A cell type-specific expression signature predicts haploinsufficient autism-susceptibility genes

Chaolin Zhang[1,2,3,6,*], Yufeng Shen[1,4,5,6*]

[1] Department of Systems Biology,

[2] Department of Biochemistry and Molecular Biophysics,

[3] Center for Motor Neuron Biology and Disease,

[4] Department of Biomedical Informatics,

[5] JP Sulzberger Genome Center

Columbia University, New York NY 10032, USA

[6] Equal contribution

[*] To whom correspondence should be addressed:

cz2294@columbia.edu (C.Z.); ys2411@cumc.columbia.edu (Y.S.)

Abstract

Recent studies have identified many genes with rare *de novo* mutations in autism, but a limited number of these have been conclusively established as disease-susceptibility genes due to lack of recurrence and confounding background mutations. Such extreme genetic heterogeneity severely limits recurrence–based statistical power even in studies with a large sample size. In addition, the cellular contexts in which these genomic lesions confer disease risks remain poorly understood. Here we investigate the use of cell-type specific expression profiles to differentiate mutations in autism patients or unaffected siblings. Using 24 distinct cell types isolated from the mouse central nervous system, we identified an expression signature shared by genes with likely gene disrupting (LGD) mutations detected by exome-sequencing in autism cases. The signature reflects haploinsufficiency of risk genes enriched in transcriptional and post-transcriptional regulators, with the strongest positive associations specific types of neurons in different brain regions, including cortical neurons, cerebellar granule cells, and striatal medium spiny neurons. Based on this signature, we assigned a D score to all human genes to prioritize candidate autism-susceptibility genes. When applied to genes with only a single LGD mutation in cases, the D-score achieved a precision of 40% as compared to the 15% baseline with a minimal loss in sensitivity. Further improvement was made by combining D score and mutation intolerance metrics from ExAC which were derived from orthogonal data sources. The ensemble model achieved precision of 60% and predicted 117 high-priority candidates. These prioritized lists can facilitate identification of additional autism-susceptibility genes.

Keywords: autism spectrum disorders; autism-susceptibility genes; *de novo* mutations; cell type-specific expression signature; D score

Introduction

Autism or autism spectrum disorders (ASDs) are very common neurodevelopmental diseases characterized by deficits in language, impaired social interaction, and repetitive behaviors with complexes such as seizures and intellectual disability (1, 2). Symptom onset is typically early (~3 years old) and the current estimate of incidence is over 1% worldwide (3), making ASDs a huge burden for affected families and for society.

Genetic risk factors are believed to play a pivotal role in ASDs, as revealed by a concordance rate up to 90% between monozygotic twins and by over 10-fold increase in the risk for a new born child if a previous sibling is affected (4). Some syndromic forms of autisms are known to be monogenic, as represented by mutations of *FMR1* (encoding FMRP) in the fragile X syndrome that is comorbid with autism and accounts for up to 5% of ASD cases (5, 6). However, most of the genomic abnormalities or mutations found in autism patents are extremely rare and frequently *de novo*. Earlier studies using microarray-based approaches identified hundreds of *de novo* copy number changes (CNVs) (7-12). More recently, *de novo* mutations in individual nucleotides, including single nucleotide variations (SNVs) and small insertions and deletions (indels) were identified by exome-sequencing (13-18) or whole genome sequencing (19, 20).

The exciting progress in these genetic studies has provided important insights into the etiopathogenesis of autism. First, at least a substantial proportion of autism risk is conferred by individually rare mutations affecting one or more disease-susceptibility genes. The number of risk loci has been estimated to be in the range of several hundred to over 1,000 genes (4, 16, 17). Second, although the complexity of the genetic landscape underlying autism is still a matter of debates, one theory supported by several lines of evidence favors that a large number of autism risk loci are individually of high penetrance (4, 21, 22). Third, analysis of the seemingly isolated candidate autism-susceptibility genes points to disruption in several convergent molecular pathways (4, 16, 23-25) that inform the neurobiological underpinnings of autism (reviewed by ref. (26)).

Strikingly, even the most frequent *de novo* mutations in single genes can explain no more than 1% of ASD cases (26). This extreme genetic heterogeneity presents a big challenge for conclusive identification of autism-susceptibility genes, which impeded further functional studies of autism neurobiology and development of therapeutic strategies. A particular group of *de novo* mutations identified by whole exome-sequencing (13-16, 18) are “likely gene-disrupting” or LGD mutations, which is a collection of severe mutations introducing frame shift, disrupted splice sites or premature stop codons. Probands clearly show a higher burden of *de novo* LGD mutations than their unaffected siblings used as control, indicating enrichment of disease-susceptibility genes disrupted by these genomic lesions. However, although about 4,000 ASD patients and their families have been sequenced so far, only about 40 genes at best have been determined as high-confidence autism-susceptibility genes based on their recurrence. Most of the remaining mutations were observed in only single patient, and ~80% of these are expected to be non-pathogenic (see Material and Methods). The signal-to-noise is even lower for genes with missense mutations (17) due to their more moderate effects and high background mutation frequency. Therefore, most autism-susceptibility genes are currently buried among a growing list of potential candidates.

Two general strategies can be used to facilitate the identification of candidate autism-susceptibility genes. One strategy is to sequence larger cohorts of ASD families. For example, the SPARK project (27) aims to recruit and analyze 50,000 individuals of autism and their families. Furthermore, whole-exome or whole-genome sequencing is also complemented by targeted re-sequencing to reduce cost (28). A caveat of this strategy is its prohibitive cost associated with recruitment and sequencing of large cohorts. A second complementary strategy is to stratify already identified mutations based on existing orthogonal information associated with the affected genes. For example, depletion of rare deleterious variants estimated from the general population reflecting severe selection pressure is effective in prioritizing deleterious mutations (29-31). Alternatively, gene regulatory and function information can also be used to help distinguish pathogenic vs. neutral mutations. The latter is based on the assumption that the common clinical phenotype of ASDs originates from certain common features shared by autism risk loci at the molecular level. Along this line, recent studies identified molecular pathways underlying autism etiopathology from analysis of shared genetic phenotypes (24, 32), protein-protein interactions (14), and gene co-expression networks (23, 25). However, these previous studies utilizing network analysis techniques were not designed to optimize the prioritization of individual autism-susceptibility genes *per se*, or to rigorously evaluate the signal-to-noise ratio associated with prediction.

In this work, we predict autism-susceptibility genes by using gene expression profiles in a wide range of specific neuronal cell types, with the assumption that different cell types have different susceptibility or relevance to autism etiopathogy. Due to the heterogeneity among many different cell types in the brain, such an analysis may reveal cell type-specific gene regulation that cannot be detected by analysis of brain tissues used in many previous studies. We identified a gene expression signature reflecting haploinsufficiency in the context of autism that was able to effectively predict whether individual LGD mutations confer disease risk. Importantly, the use of cell type-specific expression also allows us to highlight the cellular contexts of the identified risks. Furthermore, this signature is complementary to previous mutation intolerance analysis, and an ensemble model combining multiple scoring metrics gave the optimal prediction accuracy.

Results

DAMAGES analysis uncovers an expression signature of autism-susceptibility genes

Our study was motivated by a postulation that different cell types in the brain have different susceptibility and impact on autism etiopathology, which is supported by recent studies showing expression bias in candidate autism-susceptibility genes (32, 33). However, to the best of our knowledge, the effectiveness of cell type-specific expression in predicting autism-susceptibility genes, alone or in combination with other metrics, has not been systematically explored.

We developed and applied a computational framework for disease-associated mutation analysis using gene expression signatures (DAMAGES) to score the association of human genes with autism, to refine the lists of candidate autism-associated genes currently available, and to uncover features shared by these genes as a means of understanding the molecular underpinnings of the disease (Figure S1). In contrast to previous network analysis approaches, we adopted a case-control classification framework so that it can take advantage of rich resources of existing gene expression profiles and machine learning approaches to optimize and objectively evaluate the accuracy of prediction. To provide a proof of principle in this study, we focused on the utilization of gene expression profiles in a wide range of isolated central nervous system (CNS) cell types. We decided to examine a large microarray dataset that profiles cell type-specific transcripts associated with translating ribosomes in the mouse brain generated by a biochemical assay named TRAP (34, 35). In total, this dataset is composed of translational profiles of 24 specific mouse CNS cell types, including both neurons and glial cells, isolated from six different regions, together with unselected RNA representing all cell types in each of these regions (34) (Figure S1 A,C). This translational profiling approach was previously shown to give robust gene expression measurements, and to effectively identify known and novel cell-type specific markers and provide biological insights into each cell type (34).

To identify expression signatures of autism-susceptibility genes, we started with a list of 162 genes containing *de novo* LGD mutations in either ASD probands or unaffected siblings collected from four exome-sequencing studies (13-16) (Figure S1B). Our prediction model was built using these mutations representing all information available before 2013, which gave us a chance to use additional genes discovered by large-scale followup studies for objective evaluation. In total, 145 genes have mouse orthologs and were represented in the microarray dataset we used (Table S1). We assume that the 33 genes with LGD mutations in unaffected siblings confer no risk of ASD (non-disease genes). On the other hand, the other 112 genes with LGD mutations in probands represent a mixture of ~65 disease-susceptibility genes and ~47 non-disease genes (Table S1 and Material and Methods; a similar estimate provided in (25)). It is worth noting that we limited our analysis to mutations derived from unbiased genomic screens to build the model, and excluded candidates identified by more targeted or hypothesis-driven approaches or by transcriptomic analysis to avoid potential ascertainment bias. This is particularly critical for an objective assessment of DAMAGES analysis in prediction accuracy.

Given the relatively balanced representation of autism-susceptibility genes and non-disease genes (estimated to be ~65 and ~80, respectively) in the dataset, we anticipated that the contrast between these two groups of genes would represent a major axis of expression dynamics in the high-dimensional space. A principal component analysis (PCA) (36) was thus performed to identify the orthogonal axes that explain the most variance. This analysis revealed that projection of genes to the second principal component (PC) is very predictive of mutations in probands versus controls (Figure 1A-B). We note that PCA is an unsupervised method which does not incorporate information on the source of mutations so the prediction performance is not due to data overfitting. To have a more rigorous assessment of the ability of each PC or combination of PCs to differentiate potentially disease-associated versus neutral mutations, we performed a regularized linear regression analysis lasso (37) to find the PCs that are most predictive of the source of mutations. In this method, a parameter λ, specifying the penalty against more complex models, controls the number of PCs contributing to the regression model (i.e., having a non-zero regression coefficient β). We found that only PC2 has a non-zero coefficient across a wide range of λ, while the coefficients of all other PCs shrink to zero very quickly as λ increases (Figure 1C). To determine whether the regression models overfit, we performed leave-one-out cross validation (LOOCV) to predict the source of mutations using varying values of λ (and thereby different numbers of PCs in the regression models). The best performance was achieved when a single component PC2 was used for prediction (λ=5) (Figure 1D). Therefore, we decided to use the PC2 as a signature of autism-susceptibility genes, and adjusted the threshold according to LOOCV, which resulted in the final DAMAGES scores (or D scores) used for gene ranking. As a result, we were able to identify 93 genes with positive D scores, including 83 genes with LGD mutations in probands and 10 genes with LGD mutations in siblings, respectively (arrowhead in Figure 1B and Table S1). Proband-specific LGD mutations are strikingly enriched in genes with positive D scores compared to the remaining genes (Figure 1B; odds ratio=6.48, P=8.06×10^{-6}, Fisher's exact test). The ability of the gene expression signature to differentiate mutations in probands from those in siblings suggests that at least some of the CNS cell types included in the microarray dataset are strongly associated with the underlying molecular mechanisms of autism.

Validation of the expression signature using expanded autism exome sequencing data sets

We assigned a D score to all human genes represented in the microarray dataset independent of their mutation status (Table S2). This allowed us to have an independent, "prospective" evaluation of the performance of the expression signature using an expanded list of genes with LGD mutations from recent large-scale studies after our prediction model was built (Figure 1E and Table S1)(17, 18). In this expanded dataset, almost all genes with recurrent mutations in autism patients (35/38=92%) received a positive D score (exceptions are *DSCAM*, *RANBP17,* and *TCF7L2*; two genes not represented on the array were excluded). Among the genes ranked in top 25% by the D score, there is a 2.8 fold enrichment (P=3.2×10^{-12}) of LGD mutations in cases comparing to unaffected siblings, whereas there is no significant enrichment in rest of genes (rate enrichment = 1.2, P=0.16) (Table 1). Therefore, DAMAGES analysis prioritized *bona fide* autism-susceptibility genes with minimal loss.

For additional validation, we examined 528 genes compiled in the Simon Foundation Autism Research Initiative (SFARI) autism gene database, a list of potential autism-associated genes manually curated by experts according to various types of evidence available in the literature (https://gene.sfari.org) (38). Among the 483 SFARI genes represented in the microarray data, 300 (62%) have a positive D score (Figure 2A and Table S3), a very significant enrichment compared to all genes represented in the microarray dataset (odds ratio=2.34, P=6.3×10^{-20}; Fisher's exact test). In addition, this proportion is higher for genes with more evidence supporting their implication in autism (Figure 2B). For example, genes with positive D scores include 5/5 (100%) genes that are classified as strong candidates and 15/20 (75%) genes that are classified as syndromic, such as *FMR1*, *MECP2* [Rett syndrome (39)] and *TSC1/2* [tuberous sclerosis complex (40)]. Furthermore, SFARI genes received much higher ranks based on the D score, as compared to ranking by the first PC reflecting the neuron-glial distinction (P=1.2×10^{-22}; Wilcox ranksum test; see below).

CNS cell types associated with autism

To further confirm this molecular signature and gain more biological insights, we examined the loadings of different cell types on each PC (Table S4). The first PC essentially differentiates neurons versus glial cells and unselected cell types in different brain regions (Figure 3A). In contrast, the second PC predictive of autism-susceptibility genes appears to give more of a mix of different cell types and regions, although a certain bias is also clear (e.g., cortical neurons have the highest positive loadings; see below). To assess whether this pattern reflects their association with the underlying molecular mechanisms of autism and is specific for autism-susceptibility genes, we performed another PCA using all genes showing the most variation across different cell types. The first PC of the whole dataset is highly correlated with the one derived from genes with LGD mutations (R^2=0.74), and similarly differentiates neurons from glial cells and unselected cell types (Figure 3B). This result is consistent with the notion that even among genes with LGD mutations the distinction of neuronal versus glial genes dominates the expression dynamics. In contrast, the second PC identified in the global PCA has a low correlation with the second PC identified using genes with LGD mutations (R^2=0.19; Figure 3C). This observation supports the notion that the molecular signature identified using genes with LGD mutations indeed reflects certain specific features shared by autism-susceptibility genes.

Therefore, the loadings of different cell types on the signature (PC2) likely reflect their association with autism (Figure 3D). In general, none of the glial cell types included for this analysis has a positive association. On the other hand, the association of neurons with the signature varies depending on specific cell types and brain regions. Different types of cortical neurons, including interneurons, corticothalamic neurons, corticospinal and corticopontine neurons, Cck^+ neurons, and corticostriatal neurons, have large positive loadings on the signature. However, not all types of cortical neurons have a positive association, and some, such as $Pnoc^+$ interneurons, have a negative loading. Besides cortical neurons, cerebellar granule cells, striatal medium spiny neurons, but not Purkinje cells, cholinergic neurons, or motor neurons, show a strong positive loading. Altogether, these observations are not only consistent with autism being mainly an impairment of high-level cognitive functions, but also suggest that even in a given brain region, specific cell types may play very different roles in the etiopathogenesis of the disease.

Molecular functions associated with the autism-susceptibility gene expression signature

We asked whether the expression signature captures certain molecular functions shared by autism-susceptibility genes. To this end, we performed Gene Ontology (GO) analysis (41) using the top 500 protein-coding genes ranked by D scores independent of their mutation status. This analysis revealed very strong enrichment of those involved in "transcription" (Benjamini FDR=3×10^{-14}), "chromatin modification" (Benjamini FDR=7.9×10^{-6}) and "regulation of RNA metabolic process" (Benjamini FDR=2.2×10^{-4}) (Table S5). It is worth noting that "chromatin organization" is also enriched in the 83 high-priority candidate genes, although the statistical significance is marginal (Benjamini FDR=0.07). Therefore, not only are genes with LGD mutations themselves enriched in those important for transcriptional regulation, as noted previously (14, 16, 42), but they define a molecular signature represented by a larger set of genes with coherent molecular functions in both transcriptional and post-transcriptional regulation of gene expression.

The expression signature reflects haploinsufficiency

We postulate that the expression signature may reflect haploinsufficiency because it was derived from genes with heterozygous loss of function. To test this hypothesis, we examined genes covered by relatively focal *de novo* CNV events (≤50 genes) detected in ASD probands (7-9, 11, 12). Interestingly, genes with higher D scores tend to overlap with deletions than amplifications (P<0.04; Spearman correlation test). The significance is relatively marginal, presumably due to the limited spatial resolution of the CNVs. For further confirmation, we examined genes differentially expressed in post-mortem autism brains as compared to controls (43), assuming that the dosage-dependent alteration can also be caused by changes at the transcription level. Indeed, genes down-regulated in autism tend to have a positive D score ($P<2.2\times10^{-16}$), while genes upregulated in autism tend to have a negative D score ($P<2\times10^{-7}$; Figure S2A).

Recently the Exome Aggregation Consortium (ExAC) used large-scale exome sequencing data of general populations without developmental disorders to estimate metrics of haploinsufficiency (31), including the probability of being loss-of-function (LoF) intolerant (pLI) and LoF Z-score. A positive correlation ($r = 0.29$, $P < 10^{-10}$) of LoF Z score and D score was observed among genes with positive D scores (Figure S2B). This observation is again consistent with the notion that both metrics are related to haploinsufficiency although they were derived using entirely different data sources and assumptions.

Prioritizing genes in CNVs including 16p11.2

De novo CNVs detected in ASD probands typically span dozens or hundreds of genes (7-12). Therefore, although over 2,000 genes are covered by at least one CNV identified so far, it is difficult to differentiate *bona fide* autism-susceptibility genes from other passenger genes. We focused on 58 deletion CNVs for which all overlapping genes have mouse orthologs and are represented in the microarray data. Of these, 30 CNVs each have one and only one gene with a positive D score. Based on the high sensitivity (~90%, see below) of a positive D score in predicting autism-susceptibility genes, we argue that if a CNV is pathogenic, the only gene with a positive D score is the most likely causal gene. We therefore denote the CNV “likely supporting CNV” or LS-CNV of the corresponding gene. This analysis resulted in 19 genes supported by deletion LS-CNV events in one or more patients (Table S6). Of these genes, five are supported by recurrent LS-CNVs (*NRXN1*, *DPP6*, *PTPRT*, *SHANK2* and *SLC4A10*), and all of these five genes are known to have functional implications in synapse (44-48) and/or autism-related phenotypes (47). Remarkably, three genes harbor recurrent LGD mutations in ASD patients (*ANKRD11*, *CHD3* and *KMT2C;* odd ratio=130, $P<3.4\times10^{-6}$, Fisher’s exact test). Two additional genes (*NRXN1* and *SHANK2*) have singleton LGD mutations from exome sequencing (i.e., recurrent if the LS-CNV is counted).

LS-CNVs tend to span a smaller number of genes than CNVs in general. For a majority of CNVs overlapping with more genes, it is difficult to reliably distinguish susceptibility genes versus passenger genes even with the D score. Nevertheless, it is still possible to eliminate a substantial fraction of passenger genes. To illustrate this point, we examined the most frequent recurrent *de novo* CNV located in 16p11.2, which accounts for up to 1% of ASD cases (9) (14 deletions and 5 duplications in the dataset used for this analysis; Figure 4A). This region spans 26 genes (11) and all of them have mouse orthologs; deletion of the region in mice phenocopied behavior deficits observed in ASD patients (49). Among the 23 genes represented in the microarray data, nine have a positive D score (Figure 4B). Interestingly, deletion of a smaller region in this locus also segregates with ASD or ASD traits (50). This deletion encompasses five genes, including *KCTD13*, *ASPHD1* and *SEZ6L2* with a positive D score. A recent study further demonstrated that *KCTD13* is a major driver of the macrocephalic phenotype associated with ASD cases carrying the 16p11.2 CNV (51). Some of the other genes in this locus, especially the ones highlighted by DAMAGES analysis, could contribute to the additional clinical manifestations in ASDs.

An ensemble model for optimized prediction of autism-susceptibility genes

We next compared the performance of D score and ExAC scores in predicting autism-susceptibility genes and investigated if improved performance can be achieved by combining these different metrics. We focused on genes with single LGD mutations in cases and found ExAC pLI>0.9 or LoF Z-score>3 achieved similar optimal performance as D score (about 40% precision and 90% sensitivity; Figure 5A, B), as compared to a baseline 15% precision. Importantly, while predictions by the two metrics overlap, small genes with low background mutation rate (such as MECP2, pLI=0.7) tend to be missed by ExAC scores. To investigate whether ExAC scores and D score make independent contributions in gene prioritization, we performed a logistic regression to classify recurrently mutated genes in cases and genes with LGD mutations in controls, with ExAC pLI, ExAC mis-z, and D score as features. The coefficients of all three features deviate significantly from zero (Table 2), indicating that D score is complementary to ExAC scores in determining gene LGD intolerance (interestingly, gene expression in embryonic mouse brain did not any predictive power). Therefore, combining these scores would maximize the performance in candidate gene prioritization. To assess that, we applied the estimated logistic model to all genes to calculate an ensemble score (Table S2), and estimated precision-recall rates in a range of top rank thresholds, excluding the genes that are recurrently mutated. We found that the ensemble score outperforms all individual methods, with an optimal performance among the top 1,300 genes (Figure 5B,C). With ensemble score, the precision quadruples to 60% with near-maximal sensitivity (estimated to be 97%). Using this threshold, we identified 117 candidate genes with a singleton LGD mutation (Table S7). Since ASD shares substantial number of risk genes with other neurodevelopmental disorders (18, 52), we obtained the *de novo* mutation calls from the latest released data (53) from the Deciphering Developmental Disorders (DDD) project (54) to further assess the performance of ensemble score prediction. The DDD data includes 4,293 patients with severe undiagnosed developmental disorders. Among 117 candidate genes predicted by ensemble score, 65 harbor at least one LGD or damaging missense (predicted by metaSVM (55) or polyphen-2 (56) and CADD (57)) *de novo* mutations in the DDD data set. This rate is much higher than non-candidate genes with a single LGD mutation in ASD data (odds ratio = 4.6; p-value = 2×10^{-11}). Conversely, among all the genes with a single LGD mutation in ASD data, the ones with at least one damaging mutation in the DDD data have higher ensemble score than the ones without (P = 1.3×10^{-11}; KS test) (Figure 5D). These observations indicate that the candidate genes predicted by the ensemble score are much more likely to be associated with developmental disorders in general.

Gender bias of autism-associated *de novo* mutations

The incidence of autism has a strong gender bias, with a male:female ratio around 4 overall and even higher for high-functioning cases. This is reflected in the population of participants included in the exome-sequencing studies (M:F=6.4 in the SSC dataset; Table S8). However, a lower incidence of *de novo* mutations in males was previously observed (10, 16, 17). We examined the gender bias of *de novo* LGD mutations in different sets of genes, focusing on mutations identified from the Simon SSC dataset, for which the number of male and female patients is known (Table S8). Consistent with previous observations, a lowest M:F ratio was observed from genes with recurrent LGD mutations and genes with singleton LGD mutations predicted by the ensemble model (M:F~0.5, after correction for the gender bias of the participants; P<0.02, Fisher's exact test). A more moderate, but significant, gender bias was observed in genes with singleton mutations predicted by D score alone (M:F=0.62; P=0.02, Fisher's exact test). No significant gender bias was observed among singleton mutations in genes showing a negative D score (M:F=0.89; P=0.65, Fisher's exact test). These observations provided an independent line of evidence that D-score and the ensemble model can discriminate disease-susceptibility genes from non-disease genes.

Discussion

Here we present the use of cell type-specific gene expression profiles to improve prediction of autism-susceptibility genes. The molecular signature uncovered by DAMAGES analysis has several implications. First, this study echoes recent findings on convergent molecular pathways underlying autism etiopathogenesis including, approximately, three modules: synaptic structure and function, transcriptional regulation and chromatin remodeling, and Wnt signaling (reviewed by ref. (26) (18)). Conclusions of these studies were drawn from analysis of co-occurrence in genetic phenotypes (24), protein-protein interactions (14), and gene co-expression networks reflecting developmental dynamics in different brain regions (23, 25). This work extended these previous efforts by demonstrating that a robust signature of autism-susceptibility genes can be defined by their expression patterns in a range of specific CNS cell types (32, 33). Importantly, this signature reflects genes involved in transcriptional and post-transcriptional regulation and haploinsufficiency caused by LGD mutations in these genes. The importance of transcription factors and chromatin regulators in autism is now well established (14, 26). In addition, the role of post-transcriptional regulation is in line with the observation that several monogenic autism risk loci, including FMRP and MeCP2, are important regulators in RNA metabolism (58), and that candidate autism-associated genes show significant overlap with target transcripts of several neuronal RNA-binding proteins including FMRP (16, 59) and RBFOX1 (A2BP1) (43, 60, 61). Indeed, among the 822 FMRP target genes represented in the microarray dataset(59), a vast majority (718 or 87%) have a positive D score.

Second, while gene expression in brain or different brain regions was previously used to filter (13) or predict (62) candidate autism-susceptibility genes and, more recently, to reveal related pathways (23, 25), these previous methods were not optimized to predict individual autism-susceptibility genes or to provide rigorous assessment of the signal-to-noise ratio of resulted predictions. A key feature of DAMAGES analysis is that it adopts a case-control design using candidate genes derived from genomic DNA screens, which are completely independent of expression data. This design allows rigorous assessment of the biological relevance and predictive power of the uncovered signature by controlling potential confounding factors such as non-uniform mutation rates in different groups of genes. Our results demonstrated that the cell type-specific expression signature greatly increased the specificity of predicting autism-susceptibility genes with minimal loss of true hits. This is reflected in the observation that DAMAGES analysis predicted 35 out of 40 genes with recurrent LGD mutations identified so far and all 5 non-syndromic candidate genes with the highest confidence in the SFARI autism gene database. Importantly, information provided by the expression signature is complementary to the scoring metrics based on analysis of mutation frequency in the general population, and improved performance was achieved by combining the two methods.

Lastly, the cell-type specific signature captures a strong positive association with multiple types of cortical neurons, cerebellar granule cells, and striatal medium spiny neurons. This observation implies haploinsufficiency of genes that are normally highly expressed in these cell types as a converging pathogenic mechanism in autism. The implication of cortical projection neurons (25) and cells in the granule layer of the cerebellum (63) has been noted recently. In basal ganglia, striatal medium spiny neurons are known as the primary cell type vulnerable in Huntington's disease (64). In the context of autism, it has been shown that depletion of *SHANK3*, a gene highly expressed in striatum and regarded as the cause of the autism-related Phelan-McDermid Syndrome, results in ASD-like features such as impaired social interaction in mouse models (65). *SHANK3* has a singleton LGD mutation detected in the current exome-sequencing studies, and ranks among the top 5% genes genome-wide by the D score (see Figure 4A). In the cerebellum, a reduction of Purkinje cells and granule cells has been found in postmortem autistic brains and in mouse models (66, 67). Interestingly, our analysis revealed that cerebellar granule cells show a strong positive loading with a magnitude similar to those observed from cortical neurons, while Purkinje cells in cerebellum show weak loadings. These data suggest an intriguing hypothesis that different molecular mechanisms might underlie the loss of Purkinje and granule cells, although this has to be tested in future work. Finally, glial cells, especially astrocytes, show a strong negative loading. However, this should probably not exclude the contribution of these cells to autism. Instead, an alternative interpretation is that these genes may confer risks through other mechanisms than haploinsufficiency, which is supported by the observation that a subset of immune genes and glia markers are overexpressed in autism brains(43).

In summary, this study suggests the potential of utilizing gene expression and regulation information in predicting pathogenic mutations in autism. While we focused on cell type-specific expression in this work to demonstrate the proof of principle, we anticipate that additional expression profiles and functional annotations of genes, which can be easily integrated using a machine learning framework, will further improve the performance. Prioritized gene lists from such analysis can facilitate further validation by targeted re-sequencing in large cohorts (28) or more mechanistic studies using model organisms. This application will be particularly useful as the list of mutations are expected to grow steadily as a result of continuing autism exome- and genome-wide sequencing projects (4, 26) (27).

Material and Methods

Data compilation

For the current DAMAGES analysis, we used microarray gene expression profiles in 24 mouse central nervous system (CNS) cell types isolated from six brain regions, as well as unselected RNAs in each of these regions. This dataset was previously generated using a translational profiling approach named TRAP (34). For each gene, we selected the probeset with the maximum median expression across all 30 samples as a representative, if multiple probesets exist. In total, 20,870 genes with Entrez gene IDs are represented in the dataset. Expression intensities for each gene were first log2-transformed, with a pseudocount of 8 added to the intensities on the original scale.

The initial list of genes with *de novo* mutations in ASD probands and unaffected siblings were collected from four whole exome-sequencing studies (13-16). A total of 162 genes have *de novo* LGD mutations either in the probands or siblings. We determined the mouse ortholog of each human gene using the HomoloGene database (http://www.ncbi.nlm.nih.gov/homologene), complemented by manual searches. Mouse orthologs were found for 158 genes with LGD mutations, including 123 genes with LGD mutations only in probands and 35 genes with LGD mutations in siblings. Among these, a total of 145 genes, including 112 genes with LGD mutations in probands and 33 genes with LGD mutations in siblings, were represented in the microarray data, and were used for the initial model building and analysis. We also used 1,479 genes with log2 expression intensities ≥ 6 in two or more cell types and a standard deviation ≥ 2 across all cell types for PCA analysis shown in Figure 3B.

For further validation, we compiled an expanded list of genes with LGD mutations in autism patients (cases) or in their unaffected siblings (controls) from more recent exome-sequencing studies of about 3,960 cases and 1911 controls (17, 18). In total, we obtained a list of 670 genes, including 40 genes with recurrent LGD mutations in patients, 468 genes with singleton LGD mutation in patients, and 173 genes with LGD mutations in controls (note we excluded *TTN*). For genes that are not represented on the microarray data, we assigned their D score to the median value of all represented genes when we built the ensemble model (described below).

De novo copy number variation (CNV) data in ASD probands and annotations of overlapping genes were obtained from (11). This list is composed of 219 CNVs, and was compiled by the original authors from several previous studies (7-9, 11, 12). Technically redundant mutations, due to inclusion of the same patient samples in multiple studies, have already been removed from the list, so that recurrence of CNVs observed in the list is genuine. We similarly identified the mouse orthlogs of these CNV genes, and those (1,571 genes total) represented on the microarrays.

SFARI autism genes were downloaded in July 2013 from https://gene.sfari.org (38).

Principal component analysis

DAMAGES analysis adopts a supervised classification framework taking advantage of the case-control design in the recent mutation screens. For the current work, the analysis is composed of two major steps: dimension reduction of gene expression data by PCA and prediction of autism-susceptibility genes by a regularized regression method.

For PCA analysis, log2-transformed expression intensities for each gene were first standardized across the 30 cell types to obtain zero means and unit standard deviations. PCA was performed in R using the princomp package. In brief, normalized expression values of each sample (i.e., cell type) were further standardized across genes to give a data matrix denoted as $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ with n columns (genes) and p=30 rows (cell types). PCA (or equivalently, singular value decomposition) decomposes this matrix into the following form:

$$\begin{aligned} X^T &= U \Sigma V^T \\ &= \begin{pmatrix} u_{11} & \cdots & u_{1q} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nq} \end{pmatrix} \begin{pmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_q \end{pmatrix} \begin{pmatrix} v_{11} & \cdots & v_{p1} \\ \vdots & \ddots & \vdots \\ v_{1q} & \cdots & v_{pq} \end{pmatrix} \\ &= \begin{pmatrix} u_{11} & \cdots & u_{1q} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nq} \end{pmatrix} \begin{pmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_q \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_q^T \end{pmatrix} \\ &= SV^T \end{aligned} . \qquad (1)$$

Here columns of U and V are unit orthogonal vectors (eigen vectors), and $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_q$ (q≤30). With this decomposition, each gene $\mathbf{x}_g = \sum_{i=1}^{q} u_{gi}\alpha_i \mathbf{v}_i$, a linear combination of the first q principal components (PCs). Here $s_{gi} = u_{gi}\alpha_i$ is the score of gene g projected onto PC i; v_{ci} is the loading of cell type c on PC i.

For a new set of genes, the score matrix is calculated by

$$S = X^T V . \qquad (2)$$

Regularized regression analysis

We used a regularized regression analysis named lasso(37) to evaluate the contribution of each PC to prediction of each gene with respect to the source of mutations (probands versus siblings) with the following representation:

$$y_g = \sum_{i=1}^{q} \beta_i s_{gi} , \qquad (3)$$

by minimizing

$$\sum_g \left(y_g - \sum_{i=1}^{q} \beta_i s_{gi} \right)^2 + \lambda \sum_{i=1}^{q} |\beta_i| . \qquad (4)$$

The second term controls the number of non-zero regression coefficients and therefore the model complexity. As the value of λ increases, the number of non-zero coefficients decreases. Model overfitting was evaluated by a standard leave-one-out cross validation (LOOCV) procedure, in which one gene was held out and the other genes were used to build the regression model and predict the gene left-out. For this study, we define the score $D_g = s_{g2} + 0.135$ as the DAMAGES score or D score of gene g, in which the constant is determined by LOOCV.

Estimating specificity and sensitivity in predicting autism-susceptibility genes

We first estimated the number of non-disease causing genes hit by random neutral mutations in the initial list of 112 genes with LGD mutations in probands (Table S1). The five genes with recurrent LGD mutations were considered to be highly confident true positives, given the very small chance to observe such recurrent *de novo* mutations (13, 16, 25). For the remaining genes, the number of non-disease causing genes, or false positives, was estimated based on the relative frequency of LGD mutations in siblings. A case-control design was used in three studies, and the number of false positives in each study was estimated separately. For one study (15), no sibling controls were included, so the number of false positives was estimated from the false discovery rate (FDR) of the other three studies pooled together. Overall, 53 genes with non-recurrent LGD mutations (42% of 127 genes) were estimated to be false positives. Therefore, we estimated that among the 112 genes with LGD mutations represented in the microarray dataset, there are ~65 disease-causing genes and ~47 non-disease genes, respectively.

To assess the single-LGD candidate gene prioritization performance of D score, ExAC metrics, and ensemble score, we used estimated background mutation rate (30, 68) to estimate precision and recall rate. Specifically, for each gene set (with G genes) defined by various metrics, we estimated the number of true positive (i.e. disease-causing; M_T) LGD mutations based on the observed number (M_1) of LGD variants in N cases and the expected number of variants (M_0) given the background LGD mutation rate (R_i, i indexes genes):

$$M_0 = 2N \sum_{i=1}^{G} R_i \quad (5)$$

$$M_T = M_1 - M_0 \quad (6)$$

We denote the total number of true positives in all genes as M, and estimate sensitivity (recall) in each gene set by

$$S = M_T/M \quad (7)$$

and precision by

$$P = M_T/M_1 \quad (8)$$

F-measure combines precision and recall by their harmonic mean:

$$F = 2PS/(P + S) \quad (9)$$

We note that genes with recurrent mutations in ASD patients and genes with LGD mutations in controls, which were used to build ensemble regression model, were not used to estimate the precision and recall in this analysis.

Gene ontology and protein function analysis

Gene ontology (GO) analysis was performed using DAVID (41), using all protein-coding genes represented on the microarray as background.

Statistical analysis

All statistical tests and logistic regression were performed using the R software.

Acknowledgements

We would like to thank members of the Zhang and Shen labs for comments on the manuscript. This work was supported by grants from Simons Foundation (SFARI# 307711 to CZ) and NIH (R00GM95713 to CZ and U01HG008680 to YS). Computation was supported by NIH grants S10OD012351 and S10OD021764.

References

1 Devlin, B. and Scherer, S.W. (2012) Genetic architecture in autism spectrum disorder. *Curr Opin Genet Dev*, **22**, 229-237.

2 Newschaffer, C.J., Croen, L.A., Daniels, J., Giarelli, E., Grether, J.K., Levy, S.E., Mandell, D.S., Miller, L.A., Pinto-Martin, J., Reaven, J. *et al.* (2007) The epidemiology of autism spectrum disorders. *Annu Rev Public Health*, **28**, 235-258.

3 Kim, Y.S., Leventhal, B.L., Koh, Y.-J., Fombonne, E., Laska, E., Lim, E.-C., Cheon, K.-A., Kim, S.-J., Kim, Y.-K., Lee, H. *et al.* (2011) Prevalence of autism spectrum disorders in a total population sample. *Am. J. Psychiatry*, **168**, 904-912.

4 Ronemus, M., Iossifov, I., Levy, D. and Wigler, M. (2014) The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet*, **15**, 133-141.

5 Kelleher, R.J., 3rd and Bear, M.F. (2008) The autistic neuron: troubled translation? *Cell*, **135**, 401-406.

6 Verkerk, A.J.M.H., Pieretti, M., Sutcliffe, J.S., Fu, Y.-H., Kuhl, D.P.A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F. *et al.* (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, **65**, 905-914.

7 Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. *et al.* (2007) Strong association of de novo copy number mutations with autism. *Science*, **316**, 445-449.

8 Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S. *et al.* (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**, 368-372.

9 Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y. *et al.* (2008) Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet*, **82**, 477-488.

10 Levy, D., Ronemus, M., Yamrom, B., Lee, Y.-h., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K. *et al.* (2011) Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*, **70**, 886-897.

11 Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A. *et al.* (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron*, **70**, 863-885.

12 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.

13 Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L. *et al.* (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, **485**, 237-241.

14 O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D. *et al.* (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, **485**, 246-250.

15 Neale, B.M., Kou, Y., Liu, L., Ma/'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V. *et al.* (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, **485**, 242-245.

16 Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.-h., Narzisi, G., Leotta, A. *et al.* (2012) De novo gene disruptions in children on the autistic spectrum. *Neuron*, **74**, 285-299.

17 Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E. *et al.* (2014) The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, **515**, 216-221.

18 De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S. *et al.* (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, **515**, 209-215.

19 Jiang, Y.H., Yuen, R.K., Jin, X., Wang, M., Chen, N., Wu, X., Ju, J., Mei, J., Shi, Y., He, M. *et al.* (2013) Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am J Hum Genet*, **93**, 249-263.

20 Turner, T.N., Hormozdiari, F., Duyzend, M.H., McClymont, S.A., Hook, P.W., Iossifov, I., Raja, A., Baker, C., Hoekzema, K., Stessman, H.A. *et al.* (2016) Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am J Hum Genet*, **98**, 58-74.

21 Gratten, J., Visscher, P.M., Mowry, B.J. and Wray, N.R. (2013) Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. *Nat Genet*, **45**, 234-238.

22 Zhao, X., Leotta, A., Kustanovich, V., Lajonchere, C., Geschwind, D.H., Law, K., Law, P., Qiu, S., Lord, C., Sebat, J. *et al.* (2007) A unified genetic theory for sporadic and inherited autism. *Proc Natl Acad Sci U S A*, **104**, 12831-12836.

23 Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S. and Geschwind, D.H. (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*, **155**, 1008-1021.

24 Gilman, S.R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M. and Vitkup, D. (2011) Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron*, **70**, 898-907.

25 Willsey, A.J., Sanders, Stephan J., Li, M., Dong, S., Tebbenkamp, Andrew T., Muhle, Rebecca A., Reilly, Steven K., Lin, L., Fertuzinhos, S., Miller, Jeremy A. *et al.* (2013) Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell*, **155**, 997-1007.

26 Krumm, N., O'Roak, B.J., Shendure, J. and Eichler, E.E. (2014) A de novo convergence of autism genetics and molecular neuroscience. *Trends Neurosci*, **37**, 95-105.

27 SimonsFoundation. (2016), in press.

28 O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K. *et al.* (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*, **338**, 1619-1622.

29 Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. and Goldstein, D.B. (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet*, **9**, e1003709.

30 Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnstrom, K., Mallick, S., Kirby, A. *et al.* (2014) A framework for the interpretation of de novo mutation in human disease. *Nat Genet*, **46**, 944-950.

31 Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Fennell, T., O'Donnell-Luria, A., Ware, J., Hill, A., Cummings, B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*, in press.

32 Chang, J., Gilman, S.R., Chiang, A.H., Sanders, S.J. and Vitkup, D. (2015) Genotype to phenotype relationships in autism spectrum disorders. *Nat Neurosci*, **18**, 191-198.

33 Xu, X., Wells, A.B., O'Brien, D.R., Nehorai, A. and Dougherty, J.D. (2014) Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **34**, 1420-1431.

34 Doyle, J.P., Dougherty, J.D., Heiman, M., Schmidt, E.F., Stevens, T.R., Ma, G., Bupp, S., Shrestha, P., Shah, R.D., Doughty, M.L. *et al.* (2008) Application of a translational profiling approach for the comparative analysis of CNS cell types. *Cell*, **135**, 749-762.

35 Heiman, M., Schaefer, A., Gong, S., Peterson, J.D., Day, M., Ramsey, K.E., Suárez-Fariñas, M., Schwarz, C., Stephan, D.A., Surmeier, D.J. *et al.* (2008) A translational profiling approach for the molecular characterization of CNS cell types. *Cell*, **135**, 738-748.

36 Duda, R.O., Hart, P.E. and Stork, D.G. (2000) *Pattern Classification*. Wiley.

37 Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B*, **58**, 267-288.

38 Basu, S.N., Kollu, R. and Banerjee-Basu, S. (2009) AutDB: a gene reference resource for autism research. *Nucleic Acids Res*, **37**, D832-D836.

39 Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U. and Zoghbi, H.Y. (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet*, **23**, 185-188.

40 Crino, P.B., Nathanson, K.L. and Henske, E.P. (2006) The tuberous sclerosis complex. *N Engl J Med*, **355**, 1345-1356.

41 Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H. and Lempicki, R. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*, **4**, R60.

42 Ben-David, E. and Shifman, S. (2013) Combined analysis of exome sequencing points toward a major role for transcription regulation during brain development in autism. *Mol Psychiatry*, **18**, 1054-1056.

43 Voineagu, I., Wang, X., Johnston, P., Lowe, J.K., Tian, Y., Horvath, S., Mill, J., Cantor, R.M., Blencowe, B.J. and Geschwind, D.H. (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, **474**, 380-384.

44 Sudhof, T.C. (2008) Neuroligins and neurexins link synaptic function to cognitive disease. *Nature*, **455**, 903-911.

45 Lim, S.-H., Kwon, S.-K., Lee, M.K., Moon, J., Jeong, D.G., Park, E., Kim, S.J., Park, B.C., Lee, S.C., Ryu, S.-E. *et al.* (2009) Synapse formation regulated by protein tyrosine phosphatase receptor T through interaction with cell adhesion molecules and Fyn. *EMBO J*, **28**, 3564-3578.

46 Clark, B.D., Kwon, E., Maffie, J., Jeong, H.-Y., Nadal, M., Strop, P. and Rudy, B. (2008) DPP6 localization in brain supports function as a Kv4 channel associated protein. *Front Mol Neurosci*, **1**, doi: 10.3389/neuro.3302.3008.2008.

47 Won, H., Lee, H.-R., Gee, H.Y., Mah, W., Kim, J.-I., Lee, J., Ha, S., Chung, C., Jung, E.S., Cho, Y.S. *et al.* (2012) Autistic-like social behaviour in Shank2-mutant mice improved by restoring NMDA receptor function. *Nature*, **486**, 261-265.

48 Jacobs, S., Ruusuvuori, E., Sipilä, S.T., Haapanen, A., Damkier, H.H., Kurth, I., Hentschke, M., Schweizer, M., Rudhard, Y., Laatikainen, L.M. *et al.* (2008) Mice with targeted Slc4a10 gene disruption have small brain ventricles and show reduced neuronal excitability. *Proc Natl Acad Sci U S A*, **105**, 311-316.

49 Horev, G., Ellegood, J., Lerch, J.P., Son, Y.-E.E., Muthuswamy, L., Vogel, H., Krieger, A.M., Buja, A., Henkelman, R.M., Wigler, M. *et al.* (2011) Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proc Natl Acad Sci U S A*, **doi: 10.1073/pnas.1114042108**.

50 Crepel, A., Steyaert, J., De la Marche, W., De Wolf, V., Fryns, J.-P., Noens, I., Devriendt, K. and Peeters, H. (2011) Narrowing the critical deletion region for autism spectrum disorders on 16p11.2. *Am J Med Genet B Neuropsychiatr Genet*, **156**, 243-245.

51 Golzio, C., Willer, J., Talkowski, M.E., Oh, E.C., Taniguchi, Y., Jacquemont, S., Reymond, A., Sun, M., Sawa, A., Gusella, J.F. *et al.* (2012) KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature*, **485**, 363-367.

52 Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.X. *et al.* (2015) Excess of rare, inherited truncating mutations in autism. *Nat Genet*, **47**, 582-588.

53 McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M. *et al.* (2016) Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. *bioRxiv*, in press.

54 Deciphering Developmental Disorders, S. (2015) Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, **519**, 223-228.

55 Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K. and Liu, X. (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics*, **24**, 2125-2137.

56 Adzhubei, I., Jordan, D.M. and Sunyaev, S.R. (2013) Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, **Chapter 7**, Unit7 20.

57 Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M. and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, **46**, 310-315.

58 Smith, R.M. and Sadee, W. (2011) Synaptic signaling and aberrant RNA splicing in autism spectrum disorders. *Frontiers in synaptic neuroscience*, **3**, doi: 10.3389/fnsyn.2011.00001.

59 Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K.Y.S., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W. *et al.* (2011) FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, **146**, 247-261.

60 Zhang, C., Frias, M.A., Mele, A., Ruggiu, M., Eom, T., Marney, C.B., Wang, H., Licatalosi, D.D., Fak, J.J. and Darnell, R.B. (2010) Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls. *Science*, **329**, 439-443.

61 Weyn-Vanhentenryck, S., Mele, A., Sun, S., Yan, Q., Farny, N., Zhang, Z., Xue, C., Silver, P.A., Zhang, M.Q., Krainer, A.R. *et al.* (2014) HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.*, **6**, 1139-1152.

62 Kumar, A., Swanwick, C.C., Johnson, N., Menashe, I., Basu, S.N., Bales, M.E. and Banerjee-Basu, S. (2011) A brain region-specific predictive gene map for autism derived by profiling a reference gene set. *PLoS ONE*, **6**, e28431.

63 Menashe, I., Grange, P., Larsen, E.C., Banerjee-Basu, S. and Mitra, P.P. (2013) Co-expression profiling of autism genes in the mouse brain. *PLoS Comput Biol*, **9**, e1003128.

64 Ehrlich, M. (2012) Huntington's disease and the striatal medium spiny neuron: cell-autonomous andnon-cell-autonomous mechanisms of disease. *Neurotherapeutics*, **9**, 270-284.

65 Peca, J., Feliciano, C., Ting, J.T., Wang, W., Wells, M.F., Venkatraman, T.N., Lascola, C.D., Fu, Z. and Feng, G. (2011) Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. *Nature*, **472**, 437-442.

66 Tsai, P.T., Hull, C., Chu, Y., Greene-Colozzi, E., Sadowski, A.R., Leech, J.M., Steinberg, J., Crawley, J.N., Regehr, W.G. and Sahin, M. (2012) Autistic-like behaviour and cerebellar dysfunction in Purkinje cell Tsc1 mutant mice. *Nature*, **488**, 647-651.

67 Fatemi, S.H., Aldinger, K., Ashwood, P., Bauman, M., Blaha, C., Blatt, G., Chauhan, A., Chauhan, V., Dager, S., Dickson, P. *et al.* (2012) Consensus paper: pathological role of the cerebellum in autism. *Cerebellum*, **11**, 777-807.

68 Ware, J.S., Samocha, K.E., Homsy, J. and Daly, M.J. (2015) Interpreting de novo Variation in Human Disease Using denovolyzeR. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, **87**, 7 25 21-27 25 15.

Figure legends

Figure 1: A molecular signature differentiates autism-susceptibility genes and non-disease genes.

A. A total of 145 genes, including 112 genes with LGD mutations in probands (blue dots) and 33 genes with LGD mutations in siblings (green dots) are projected onto the two-dimensional space defined by the first two principal components (PCs).

B. The second principal components differentiates autism-susceptibility genes and non-disease genes. In the heatmap on the left, the PC2 score and loading were used to rank genes and arrays respectively. The source of mutation in each gene is indicated with genes shown in the same order. The arrowhead indicates the threshold used for further analysis in this study, and this threshold gives the same list of genes as those predicted by LOOCV using lasso model with λ=5 (panels **C,D** below). The number of genes with D score>0 or <0 is shown on the right.

C. A regularized regression analysis is used to evaluate the relevance of each PC in predicting the source of mutations (i.e., probands versus siblings). Parameter λ controls the number of non-zero regression coefficients and therefore the complexity of the models. Each curve represents the trace of one coefficient when varying values of λ are used. The trace for the coefficient of PC2 is indicated. Dotted lines indicate shrinkage of coefficients to zero.

D. Leave-one-out cross-validation (LOOCV) is used to evaluate the specificity and sensitivity in predicting the source of LGD mutations. Prediction is based on genes ranked by the regression score. Each receiving operating characteristic (ROC) curve is derived from one specific value of the regularizing parameter λ. The best performance is achieved when only PC2 is used for prediction (λ=5). The arrowhead indicates the turning point of prediction performance when different thresholds of varying stringency are used for prediction.

E. Summary of prediction using an expanded list of genes with LGD mutations in ASD patients and unaffected siblings.

Figure 2: The DAMAGES molecular signature refines the list of candidate genes in the SFARI autism gene database.

A. All genes represented in the microarray dataset, except those with average $\log_2$ intensity <4.5, are projected onto the first two PCs, and shown as a smoothed scatter plot. The gray-scale intensity reflects the local density of genes. A total of 483 genes from the SFARI autism gene database represented in the microarray dataset (asterisks) are overlaid. A subset of these genes were manually scored by experts by considering strength of existing evidence, and these scored genes are distinguished using different colors. A subset of syndromic ASD genes and the five strong candidate genes are highlighted.

B. The percentage of scored genes in each group with a positive or negative DAMAGES score (D score) is shown. The color codes are the same as in (**A**). The number of genes in each group is indicated in the parentheses following the gene categories.

Figure 3: The DAMAGES molecular signature reveals CNS cell types associated with autism.

A. The loadings of different cell types on the first two PCs derived from genes with LGD mutations are shown. Each dot represents a cell type. Different colors represent the brain regions used to isolate the specific types of cells, with the same color codes as shown in **Figure S1A**. Neurons, glial cells and unselected RNA samples are represented by triangles, circles, and squares, respectively.

B. The loadings of cell types on PC1 derived from genes with LGD mutations (x-axis) are plotted against that derived from the whole dataset (y-axis). The asterisk indicates cerebellar Grp$^+$ cells that are known to include both unipolar brush cells and Bergmann glial cells (34). The squared Pearson correlation between the two signatures is indicated.

C. Similar to (**B**), except that the loadings on PC2 are plotted.

D. Loadings of all cell types on PC2 derived from genes with LGD mutations (DAMAGES signature) are plotted. The color codes and abbreviation of each brain region are the same as shown in **Figure S1A**. UB: unbound RNA without selection for specific cell types.

Figure 4: Prioritized candidate autism-susceptibility genes with recurrent CNVs in chromosome 16p11.2.

A. A UCSC genome browser view of the region (hg19: chr16:29,350,841-30,433,540) is shown, with *de novo* CNV events displayed above the RefSeq genes. Duplication and deletion CNVs are shown in red and blue, respectively. The dotted box indicates the region with 26 genes affected in almost all CNVs.

B. The 26 genes are ranked by their D scores. Three genes not represented in the microarray data are indicated by n.a..

Figure 5: Comparison and integration of D score with ExAC scores.

A. Distribution of D scores and ExAC LOF Z-scores for genes with recurrent LGD mutations in ASD cases and controls.

B. Performance of prediction as measured by Precision and Recall using different scoring metrics or an ensemble model as a function of varying cutoffs.

C. Similar to (**B**), but the F measure is shown.

D. Distribution of ensemble score among genes in which damaging *de novo* mutations observed (red) in the DDD data set versus the ones not observed (blue).

Table 1: Enrichment of LGD *de novo* mutations in cases among genes grouped by D score.

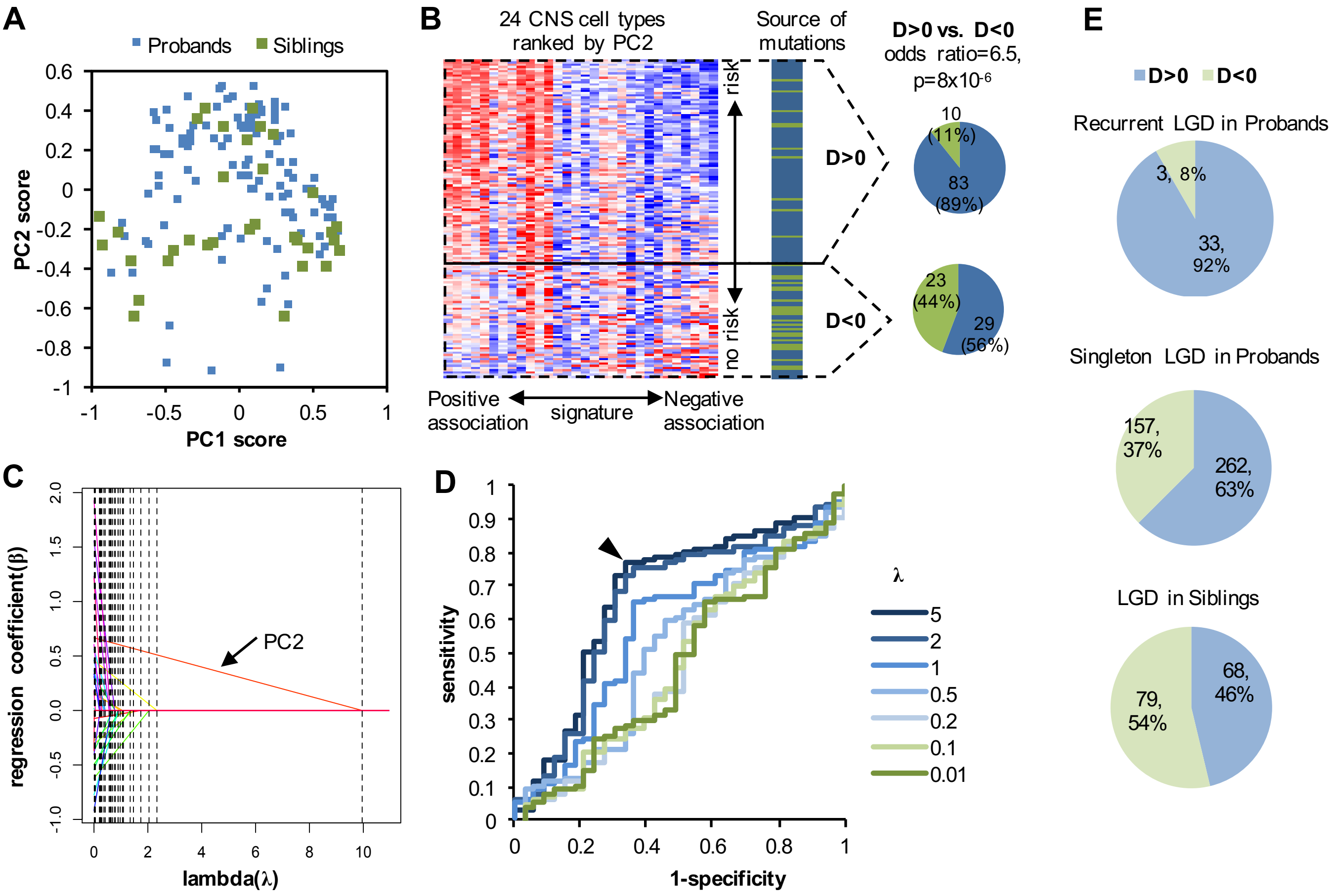
D score rank percentile	Number of LGD mutations in cases	Number of LGD mutations in controls	Rate enrichment	P-value
Top 25% (n-genes = 3928)	257	45	2.76	3.2×10^{-12}
Bottom 75%	318	132	1.16	0.16

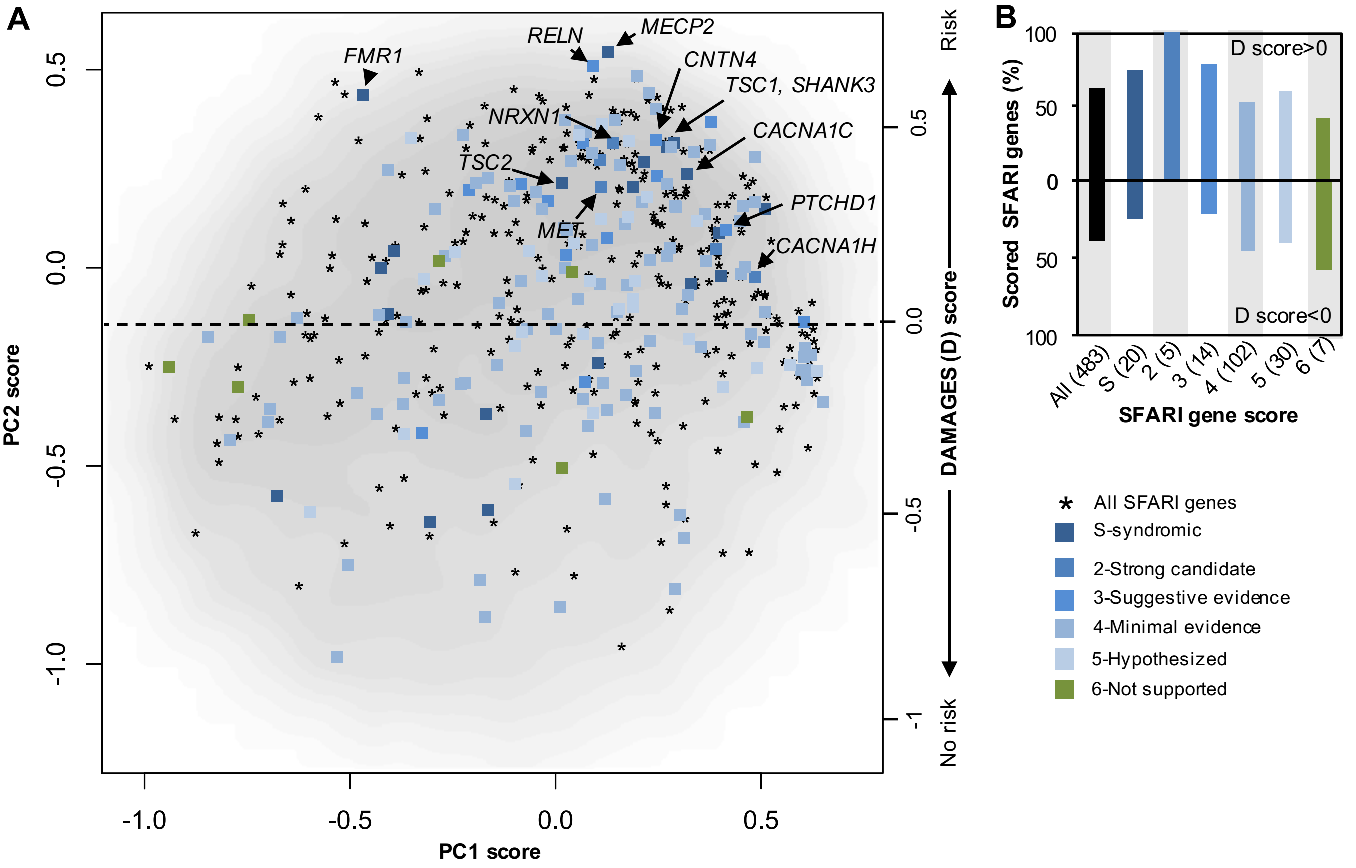
P-values were calculated by binomial tests.

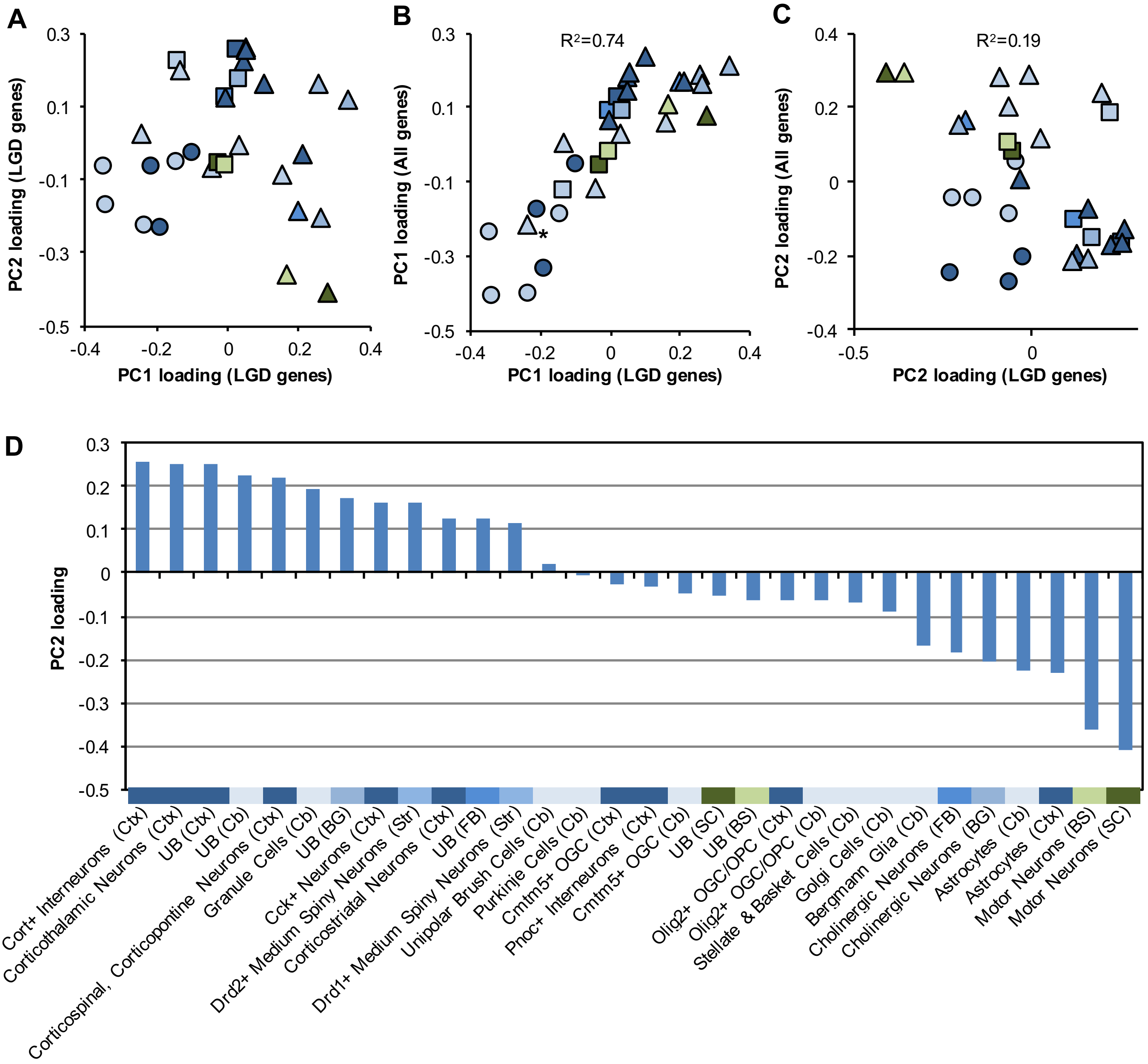
Table 2: Logistic regression for classification of recurrently LGD-mutated genes in ASD and genes LGD-mutated in unaffected siblings.

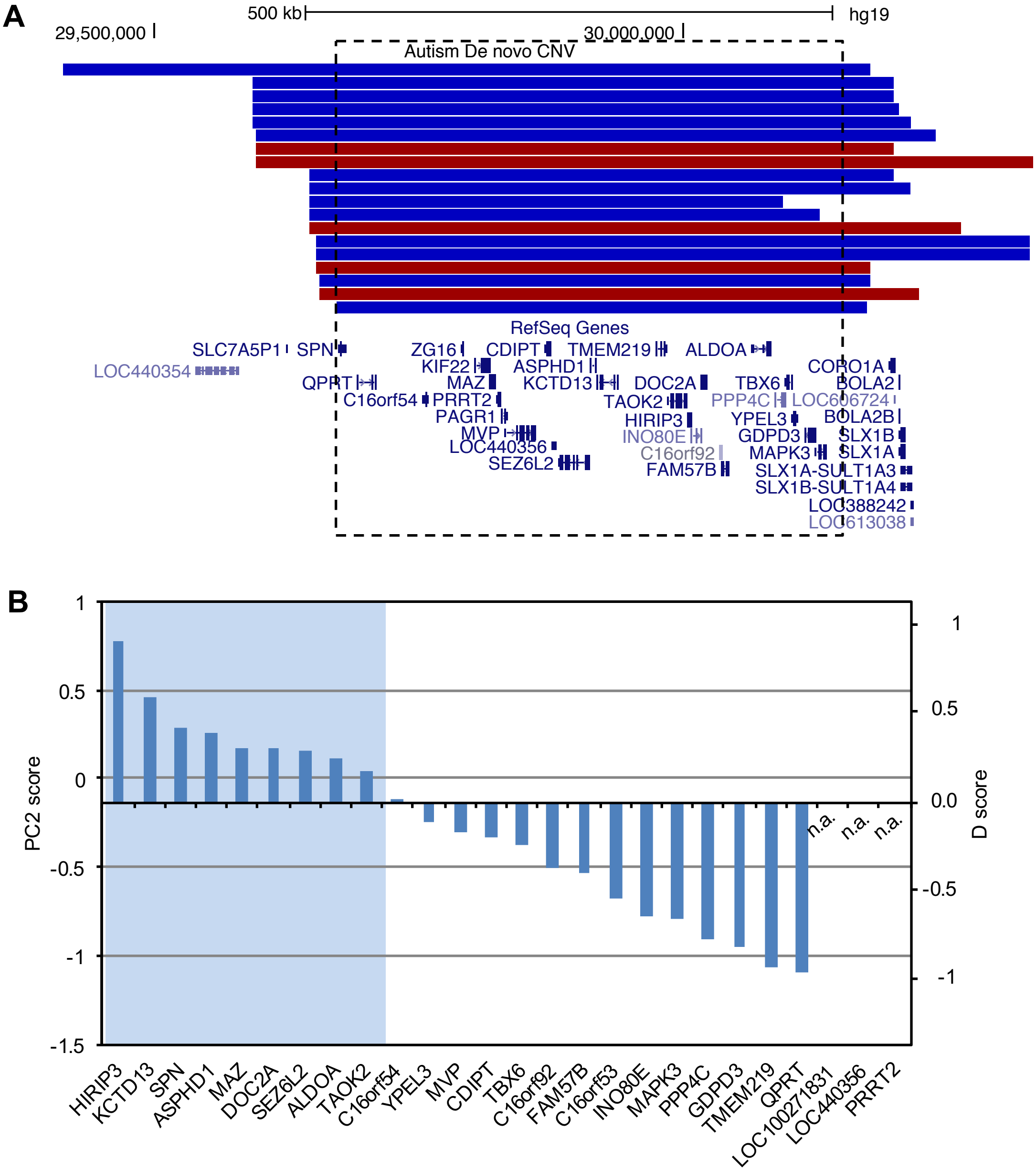
Features	Estimate	P-value
pLI	2.00	0.0029
Mis_z	0.290	0.019
D score	2.90	0.0015

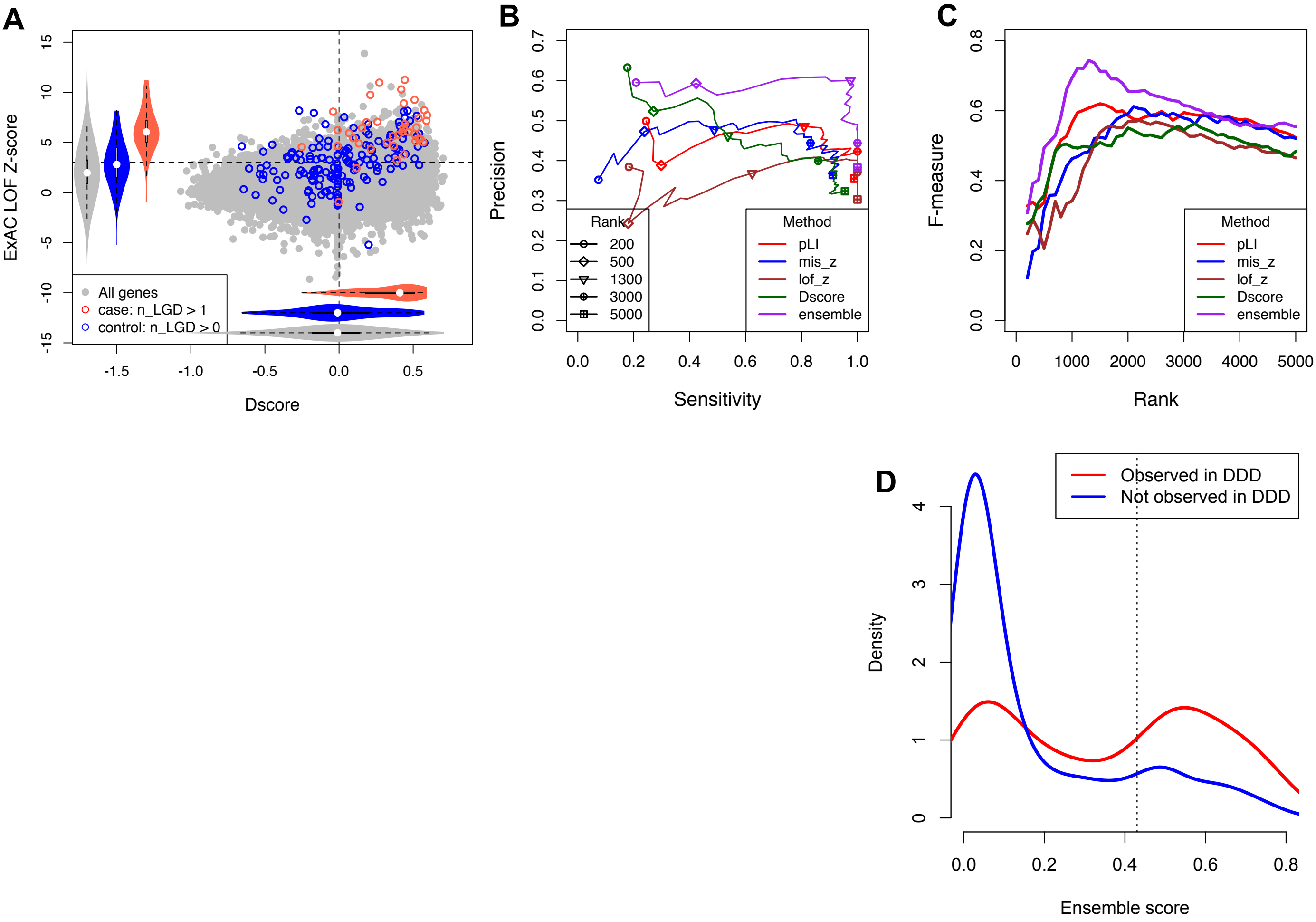
Brain E9.5 is not significant given D score and other features (P-value = 0.92).


A
Probands
Siblings
PC2 score
PC1 score
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1
-0.5
0
0.5
1
B
24 CNS cell types ranked by PC2
Source of mutations
risk
no risk
D>0
D<0
D>0 vs. D<0
odds ratio=6.5,
p=8x10-6
10 (11%)
83 (89%)
23 (44%)
29 (56%)
Positive association
signature
Negative association
C
regression coefficient(β)
lambda(λ)
PC2
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0
2
4
6
8
10
D
sensitivity
1-specificity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
λ
5
2
1
0.5
0.2
0.1
0.01
E
D>0
D<0
Recurrent LGD in Probands
3, 8%
33, 92%
Singleton LGD in Probands
157, 37%
262, 63%
LGD in Siblings
79, 54%
68, 46%


A
B
FMR1
RELN
MECP2
CNTN4
TSC1, SHANK3
NRXN1
CACNA1C
TSC2
PTCHD1
MET
CACNA1H
PC2 score
0.5
0.0
-0.5
-1.0
PC1 score
-1.0
-0.5
0.0
0.5
Risk
No risk
DAMAGES (D) score
0.5
0.0
-0.5
-1
Scored SFARI genes (%)
100
50
0
50
100
D score>0
D score<0
All (483)
S (20)
2 (5)
3 (14)
4 (102)
5 (30)
6 (7)
SFARI gene score
* All SFARI genes
S-syndromic
2-Strong candidate
3-Suggestive evidence
4-Minimal evidence
5-Hypothesized
6-Not supported


A
PC2 loading (LGD genes)
PC1 loading (LGD genes)
0.3
0.1
-0.1
-0.3
-0.5
-0.4
-0.2
0
0.2
0.4
B
R²=0.74
PC1 loading (All genes)
PC1 loading (LGD genes)
*
C
R²=0.19
PC2 loading (All genes)
PC2 loading (LGD genes)
0.4
0.2
0
-0.2
-0.4
-0.5
D
PC2 loading
Cort+ Interneurons (Ctx)
Corticothalamic Neurons (Ctx)
UB (Ctx)
UB (Cb)
Corticospinal, Corticopontine Neurons (Ctx)
Granule Cells (Cb)
UB (BG)
Cck+ Neurons (Ctx)
Drd2+ Medium Spiny Neurons (Str)
Corticostriatal Neurons (Ctx)
UB (FB)
Drd1+ Medium Spiny Neurons (Str)
Unipolar Brush Cells (Cb)
Purkinje Cells (Cb)
Cmtm5+ OGC (Ctx)
Pnoc+ Interneurons (Ctx)
Cmtm5+ OGC (Cb)
UB (SC)
UB (BS)
Olig2+ OGC/OPC (Ctx)
Olig2+ OGC/OPC (Cb)
Stellate & Basket Cells (Cb)
Golgi Cells (Cb)
Bergmann Glia (Cb)
Cholinergic Neurons (FB)
Cholinergic Neurons (BG)
Astrocytes (Cb)
Astrocytes (Ctx)
Motor Neurons (BS)
Motor Neurons (SC)


A
29,500,000
500 kb
30,000,000
hg19
Autism De novo CNV
RefSeq Genes
SLC7A5P1
SPN
ZG16
CDIPT
TMEM219
ALDOA
LOC440354
KIF22
ASPHD1
CORO1A
QPRT
MAZ
KCTD13
DOC2A
TBX6
BOLA2
C16orf54
PRRT2
TAOK2
PPP4C
LOC606724
PAGR1
HIRIP3
YPEL3
BOLA2B
MVP
INO80E
GDPD3
SLX1B
LOC440356
C16orf92
MAPK3
SLX1A
SEZ6L2
FAM57B
SLX1A-SULT1A3
SLX1B-SULT1A4
LOC388242
LOC613038
B
PC2 score
1
0.5
0
-0.5
-1
-1.5
D score
1
0.5
0.0
-0.5
-1
n.a.
n.a.
n.a.
HIRIP3
KCTD13
SPN
ASPHD1
MAZ
DOC2A
SEZ6L2
ALDOA
TAOK2
C16orf54
YPEL3
MVP
CDIPT
TBX6
C16orf92
FAM57B
C16orf53
INO80E
MAPK3
PPP4C
GDPD3
TMEM219
QPRT
LOC100271831
LOC440356
PRRT2


A
ExAC LOF Z-score
15
10
5
0
-5
-10
-15
-1.5
-1.0
-0.5
0.0
0.5
Dscore
All genes
case: n_LGD > 1
control: n_LGD > 0
B
Precision
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.0
0.2
0.4
0.6
0.8
1.0
Sensitivity
Rank
200
500
1300
3000
5000
Method
pLI
mis_z
lof_z
Dscore
ensemble
C
F-measure
0.0
0.2
0.4
0.6
0.8
0
1000
2000
3000
4000
5000
Rank
Method
pLI
mis_z
lof_z
Dscore
ensemble
D
Observed in DDD
Not observed in DDD
Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
Ensemble score