Supplementary Figure 1

Illustration of the working of network-based SVM to confidently predict a new (and now confirmed) ASD gene.

Gene CTNND2's brain network neighborhood that enabled its prediction by the SVM. E1-E4 denote genes with various levels of evidence (high to low) for known association with ASD. CTNND2 gene ranks 16th in our prediction and has been recently discovered as a high-confidence gene associated with ASD in female-enriched multiplex (Turner et al. (2015) Nature 520(7545), 51–6), although our classifier did not see this gene during training (at any level of confidence). In our brain network, CTNND2 is tightly linked to high-evidence genes SHANK2 and NRXN1 (E1), and several lower-evidence genes such as ATP2B2, DPP6 and SNAP25. In addition, it also shares common neighbors with E1 genes SHANK2, GRIN2B and NRXN1. The combination of local connectivity and global interaction pattern together enable the SVM to accurately predict CTNND2 as an ASD-related gene. Both DAWN (Liu et al. (2015) The Annals of Applied Statistics 9(3), 1571–1600) (DAWN-2015 rank 6079/8488) and NETBAG+ (Chang et al. (2014) Nature Neuroscience 18(2), 191–198), recent network-based methods focused on autism, fail to predict this gene's relevance for autism, mainly because previous genetic studies have not strongly linked this gene to ASD. This is evident from CTNND2's high P-value (0.90) using TADA-2014 (De Rubeis et al. (2014) Nature 515(7526), 209–15), a method to summarize prior ASD genetic data (all types of mutations) into gene P-values for ASD-risk. Our method, on the other hand, without any previous genetic evidence about this gene (this gene is also not in our gold standard), ranks it 16th out of more than 25,000 genes in the genome.#
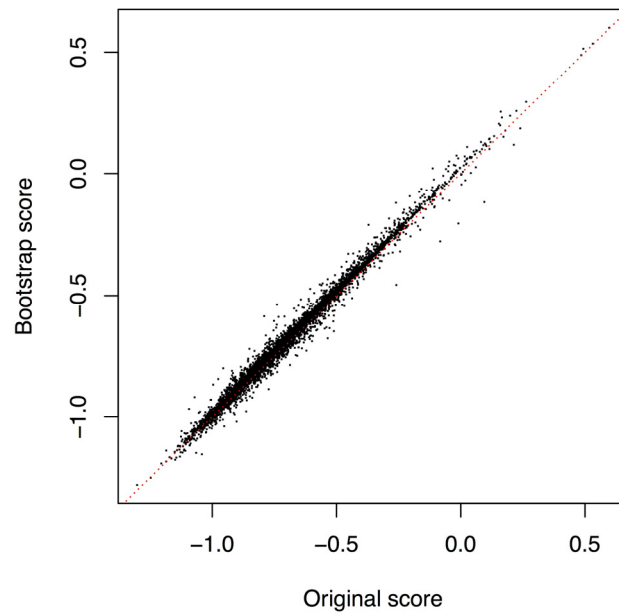
Supplementary Figure 2

Robustness of ASD-gene predictions to changes in our gold-standard gene sets.

To test the robustness of our predictions, we made predictions by subsampling the gold standard, each time using 4/5 of the negatives and positives. The rank based correlation coefficient between our original prediction and the 100 sets of predictions made on the resampled gold standard is 0.993, indicating that our genome-wide predictions are highly robust to noise in the gold standard.
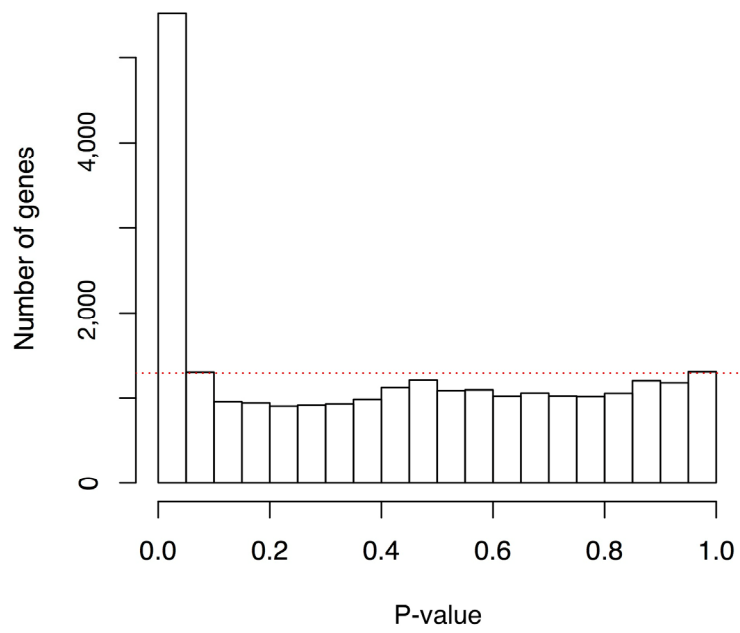
Supplementary Figure 3

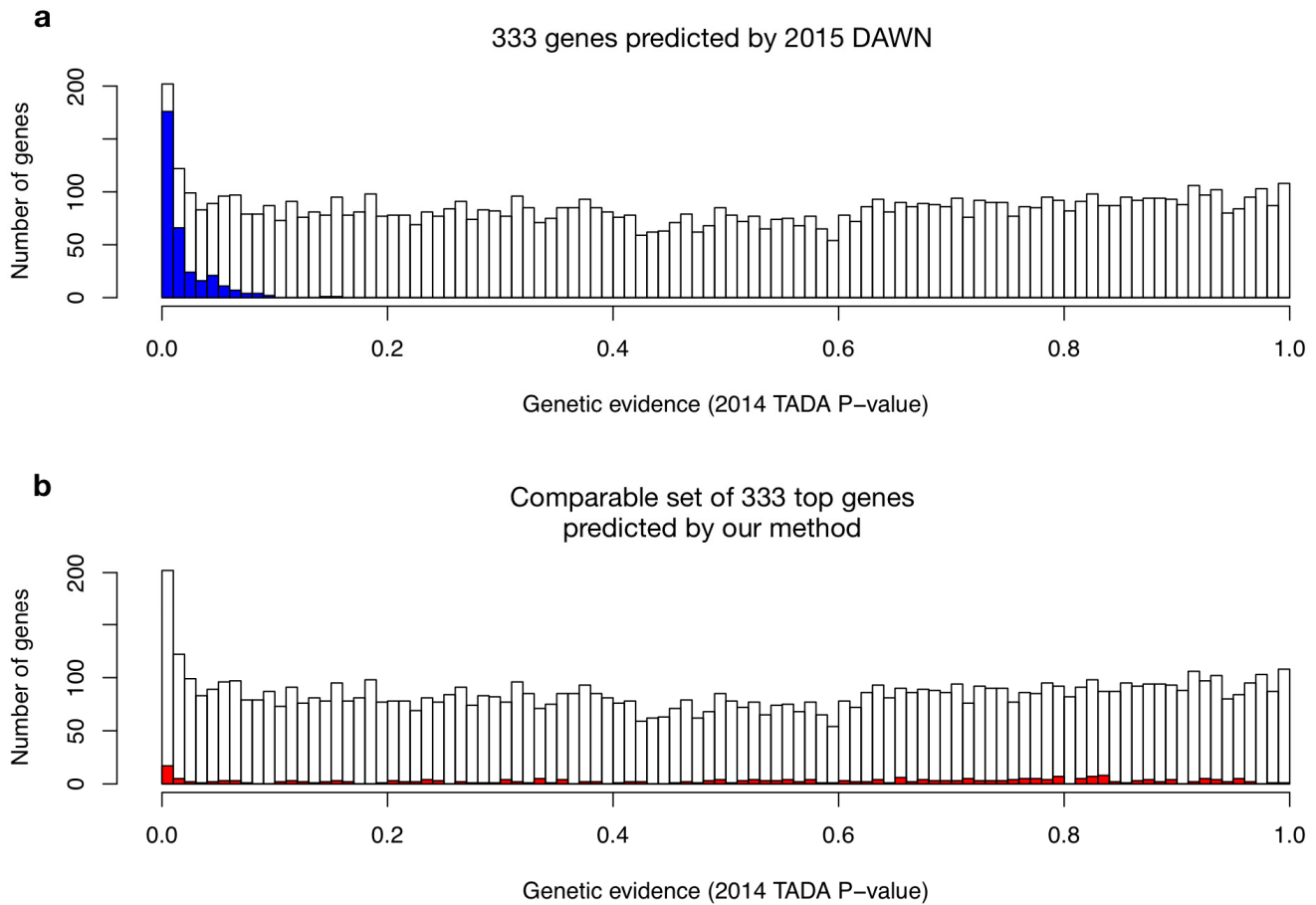Permutation-based *P*-values for our genome-wide predictions.

In order to improve the interpretability of our genome-wide ranking, we calculated a permutation-based *P*-value and a corresponding *Q*-value for each gene (see *Methods*). The plot shows the distribution of these *P*-values, with the red dashed line indicating mean frequency.
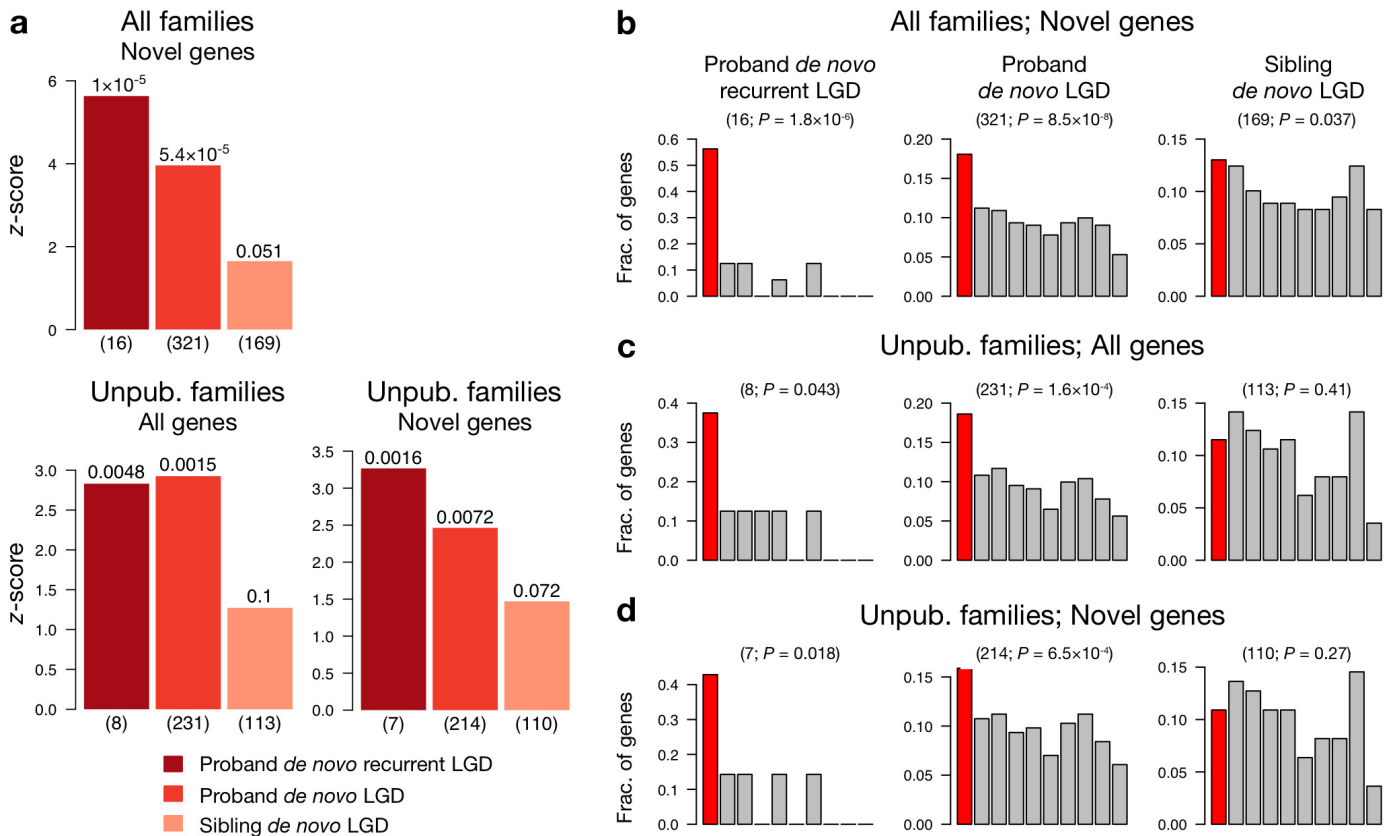
Supplementary Figure 4

Extended evaluation of autism-gene predictions on empirical data from an independent sequencing study.
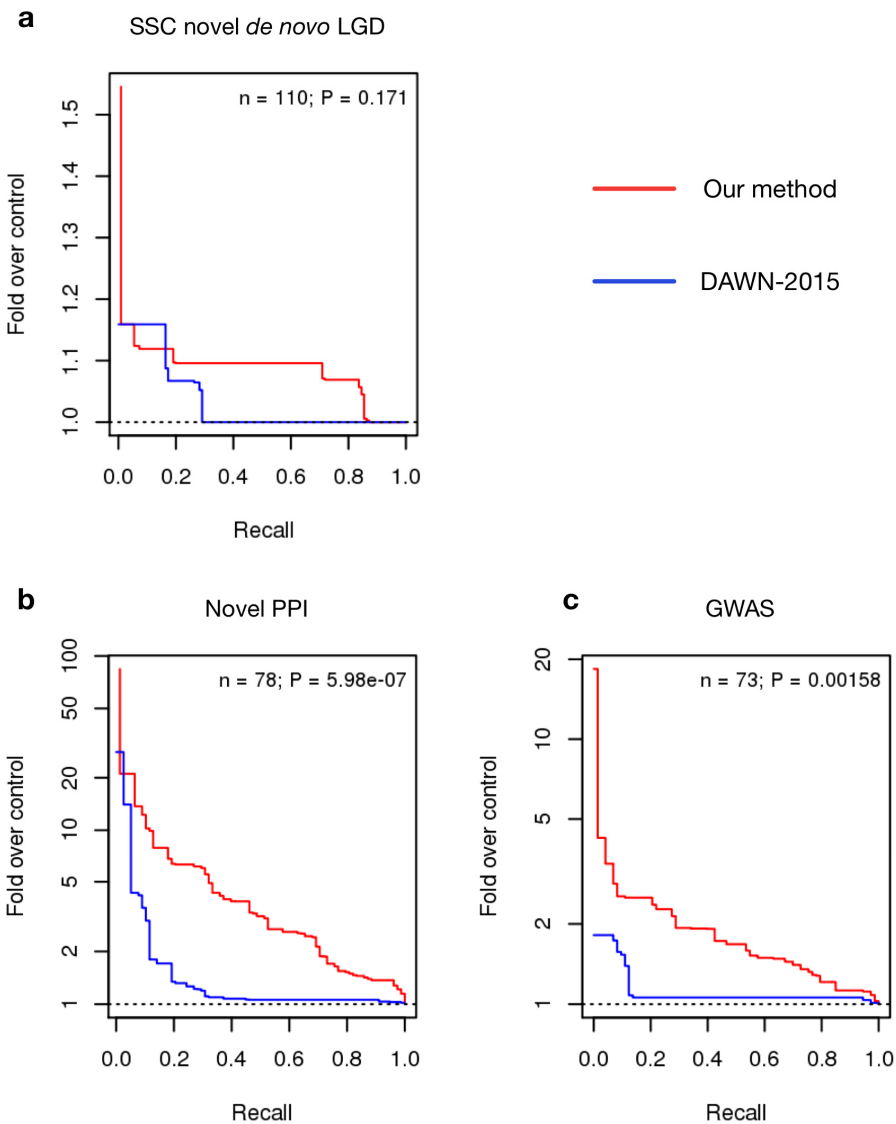
All evaluations below were performed and presented similar to those in Figure 2b and 2c. The resulting trends – significant enrichment of proband LGD mutations and non-enrichment of sibling LGD mutations – are consistent with those observed in Figure 2. **(a)** Rank-based enrichment test (without top-decile cutoffs) on data from all and unpublished families, for 'All' and 'Novel' ASD genes (Fig. 2, and Fig S4 b, c and d), showing trends similar to those presented in Figure 2. 'All families – All genes' is in Figure 2b, and the rest are here. All plots present the z-score quantifying the enrichment of the gene-set of interest towards the top of our genome-wide ranking of genes (see *Methods*). The three mutation gene-sets in each case are colored differently (labeled in the legend below) with the number of genes in parenthesis below. The *P*-values recorded at top of each bar were calculated using a permutation test described in *Methods* in detail. **(b)** Novel *de novo* LGDs from all families: This data set is derived from mutations recorded from all families published in 2014, but restricted to only genes that were not part of our training gold-standard (completely 'novel' ASD genes). The total number of genes in the gene-set and the *P*-value from the binomial test are given in parenthesis just below. **(c)** All *de novo* LGDs from unpublished families: This data set is derived from mutations recorded only from SSC families that were unpublished in the 2012 studies and subsequently published in 2014 (all genes from completely unseen families). **(d)** Novel *de novo* LGDs from unpublished families: This data set is derived from mutations recorded from families that were unpublished in the 2012 studies and subsequently published in 2014, further restricted to only genes that were not part of our training gold-standard (completely 'novel' ASD genes from unseen families).

a

333 genes predicted by 2015 DAWN

Number of genes

Genetic evidence (2014 TADA P−value)

b

Comparable set of 333 top genes
predicted by our method

Number of genes

Genetic evidence (2014 TADA P−value)

Supplementary Figure 5

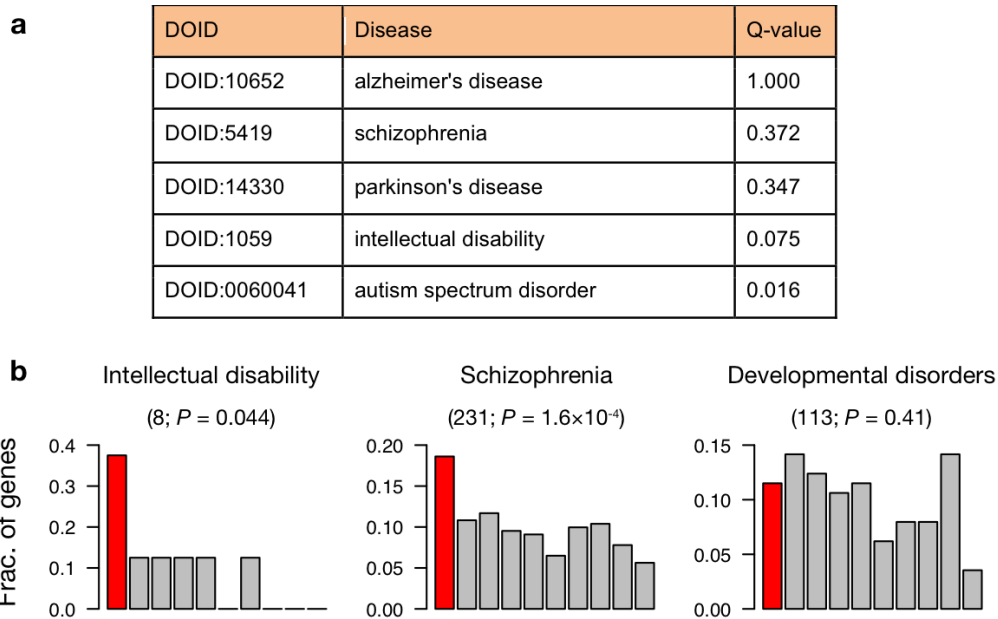Histograms of prior genetic evidence scores and the distributions of top genes as predicted by DAWN-2015 and our method.

(a) Distribution of 2014 TADA (De Rubeis et al. (2014) Nature 515(7526), 209–15) *P*-values of 8488 genes (white), which is the input genetic evidence for DAWN-2015 (Liu et al. (2015) The Annals of Applied Statistics 9(3), 1571–1600). Overlaid on top (blue) is the distribution of TADA *P*-values of DAWN's top 333 predicted genes. **(b)** Similar to (a), here overlaid on top (red) with the distribution of 2014 TADA *P*-values of a comparable set of our top 333 predicted genes.

a SSC novel *de novo* LGD

n = 110; P = 0.171

—— Our method

—— DAWN-2015

b Novel PPI

n = 78; P = 5.98e-07

c GWAS

n = 73; P = 0.00158

Supplementary Figure 6

Comparison of our method to a previous method (DAWN).

We compare our method to DAWN-2015 (Liu et al. (2015) The Annals of Applied Statistics 9(3), 1571–1600) by evaluating the ASD gene ranking produced by the two methods in their ability to prioritize (**a**) novel LGD-targets, (**b**) novel protein-protein interaction (PPI) partners, and (**c**) ASD-associated genes identified by genome-wide association studies (GWAS). Precision is presented as fold-over-random, measured as observed precision over expected baseline-precision (calculated as Precision/[P/(P+N)], where P is the number of positives and N is the number of negatives). For (**a**), gene targets of novel *de novo* LGDs observed in SSC probands and siblings were, respectively, used as positive and negative examples. For (**b**), novel PPI partners of potential ASD genes identified in a genome-wide assay (Corominas et al. (2014) Nature Communications 5) were used as positives, and all other proteins tested in that assay were used as negatives. For (**c**), genes associated with autism as documented in the GWAS catalog were used as positives, while all other genes in the catalogue were used as negatives. *P*-values were calculated by Wilcoxon rank-sum test.

a

DOID	Disease	Q-value
DOID:10652	alzheimer's disease	1.000
DOID:5419	schizophrenia	0.372
DOID:14330	parkinson's disease	0.347
DOID:1059	intellectual disability	0.075
DOID:0060041	autism spectrum disorder	0.016

b

Intellectual disability $(8; P = 0.044)$ — Schizophrenia $(231; P = 1.6 \times 10^{-4})$ — Developmental disorders $(113; P = 0.41)$

Supplementary Figure 7

Specificity of our genome-wide ranking to ASD.

(a) Enrichment of ASD-gene ranking on genes associated with various neurological/brain diseases. To test the specificity of our genome-wide ranking to ASD, we tested the ranking of genes associated with five neurological diseases annotated in the OMIM database. We found no significant enrichment for unrelated diseases, indicating that our top-ranked genes are indeed specific for ASD, and distinct from genes associated with other neuronal disease. **(b)** Evaluation of ASD-gene ranking on genes linked to disorders closely related to ASD. We obtained genes associated with intellectual disability (ID; left; (Parikshak et al. (2013) Cell 155(5), 1008–21)), schizophrenia (middle; (Fromer et al. (2014) Nature 506(7487), 179–84)), and developmental disorders (DDD; right; (TDDD Study (2015) Nature 519(7542), 223–8)) identified in large sequencing studies, removed the genes in our ASD positive gold standards, and test their distribution in our genome-wide ASD-gene ranking. We observe the expected significant overlap with ID and schizophrenia, while noting that our ranking prioritizes hundreds of additional genes not implicated in these disorders. The enrichment of genes in the DDD dataset is not surprising because the underlying cohort includes cases with ASD as well as several other disorders that have comorbidity with ASD (e.g. ID, heart development disorders).

a RVIS score distribution   **b** Constrained gene distribution
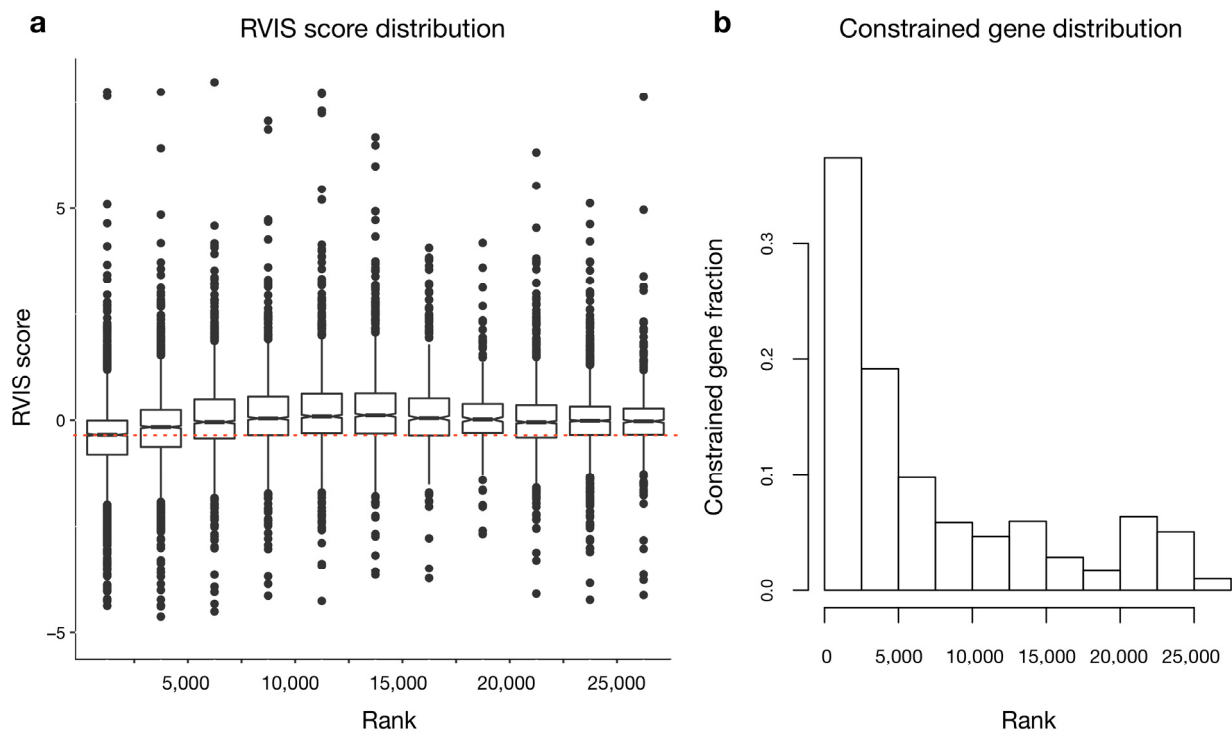
Supplementary Figure 8

Top decile genes tend to be more constrained (have less common variation).

(a) Boxplot of distribution of RVIS score (Petrovski et al. (2013) PLoS Genetics 9(8), e1003709) as a function of our autism gene ranking. **(b)** Histogram of fraction of constrained genes (out of 1,003) (Samocha et al. (2014) Nature Genetics 46(9), 944–50), along autism gene ranking. Testing the two constrained sets against our predictions (see *Methods*) showed similar results both within the top-decile (using Fisher's exact test) or without any cutoff (using a rank-based permutation test): top-ranked genes do tend to be more constrained relative to all the other genes (RVIS, top-decile Wilcoxon test $P < 2.2e\text{-}16$, rank-based permutation test $P = 1e\text{-}6$; Constrained set, Fisher's exact test $P = 1.36e\text{-}86$, rank-based permutation test $P = 1e\text{-}6$).
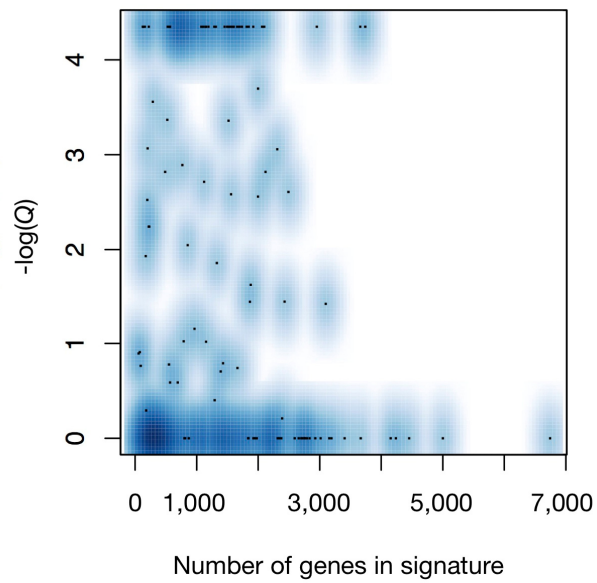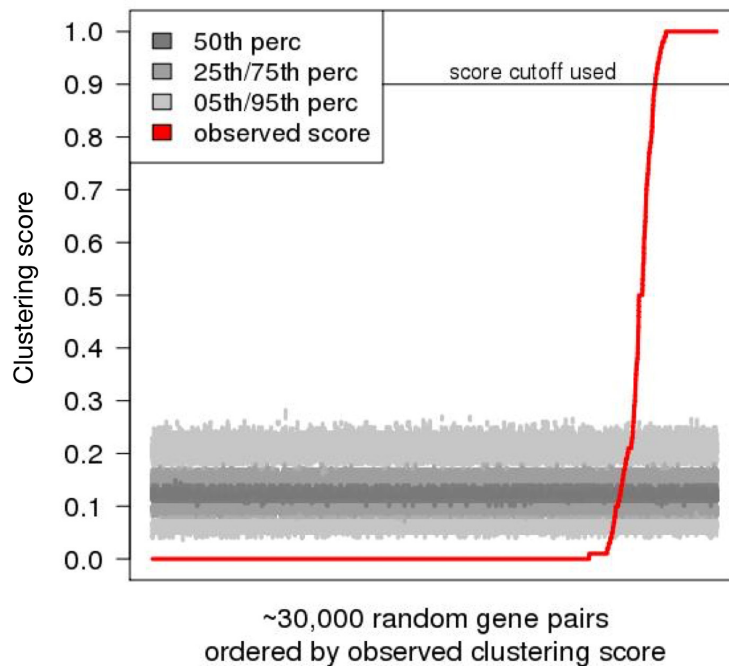
Supplementary Figure 9

Spatiotemporal analysis is not biased by windows with dramatic changes.

Spatiotemporal signature for each window was derived by controlling for both brain region and developmental stage. The permutation test used to identify association of each signature with ASD also controls for the number of genes in that signature. The plot shows that ASD-association of all signatures (indicated using the negative logarithm of the enrichment Q-value) as a function of the number of genes in each signature, demonstrating that there is no correlation between signature-size and ASD-association.

Supplementary Figure 10

Statistical analysis of ASD-gene clustering in the brain network.

Results of a permutation test to evaluate the clustering of the ASD genes in a randomized brain network. Shared k-nearest-neighbor-based analysis was repeated a 100 times, each time randomizing the k-nearest neighbors of each node in the network of top ASD genes. For a random set of nearly 30,000 gene pairs, the bulk of the clustering scores ranged between 0 and 0.3 based on random k neighbors, significantly lower than the cutoff of 0.9 used in our analysis of the real network. Our ASD functional modules thus are significantly and substantially more cohesive than random.
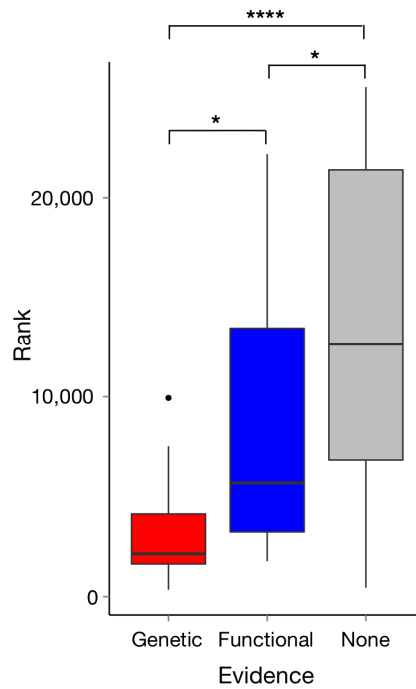
Supplementary Figure 11

Comparison of predicted ASD ranks of genes within autism-associated CNVs that have prior genetic and functional evidence.

Boxplots show the distribution of ASD ranks of genes within the 8 ASD-associated CNVs that have different types of prior evidence: genetic (strong; red; $n = 13$), functional (weak; blue; $n = 8$), or none (grey; $n = 166$). *: $P \leq 0.05$; ***: $P \leq 0.001$.
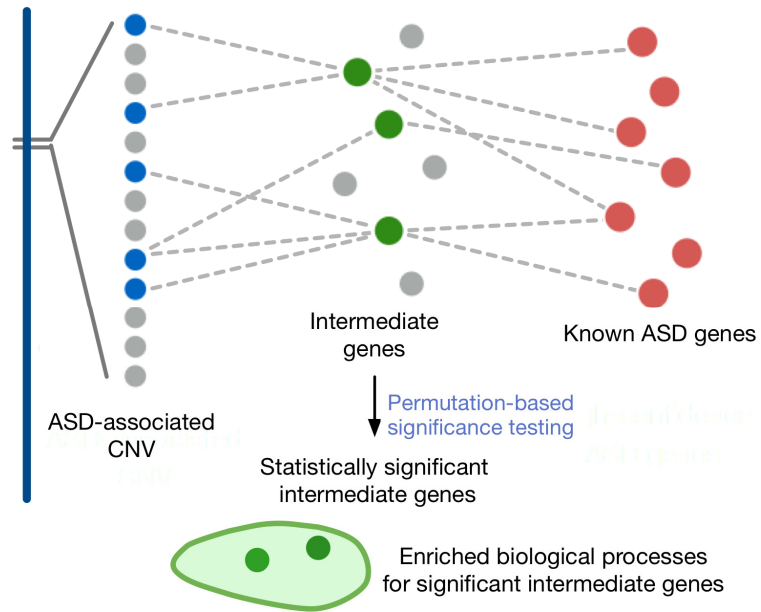
Supplementary Figure 12

Illustration of CNV diagram.

Top decile genes within ASD-associated CNVs (blue circles), and known ASD genes (red circles) are linked in the brain network via intermediate genes (green circles). Statistically significant intermediate genes are identified using a permutation test against random genomic intervals. The biological processes enriched among significant intermediate genes (green bubble) that are shared with multiple ASD-associated CNVs are shown in Figure 6.
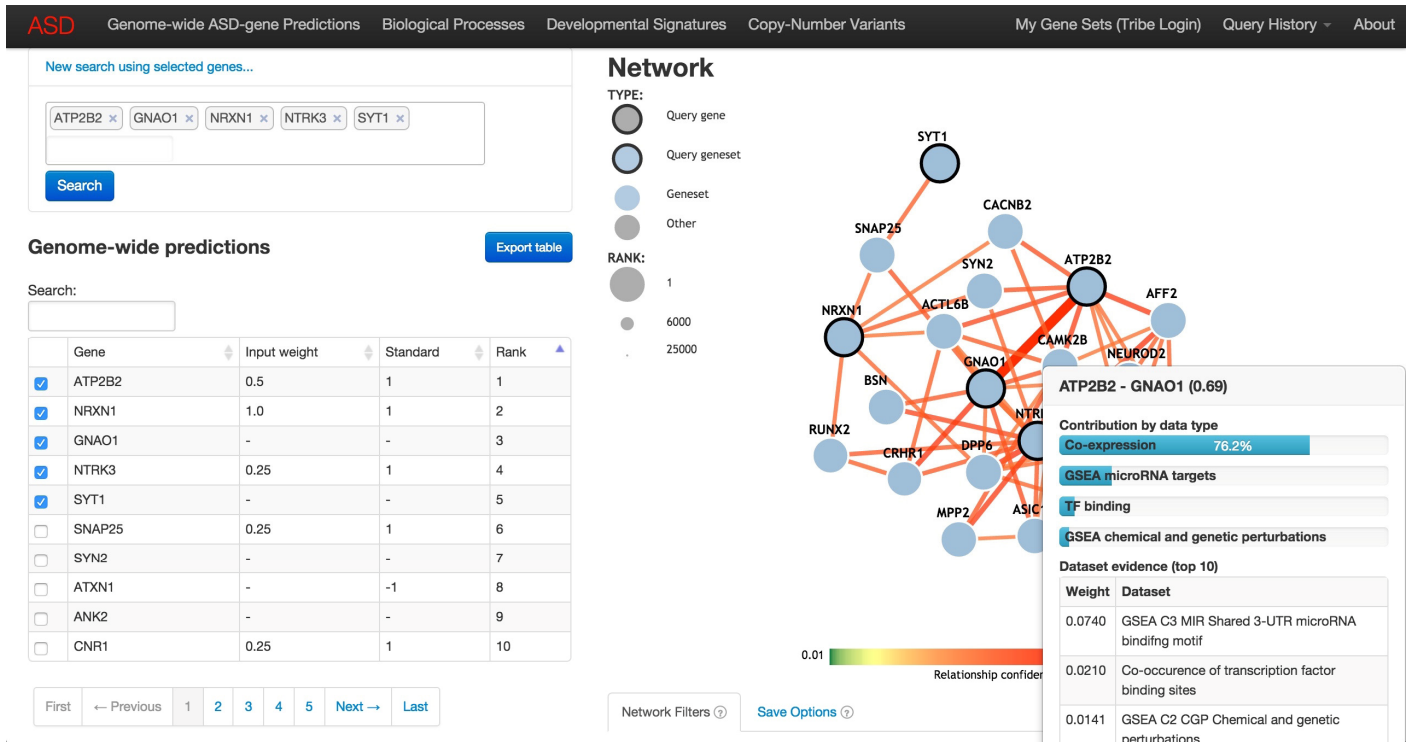
Supplementary Figure 13

Illustration of network visualization in the ASD web server.

The interactive ASD web-interface enables biologists to explore their genes of interest (left) in the human brain-specific network (right). Users can easily explore the contribution of each of the different data types to any predicted brain-specific interaction (window at bottom right), including a summary of which data types contributed the most – for example, co-expression and GSEA miRNA targets in this case – as well as evidence weights for each individual dataset. Users can click on any dataset to be redirected to the underlying data.
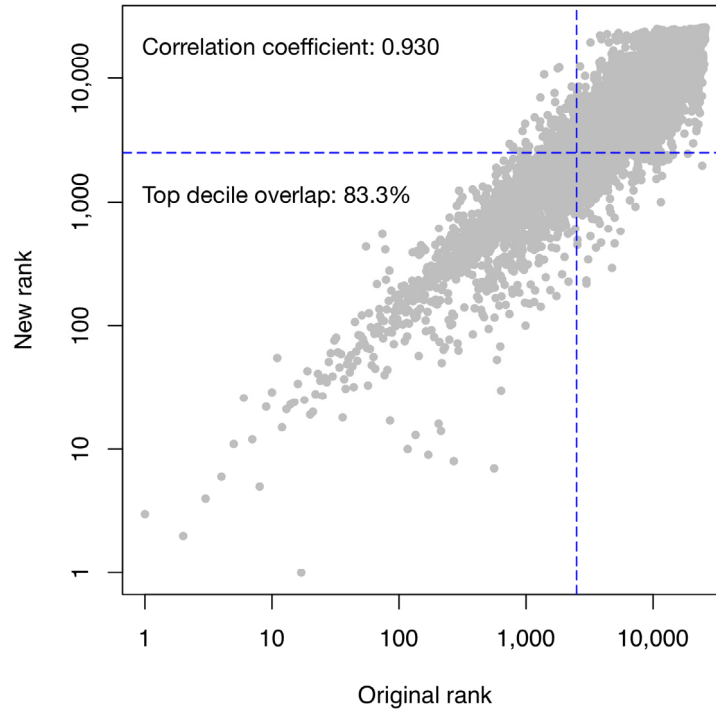
Supplementary Figure 14

New predictions based on training with updated gold standards correspond closely to the predictions used in this study.

As an example of how we can regularly update the results in our web-server based on newly identified genes, we have made available a new set of predictions at http://asd.princeton.edu/v2 by training on an updated version of the SFARI gene database that includes all the results from the 2014 study (Iossifov et al. (2014) Nature 515(7526), 216–21). The scatter plot shows the original (v1) and new ranks (v2) of each gene in the genome. The dotted lines correspond to the top-decile of the predicted genes. The new predictions are overall quite consistent with the original one used throughout this study, having a correlation coefficient of 0.93 between the genome-wide rankings with 83.3% of top decile genes consistent between v1 and v2. In addition, as our web server demonstrates, analyses results (including GO enrichments of predictions and neurodevelopmental analysis) also remain nearly identical.
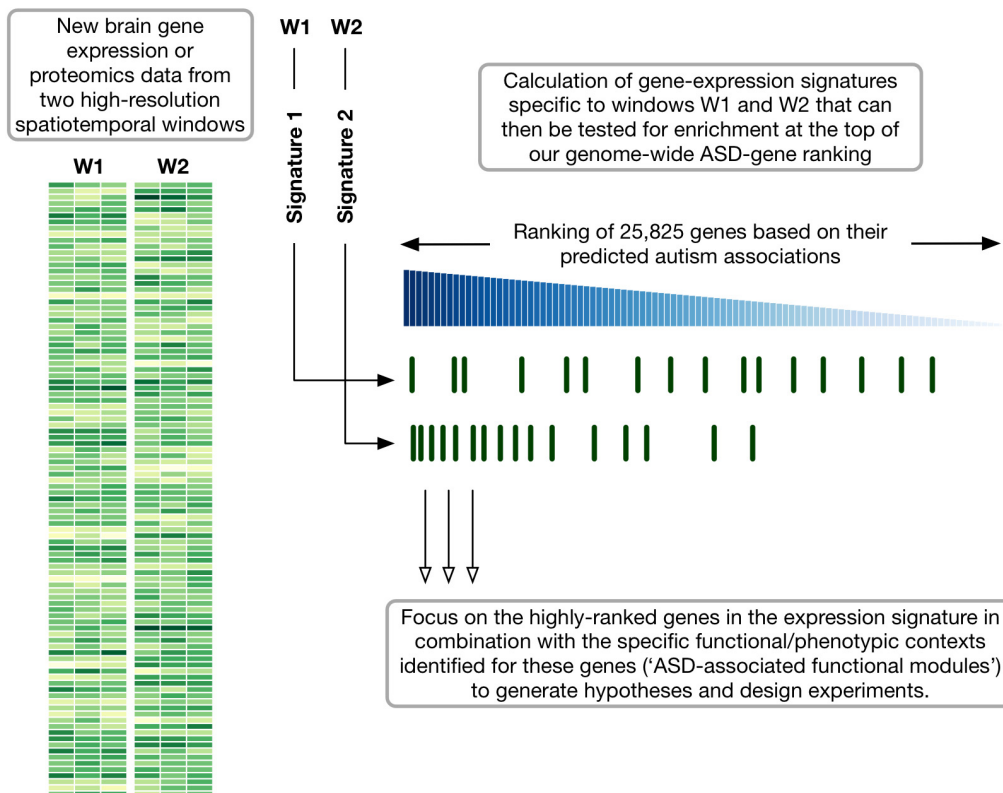
Supplementary Figure 15

A potential use case: estimating the relevance of a high-resolution spatiotemporal window in the brain to autism.

A biomedical researcher could use our predictions as a framework for analyzing their data from high- or low-throughput assays, allowing high-resolution study of autism genetics in the functional and physiological contexts of their interest. For example, a researcher who has generated gene expression or proteomics data from a new high-resolution spatiotemporal window in the brain (either human or in a model organism) can use our approach to assess the molecular relevance of that window to ASD. This approach – identifying a characteristic gene signature from the samples and estimating its enrichment in our genome-wide ASD-gene ranking (similar to results presented in Figure 3) – can provide results of increasing specificity with increasing spatiotemporal resolution in the available data. (S)he could then focus on the highly-ranked genes in her/his expression signature in combination with the specific functional contexts identified for these genes ('ASD-associated functional modules' as presented in Figure 4) to generate hypotheses and design experiments to further characterize the genes expressed in the window in relation to autism.