*; Figures S3 and S4; Table S7). In addition, these genes are related to biological processes previously associated with abnormal neurodevelopment, such as histone modification, dendrite development, synaptic transmission, muscle tissue development, semaphorin-

plexin signaling pathway, or regulation of neural projection (Table S8).

qDSGs converge at PPI networks

PPI enrichment was evaluated for the 294 proteins encoded by qDSGs using STRING. Following a conservative approach for this protein interaction analysis, we created the PPI networks using only strong interaction evidence from databases and experiments. This analysis confirmed

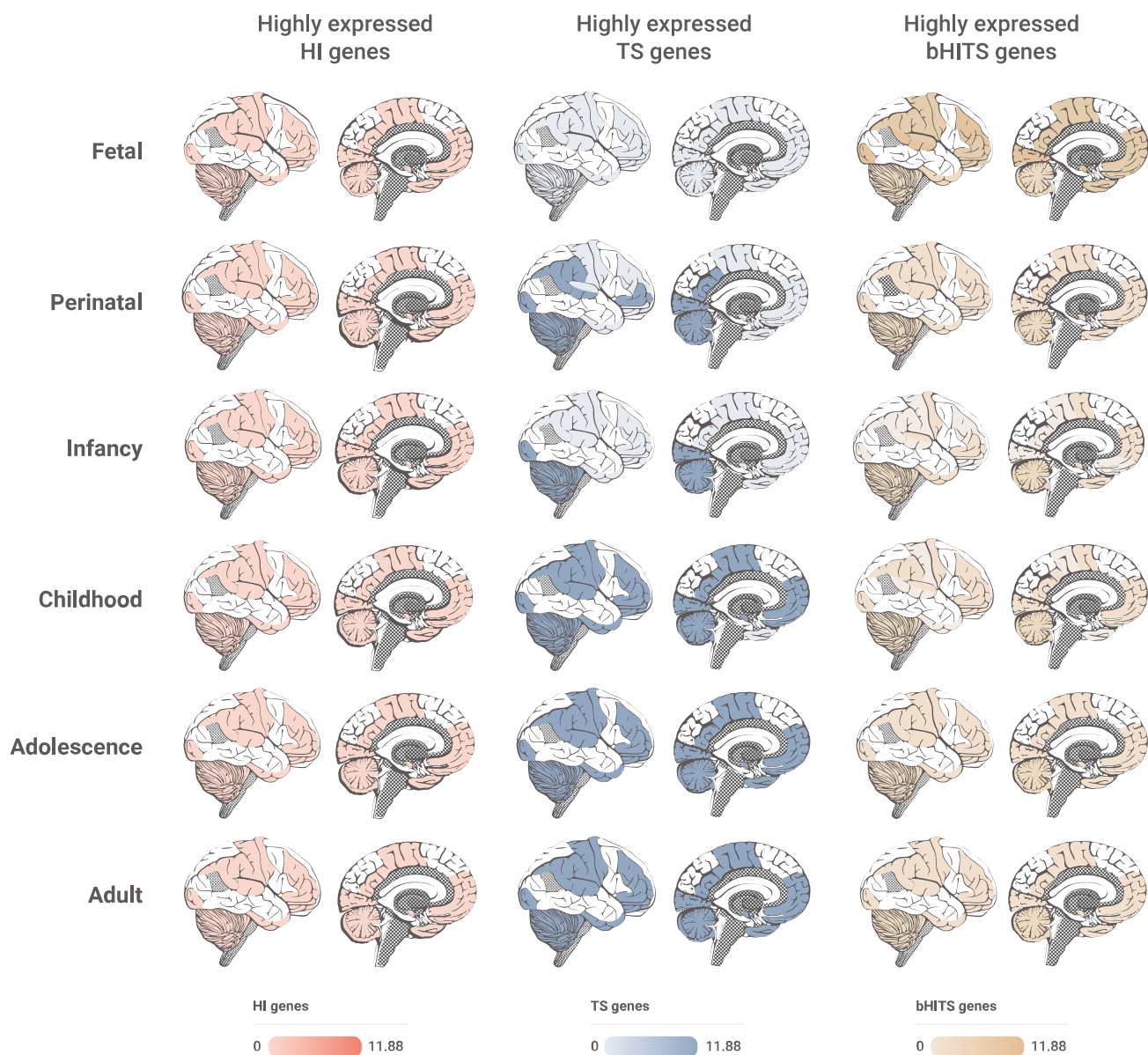


Figure 4. qDSGs are significantly highly expressed in key areas of the brain across all developmental stages

Surface and slice representation of odds ratio resulting from the enrichment analysis of qDSGs in the highly expressed ranking. Only statistically significant areas enriched with highly expressed qDSGs are colored.

a statistically significant enrichment in the interaction of the protein products of qDSGs, (expected/observed number of edges = 286/328; $p = 6.95 \times 10^{-4}$, average node degree = 2.56; cluster coefficient = 0.351; Figure 5; Table S9).

Interestingly, a high degree of direct and/or indirect interaction (distance of 1 and 2 nodes in the network, respectively) of proteins encoded by qDSGs that are located within the same NDD-risk CNV region was also found. In particular, this was observed for 13 out of the 36 (36.11%) NDD-risk CNV regions reported in this study that have at least one qDSG. For example, for the 16p13.3 region, among the 12 qDSGs identified within this region, 5 are indirectly connected and two are directly connected (Table S10). This is consistent with the fact that a higher probability of interaction is expected for genes that are close

to each other on the same chromosome, due to the three-dimensional organization of the genome.²⁶ Noteworthy, a trend toward higher degree of protein connectivity is also observed when randomly selecting a high-confidence NDD-risk qDSG from each NDD-risk region (expected/observed number of edges = 3/6; $p = 0.08$, average node degree = 0.5; cluster coefficient = 0.333), indicating functional convergence also across NDD-risk regions reported in this study.

qDSGs are associated with the manifestation of specific clinical phenotypes

We finally investigated whether deletion or duplication of qDSGs are associated with the manifestation of specific phenotypes in carrier affected individuals. Using Decipher

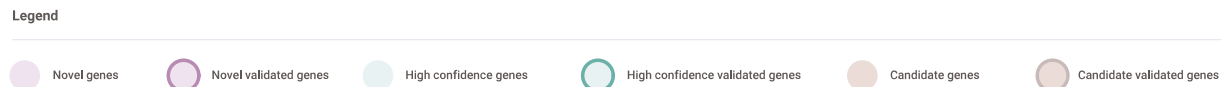


Figure 5. PPI network depicting interactions between protein products of qDSGs
Protein-protein interaction network depicting interactions of qDSGs where nodes are colored based on their evidence strength of NDD-risk. Circled nodes stand for validated NDD-risk for those genes in our analysis.

data, we found a total of 1,786 unique phenotypes for 3,708 affected individuals with deletions or duplications affecting qDSGs. Out of the 40,838 evaluated gene-phenotype relationships, a total of 1,761 were found to be statistically significant surpassing Bonferroni multiple test correction ($p < 1.22 \times 10^{-6}$). Interestingly, while 10.39% (183) of these gene-phenotype relationships can also be found reported in the Human Phenotype Ontology (HPO) database, 89.61%

(1,578) of them are novel. Besides, we found that the phenotypes significantly associated with the highest number of qDSGs (Figure 6A) are common clinical signs and symptoms reported in affected individuals with NDDs, such as hypertelorism, hypoplasia of the corpus callosum, lissencephaly, or hypotonia. We also found that most of the genes with the highest number of significant gene-phenotype associations were candidate or high-confidence

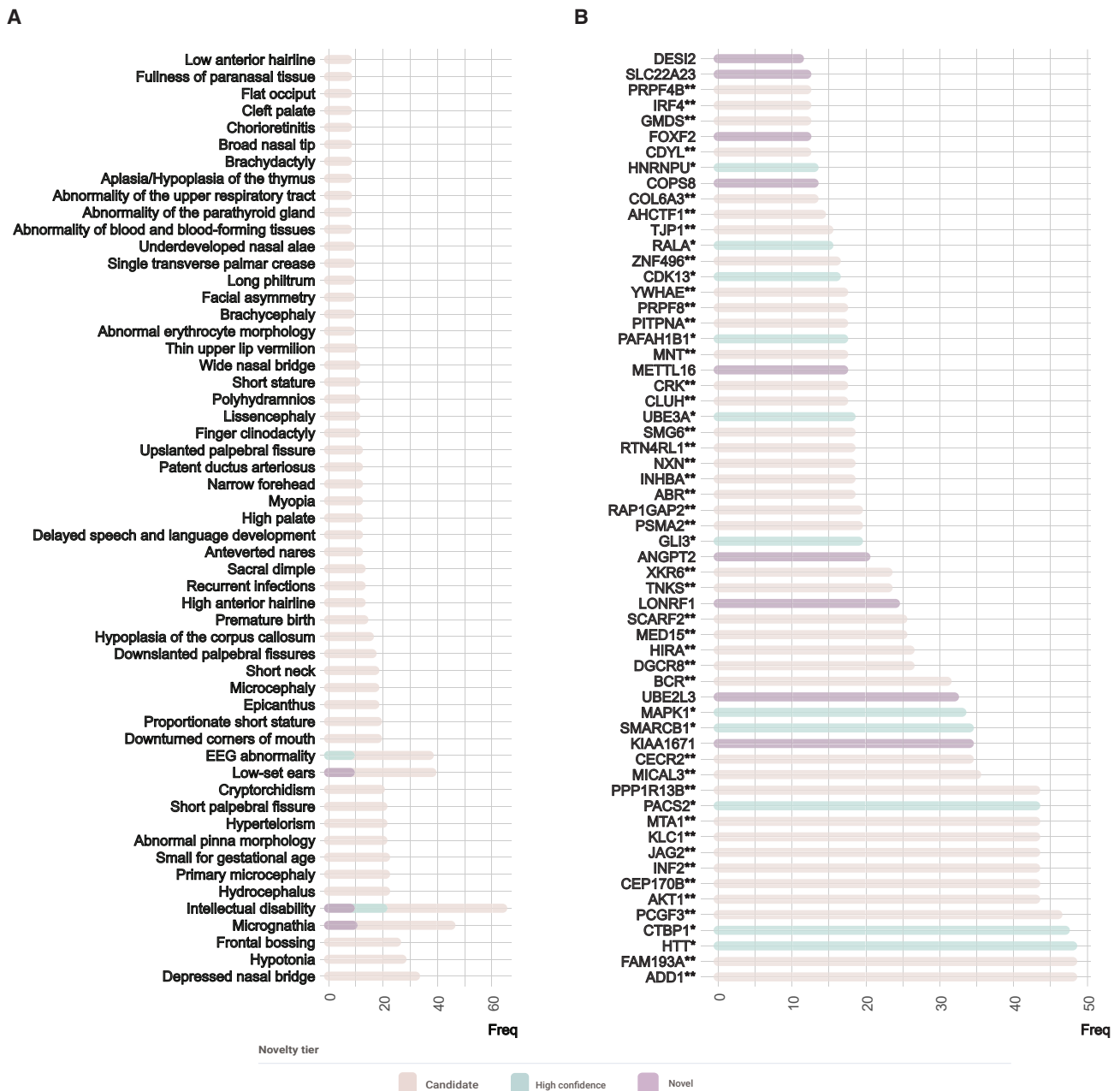


Figure 6. Phenotypes associated to most qDSGs are NDD-risk clinical signs and symptoms

(A) Histogram listing the top 60 HPO terms associated with NDD-risk genes, according to the type and number of genes they are associated to.

(B) Histogram listing the top 60 NDD-risk genes, according to the number of HPO phenotypes they are associated.

genes (Figure 6B), although some of the novel validated genes such as *COPS8* were also associated with a high number of statistically significant gene-phenotype relationships (see Table S3 for further details on the specific HPO terms associated to these genes).

Discussion

Large and rare CNVs are associated with risk for a wide range of NDDs and contribute to the shared etiology of disorders

such as autism (ASD), schizophrenia (SCZ), attention deficit hyperactivity disorder (ADHD), obsessive-compulsive disorder (OCD), bipolar disorder, or intellectual disabilities (ID). By conducting a comprehensive meta-analysis of available datasets, we have validated the association of 17 previously known NDD-risk regions and found 24 novel NDD-risk regions, of which 5 partially overlap with previously known NDD-risk regions.

Interestingly, for several NDD-risk regions that we identified, deletions and/or duplications have been described in association with more than one NDD, such as ASD, ADHD,

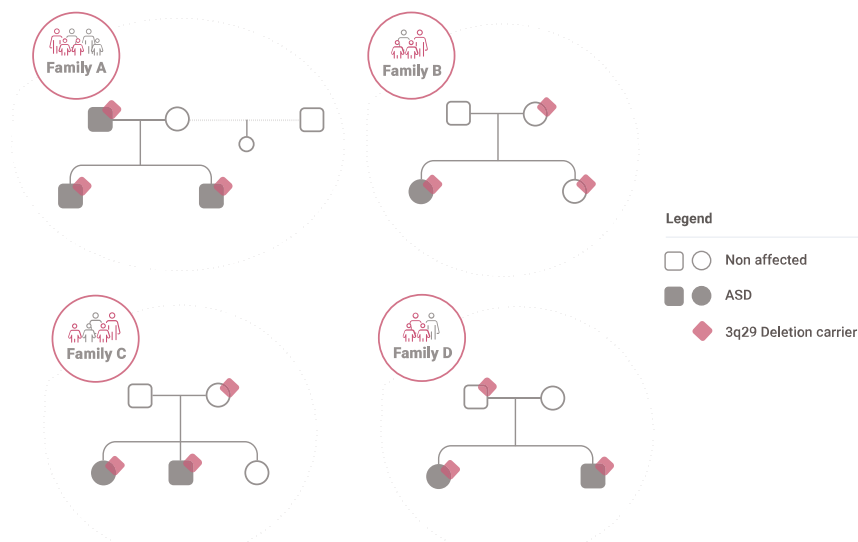


Figure 7. Pedigree for deleted CNVs on 3q29, affecting qDSG ACAP2, which was identified in four different families with a transmission rate of 1

SCZ, or ID (see [supplemental information*¹](#)). For example, reciprocal structural variants (duplications and deletions) at region 1q21 are found at high incidence in ASD, SCZ, ADHD, OCD, ID, or epilepsy. Although there is wide phenotypical variation among affected individuals with a CNV in these region,^{27,28} duplication events are more associated with autism, while high rates of deletions are shown to be more related with ID, SCZ, and epilepsy.²⁹ Besides, ADHD risk has been associated to affected individuals carrying both deletions and duplications, although affected individuals with a 1q21.1 deletion have higher scores for ADHD traits.^{30,31} The 15q locus is another highly complex region of the human genome, being the second most common region affected by duplications in affected individuals with autism.³² Duplications in subregions such as 15q11-q13, 15q13.2-q13.3, or 15q13.3 also show a high range of clinical heterogeneity, and beyond autism, a risk association signal is also found for affected individuals with ADHD,^{33,34} SCZ,^{35–38} or OCD.^{39–41} This fact also highlights shared etiologies among these heterogeneous groups of disorders and supports the rationale of grouping affected individuals with NDDs for variant and gene association analysis as done in the current study.⁴² Thus, even though most affected individuals in our study are diagnosed with ASD, the findings reported here are overall related with risk for abnormal brain development, which characterizes affected individuals with NDDs. Consistently, for all high-confidence NDD-risk regions identified in this study we find affected individual CNV carriers holding diagnoses of developmental delay (Table S1).

Substantial evidence indicates that the mechanism through which these disease phenotypes are triggered is the dosage sensitivity of genes affected by CNV. These CNVs are known to affect broad regions of the genome, and pinpointing which genes are responsible for causing the phenotypic profile remains a challenge. Through our results, we confirmed previous evidence that indicates that dosage variation is a key mechanism leading to this

type of disorder, since NDD-risk regions, both novel and previously reported, were significantly enriched for dosage-sensitive genes previously associated with risk for ASD and NDDs.

In addition, we observed that deletion variants affecting HI genes enclosed in the NDD-associated regions are more transmitted to offspring with an ASD diagnosis than expected by chance. Furthermore, HI and/or TS genes within NDD-risk regions

(qDSGs) were found to be significantly involved in biological process or functions disrupted in affected individuals with NDDs, such as synapsis, dendrite morphology, axon guidance, histone/RNA binding or semaphorin pathway, among others. Many of these pathways are essential in the central nervous system throughout an individual's entire life span. Linked to this fact, we found a significant overrepresentation of these dosage-sensitive genes in the top 5% percentile of more expressed genes across all stages of development in brain tissue. In particular, we found that 36 qDSGs are highly expressed in the brain from the fetal stage until adulthood, and they are involved in regulation of core biological pathways, such as mitochondrion organization, histone modification, selective molecular interaction, or RNA splicing, as well as in specific nervous system processes, such as neuromuscular control, dendrite development, or modulation of synaptic transmission. These findings provide evidence that mechanisms of gene dosage dysfunction during brain development can persist in adult individuals, which suggests that some targeted pharmacological therapies might still be beneficial in adulthood.

In addition, in independent cohorts to those used for the meta-analysis, gene-NDD-risk association was validated for 162 out of 294 qDSGs, including 22 novel NDD-risk genes. Importantly, multiple NDD-risk genes were identified and/or validated per CNV for most of the NDD-risk CNVs identified in this study, which provides additional evidence for the pathophysiology of NDDs caused by large and rare CNVs not necessarily fitting into a monogenic model. This complex interaction model where candidate genes within each region interact with each other to influence a broad clinical profile is also ratified by the high connectivity we have observed in the PPI network, which also applies for genes within the same risk regions. Such a multi-genetic pathogenicity model is consistent with a complex genetic architecture of NDDs, in which hundreds of genes with varying size effects and complex interactions exert an impact on developmental traits.

Among the 22 novel NDD-risk genes identified in this study, there is the HI gene *COPS8*, located in novel risk region 2q37.3. This gene is involved in both MAPK cascade regulation and histone modification, and it interacts in the described PPI network with genes previously reported in association with autism, such as *COPS3*, *RBX1*, and *PSMD11*. Deletions in the 2q37.3 region have been previously reported in affected individuals with autism⁴³ and other NDDs^{44,45} but *COPS8* had never been associated as a critical gene for this region. Another identified novel NDD-risk gene is *DESI2*, located in the previously known NDD-risk region 1q43-q44. *DESI2* regulates the modification of proteins by small ubiquitin-like modifier which is an essential posttranslational regulatory mechanism whose defects have been proven to have an important influence in axonal trafficking, neural plasticity, brain development, and neurodegeneration.⁴⁶ We found several NDD-related clinical signs and symptoms as significantly associated to this gene, such as EEG abnormality, epicanthus, micrognathia, prominent metopic ridge, microcephaly, or hypoplasia of the corpus callosum (Table S3).

On the other hand, we also validated candidate genes that had not been established yet as high-confidence NDD-risk genes. For instance, we confirmed *ACAP2* as a high-confidence NDD-risk gene located in the 3q29 region, which was previously well known to be associated with risk for different NDDs, but for which the driver gene had not been yet identified. We found deletions of this gene to be transmitted to all the children with ASD in four multiplex families from iHART (Figure 7). Animal knockout models have shown that *ACAP2* is required to promote tunneling nanotubes and vesicle trafficking,⁴⁷ cytoskeletal structures that have been shown to be essential in neuronal morphology, including axon growth and guidance, dendrite development and plasticity, and synapse formation (see supplemental information^{*2}).

This study also provides additional association evidence for genes that have been extensively reported in the literature as high-confidence NDD-risk genes, such as *MTOR*. Pharmacological inhibition of mTOR has been described as effective in rescuing synaptic surplus and functional hyperconnectivity observed in some affected individuals with autism.⁴⁸ *MTOR* is known to be involved in reward/aversion processing as well as in positive/negative reinforcement learning, both circuits being disrupted in neuropsychiatric disorders (SCZ, ADHD, ASD, Tourette syndrome, conduct disorder/oppositional defiant disorder, Fragile X syndrome, Prader-Willi syndrome, Williams syndrome, Angelman syndrome, and Rett syndrome).^{49,50}

By conducting validation of the findings in independent cohorts and through multiple gene set enrichment analyses we have mitigated limitations of the current study, including: (1) scarcity of information about potential confounding factors such as individual gender and ancestry and (2) resolution of the array platforms used to ascertain the CNVs. Overall, the research findings described in this study represent a step forward to increase

genetic diagnostic yield in NDDs, enable the characterization of new NDD syndromes, and to expand the knowledge bases on molecular processes implicated in these conditions.

Web resources

Brain Span: www.brainspan.org/maseq
GeneTrek: <https://genetrek.pasteur.fr>
SFARI: <https://gene.sfari.org>
STRING: <https://www.string-db.org>

Supplemental information

Supplemental information can be found online at <https://doi.org/10.1016/j.xhgg.2024.100316>.

Acknowledgments

This study makes use of data generated by the DECIPHER community. A full list of centers who contributed to the generation of the data is available from <https://deciphergenomics.org/about/stats> and via email from contact@deciphergenomics.org. DECIPHER is hosted by EMBL-EBI and funding for the DECIPHER project was provided by the Wellcome Trust (grant no. WT223718/Z/21/Z). Access to the individual CNV calls from the iHart cohort¹⁷ was granted by the board of trustees of the Leland Stanford Junior University. We acknowledge the Hartwell Foundation for supporting whole-genome sequencing and the creation of the iHART database, as well as the Simons Foundation who provided additional support for the genome sequencing. We appreciate obtaining access to genetic data on SFARI Base (<https://base.SFARI.org>) and the Database of Genomic Variants (<http://dgv.tcag.ca/dgv/app/home>). S.A.C. was supported by industrial doctorate grant 2021DI056 from the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) of the Generalitat de Catalunya.

Author contributions

S.A. was primarily responsible for conducting the experiments, analyzing the data, and writing the initial draft of the manuscript. X.G. contributed to the study design, data processing and assisted in revising the manuscript. L.D. was involved in the experimental design and provided critical revisions to the manuscript. M.C. contributed to the project conceptualisation, supervised the study, and assisted in the interpretation of data. E.G. contributed to data analysis, and helped in the drafting and revision of the manuscript. L.P.C. led the overall project, was responsible for the study design and supervision, and critically reviewed and revised the manuscript.

Declaration of interests

S.A.C., X.G., L.P.-C., E.G., and L.D. are STALICLA employees, a company advancing treatment candidates for subgroups of affected individuals with NDDs.

Received: February 2, 2024

Accepted: June 3, 2024

References

1. Itsara, A., Wu, H., Smith, J.D., Nickerson, D.A., Romieu, I., London, S.J., and Eichler, E.E. (2010). De novo rates and selection of large copy number variation. *Genome Res.* 20, 1469–1481.
2. Pos, O., Radvanszky, J., Buglyo, G., Pos, Z., Rusnakova, D., Nagy, B., and Szemes, T. (2021). DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects. *Biomed. J.* 44, 548–559.
3. Bacchelli, E., Cameli, C., Viggiano, M., Iglizzi, R., Mancini, A., Tancredi, R., Battaglia, A., and Maestrini, E. (2020). An integrated analysis of rare CNV and exome variation in Autism Spectrum Disorder using the Infinium PsychArray. *Sci. Rep.* 10, 3198.
4. Shishido, E., Aleksic, B., and Ozaki, N. (2014). Copy-number variation in the pathogenesis of autism spectrum disorder. *Psychiatry Clin. Neurosci.* 68, 85–95.
5. Barone, J., Smith, M., Kendall, K., Owen, M., O'Donovan, M., and Kirov, G. (2019). The rate of *de novo* CNVs in healthy controls.
6. Girirajan, S., Brkanac, Z., Coe, B.P., Baker, C., Vives, L., Vu, T.H., Shafer, N., Bernier, R., Ferrero, G.B., Silengo, M., et al. (2011). Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genet.* 7, e1002334.
7. Nishino, J., Ochi, H., Kochi, Y., Tsunoda, T., and Matsui, S. (2018). Sample Size for Successful Genome-Wide Association Study of Major Depressive Disorder. *Front. Genet.* 9, 227.
8. Collins, R.L., Glessner, J.T., Porcu, E., Lepamets, M., Brandon, R., Lauricella, C., Han, L., Morley, T., Niestroj, L.M., Ulirsch, J., et al. (2022). A cross-disorder dosage sensitivity map of the human genome. *Cell* 185, 3041–3055.e25.
9. Marshall, C.R., Howrigan, D.P., Merico, D., Thiruvahindrapuram, B., Wu, W., Greer, D.S., Antaki, D., Shetty, A., Holmans, P.A., Pinto, D., et al. (2017). Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* 49, 27–35.
10. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* 87, 1215–1233.
11. Zhang, F., Gu, W., Hurles, M.E., and Lupski, J.R. (2009). Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10, 451–481.
12. Rice, A.M., and McLysaght, A. (2017). Dosage-sensitive genes in evolution and disease. *BMC Biol.* 15, 78.
13. Leblond, C.S., Le, T.L., Malesys, S., Cliquet, F., Tabet, A.C., Delorme, R., Rolland, T., and Bourgeron, T. (2021). Operative list of genes associated with autism and neurodevelopmental disorders based on database review. *Mol. Cell. Neurosci.* 113, 103623.
14. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47, W191–W198.
15. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29.
16. Gene Ontology Consortium (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* 49, D325–D334.
17. Ruzzo, E.K., Perez-Cano, L., Jung, J.Y., Wang, L.K., Kashef-Haghighi, D., Hartl, C., Singh, C., Xu, J., Hoekstra, J.N., Leventhal, O., et al. (2019). Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. *Cell* 178, 850–866.e26.
18. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* 84, 524–533.
19. MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* 42, D986–D992.
20. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443.
21. Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11, R106.
22. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Ru-beis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* 180, 568–584.e23.
23. Trost, B., Thiruvahindrapuram, B., Chan, A.J.S., Engchuan, W., Higginbotham, E.J., Howe, J.L., Loureiro, L.O., Reuter, M.S., Roshandel, D., Whitney, J., et al. (2022). Genomic architecture of autism from comprehensive whole-genome sequence annotation. *Cell* 185, 4409–4427.e18.
24. D'Souza, L., Channakkar, A.S., and Muralidharan, B. (2021). Chromatin remodelling complexes in cerebral cortex development and neurodevelopmental disorders. *Neurochem. Int.* 147, 105055.
25. Sokpor, G., Castro-Hernandez, R., Rosenbusch, J., Staiger, J.F., and Tuoc, T. (2018). ATP-Dependent Chromatin Remodeling During Cortical Neurogenesis. *Front. Neurosci.* 12, 226.
26. Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23, 324–328.
27. Grunblatt, E., Oneda, B., Ekici, A.B., Ball, J., Geissler, J., Uebe, S., Romanos, M., Rauch, A., and Walitza, S. (2017). High resolution chromosomal microarray analysis in paediatric obsessive-compulsive disorder. *BMC Med. Genom.* 10, 68.
28. Mefford, H.C., Sharp, A.J., Baker, C., Itsara, A., Jiang, Z., Buysse, K., Huang, S., Maloney, V.K., Crolla, J.A., Baralle, D., et al. (2008). Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N. Engl. J. Med.* 359, 1685–1699.
29. Coe, B.P., Girirajan, S., and Eichler, E.E. (2012). The genetic variability and commonality of neurodevelopmental disease. *Am. J. Med. Genet. C Semin. Genet. Genet.* 160C, 118–129.
30. Gudmundsson, O.O., Walters, G.B., Ingason, A., Johansson, S., Zayats, T., Athanasiu, L., Sonderby, I.E., Gustafsson, O., Nawaz, M.S., Jonsson, G.F., et al. (2019). Attention-deficit hyperactivity disorder shares copy number variant risk with schizophrenia and autism spectrum disorder. *Transl. Psychiatry* 9, 258.
31. Linden, S.C., Watson, C.J., Smith, J., Chawner, S.J.R.A., Lancaster, T.M., Evans, F., Williams, N., Skuse, D., Raymond, F.L., Hall, J., et al. (2021). The psychiatric phenotypes of 1q21 distal deletion and duplication. *Transl. Psychiatry* 11, 105.

32. Depienne, C., Moreno-De-Luca, D., Heron, D., Bouteiller, D., Gennetier, A., Delorme, R., Chaste, P., Siffroi, J.P., Chantot-Bastarud, S., Benyahia, B., et al. (2009). Screening for genomic rearrangements and methylation abnormalities of the 15q11-q13 region in autism spectrum disorders. *Biol. Psychiatry* 66, 349–359.
33. Thomas, J.A., Johnson, J., Peterson Kraai, T.L., Wilson, R., Tartaglia, N., LeRoux, J., Beischel, L., McGavran, L., and Hagerman, R.J. (2003). Genetic and clinical characterization of patients with an interstitial duplication 15q11-q13, emphasizing behavioral phenotype and response to treatment. *Am. J. Med. Genet.* 119A, 111–120.
34. Williams, N.M., Franke, B., Mick, E., Anney, R.J.L., Freitag, C.M., Gill, M., Thapar, A., O'Donovan, M.C., Owen, M.J., Holmans, P., et al. (2012). Genome-wide analysis of copy number variants in attention deficit hyperactivity disorder: the role of rare variants and duplications at 15q13.3. *Am. J. Psychiatry* 169, 195–204.
35. Bassett, A.S. (2011). Parental origin, DNA structure, and the schizophrenia spectrum. *Am. J. Psychiatry* 168, 350–353.
36. Costain, G., Lionel, A.C., Merico, D., Forsythe, P., Russell, K., Lowther, C., Yuen, T., Husted, J., Stavropoulos, D.J., Speevak, M., et al. (2013). Pathogenic rare copy number variants in community-based schizophrenia suggest a potential role for clinical microarrays. *Hum. Mol. Genet.* 22, 4485–4501.
37. Ingason, A., Kirov, G., Giegling, I., Hansen, T., Isles, A.R., Jakobsen, K.D., Kristinsson, K.T., le Roux, L., Gustafsson, O., Craddock, N., et al. (2011). Maternally derived microduplications at 15q11-q13: implication of imprinted genes in psychotic illness. *Am. J. Psychiatry* 168, 408–417.
38. Rees, E., Walters, J.T.R., Georgieva, L., Isles, A.R., Chambert, K.D., Richards, A.L., Mahoney-Davies, G., Legge, S.E., Moran, J.L., McCarroll, S.A., et al. (2014). Analysis of copy number variations at 15 schizophrenia-associated loci. *Br. J. Psychiatry* 204, 108–114.
39. Cox, D.M., and Butler, M.G. (2015). The 15q11.2 BP1-BP2 microdeletion syndrome: a review. *Int. J. Mol. Sci.* 16, 4068–4082.
40. Vanlerberghe, C., Petit, F., Malan, V., Vincent-Delorme, C., Bouquillon, S., Boute, O., Holder-Espinasse, M., Delobel, B., Duban, B., Vallee, L., et al. (2015). 15q11.2 microdeletion (BP1-BP2) and developmental delay, behaviour issues, epilepsy and congenital heart disease: a series of 52 patients. *Eur. J. Med. Genet.* 58, 140–147.
41. Zarrei, M., Burton, C.L., Engchuan, W., Young, E.J., Higginbotham, E.J., MacDonald, J.R., Trost, B., Chan, A.J.S., Walker, S., Lamoureux, S., et al. (2019). A large data resource of genomic copy number variation across neurodevelopmental disorders. *NPJ Genom. Med.* 4, 26.
42. Gandal, M.J., Haney, J.R., Parikshak, N.N., Leppa, V., Ramaswami, G., Hartl, C., Schork, A.J., Appadurai, V., Buil, A., Werge, T.M., et al. (2019). Shared Molecular Neuropathology Across Major Psychiatric Disorders Parallels Polygenic Overlap. *Focus* 17, 66–72.
43. Wassink, T.H., Piven, J., Vieland, V.J., Jenkins, L., Frantz, R., Bartlett, C.W., Goedken, R., Childress, D., Spence, M.A., Smith, M., and Sheffield, V.C. (2005). Evaluation of the chromosome 2q37.3 gene CENTG2 as an autism susceptibility gene. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 136B, 36–44.
44. Aldred, M.A., Sanford, R.O.C., Thomas, N.S., Barrow, M.A., Wilson, L.C., Brueton, L.A., Bonaglia, M.C., Hennekam, R.C.M., Eng, C., Dennis, N.R., and Trembath, R.C. (2004). Molecular analysis of 20 patients with 2q37.3 monosomy: definition of minimum deletion intervals for key phenotypes. *J. Med. Genet.* 41, 433–439.
45. Ogura, K., Takeshita, K., Arakawa, C., Shimojima, K., and Yamamoto, T. (2014). Neuropsychological profiles of patients with 2q37.3 deletion associated with developmental dyspraxia. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 165B, 684–690.
46. Droscher, M., Chaugule, V.K., and Pichler, A. (2013). SUMO rules: regulatory concepts and their implication in neurologic functions. *NeuroMolecular Med.* 15, 639–660.
47. Bhat, S., Ljubojevic, N., Zhu, S., Fukuda, M., Echard, A., and Zurzolo, C. (2020). Rab35 and its effectors promote formation of tunneling nanotubes in neuronal cells. *Sci. Rep.* 10, 16803.
48. Pagani, M., Barsotti, N., Bertero, A., Trakoshis, S., Ulysse, L., Locarno, A., Miseviciute, I., De Felice, A., Canella, C., Supekar, K., et al. (2021). mTOR-related synaptic pathology causes autism spectrum disorder-associated functional hyperconnectivity. *Nat. Commun.* 12, 6084.
49. Bryant, C.D., and Yazdani, N. (2016). RNA-binding proteins, neural development and the addictions. *Genes Brain Behav.* 15, 169–186.
50. Qiu, T., Ruan, M.A.R., Lynch, W.B., Cox, J.W., Beierle, J.A., Yao, E.J., Kandola, A., Chen, M.M., Kelliher, J.C., Babbs, R.K., et al. (2021). The methamphetamine-induced RNA targetome of hnRNP H in Hnrnp1 mutants showing reduced dopamine release and behavior. Preprint at bioRxiv. <https://doi.org/10.1101/2021.07.06.451358>.