Supplementary Materials for

Genome-wide *de novo* risk score implicates promoter variation in autism spectrum disorder

Joon-Yong An[1,*], Kevin Lin[2,*], Lingxue Zhu[2,*], Donna M. Werling[1,*], Shan Dong[1], Harrison Brand[3,4,5], Harold Z. Wang[3], Xuefang Zhao[3,4,5], Grace B. Schwartz[1], Ryan L. Collins[3,4,6], Benjamin B. Currall[3,4,5], Claudia Dastmalchi[1], Jeanselle Dea[1], Clif Duhn[1], Michael C. Gilson[1], Lambertus Klei[7], Lindsay Liang[1], Eirene Markenscoff-Papadimitriou[1], Sirisha Pochareddy[8], Nadav Ahituv[9,10], Joseph D. Buxbaum[11,12,13,14], Hilary Coon[15,16], Mark J. Daly[5,17,18], Young Shin Kim[1], Gabor T. Marth[19,20], Benjamin M. Neale[5,17,18], Aaron R. Quinlan[16,19,20], John L. Rubenstein[1], Nenad Sestan[8], Matthew W. State[1,10], A. Jeremy Willsey[1,21,22], Michael E. Talkowski[3,4,5,23,†], Bernie Devlin[7,†], Kathryn Roeder[2,24,†], and Stephan J. Sanders[1,10,†]

Correspondence to: talkowski@chgr.mgh.harvard.edu (M. E. T.), devlinbj@upmc.edu (B. D.), kathryn.roeder@gmail.com (K. R.), stephan.sanders@ucsf.edu (S. J. S.)

This PDF file includes:

Materials and Methods

Sample information

Whole-genome sequencing (WGS) data were generated for 7,608 samples from 1,902 quartet families, each composed of a mother, a father, a child affected by autism spectrum disorder (ASD), and an unaffected sibling control. A complete list of families is provided in Table S1; 1,872 of these families were from the Simons Simplex Collection (SSC) (*18*) with the remaining 30 families from the Korean ASD cohort (*38*). The SSC families were generated in the following phases and batches:

- 40 families from SSC WGS data Pilot; included in the 519 families in our prior publication (*15*);
  Accession ID: SFARI_SSC_WGS_p

- 479 families from SSC WGS data Phase 1; included in the 519 families in our prior publication (*15*);
  Accession ID: SFARI_SSC_WGS_1

- 588 families from SSC WGS data, Phase 2;
  Accession ID: SFARI_SSC_WGS_2

- 539 families from SSC WGS data, Phase 3 Batch 1;
  Accession ID: SFARI_SSC_WGS_2

- 226 families from SSC WGS data, Phase 3 Batch 2;
  Accession ID: SFARI_SSC_WGS_2

WGS data generation

Details of data generation for the first 519 families are described in our previous publication (*15*). For the 1,383 new families, whole blood-derived DNA from all four family members was transferred from the Rutgers University Cell and DNA Repository (SSC samples) and UCSF (Korean samples) to the New York Genome Center (NYGC). PCR-free library preparation was used for all samples. WGS data were generated on Illumina Hi-Seq X Ten at NYGC and processed by the CCDG compliant pipeline (`https://github.com/CCDG/Pipeline-Standardization`). CRAM files and joint genotyping VCF are available via SFARI Base (`https://www.sfari.org/resource/sfari-base/`). Sequencing metrics were generated using the Picard CollectWgsMetrics module (version 2.17.2; `http://broadinstitute.github.io/picard`), see Table S1.

De novo variant detection

We used our previously described pipeline (*15*) to detect *de novo* SNVs and indels in the new 1,383 quartet families. Specifically, we used hail (https://github.com/hail-is/hail) to find Mendelian violations within each pedigree and then ran TrioDeNovo (*39*) to identify a list of putative *de novo* mutations. To refine this list to high quality *de novo* variants, we applied the sequential ROC method outlined previously (*15*). The following data-derived quality filters were applied for SNVs: passed VQSR, GQ$\geq$99, QUAL$\geq$200, SOR$\leq$2.5, DP$\geq$10, ReadPosRankSum$\geq$-1.4, GQM$\geq$50, QD$\geq$3, AB$\geq$0.22, parental GQ$\geq$30, parental AB$\geq$0.95, and MQ$\geq$60. The following data-derived quality filters were applied for indels: passed VQSR, GQ$\geq$99, QUAL$\geq$200, SOR$\leq$3, DP$\geq$10, ReadPosRankSum$\geq$-1.7, GQM$\geq$40, QD$\geq$4, AB$\geq$0.20, parental GQ$\geq$30, parental DP$\geq$16, DP$\leq$50 and MQ$\geq$50.

Since the variant calls from the phase 3 batch 1, Korean ASD families, and previous 519 families are based on hg19, 108,786 *de novo* mutations were lifted over to hg38 coordinates

using the CrossMap software (*40*) with 99.9% of mutations being successfully converted. The rate of *de novo* variants from the current study is comparable to previous reports (Fig. S1). The complete list of 255,106 high quality *de novo* mutations is given in Table S2.

Validation of *de novo* calls

Additional *de novo* calls were validated using PCR amplification and Sanger sequencing. We selected 56 conserved *de novo* mutations observed in the same promoter in either two cases (34 mutations at 17 loci) or two controls (22 mutations at 11 loci). Primers were designed using an in-house script that is based on Primer3 (*41*) and in-silico PCR. DNA was extracted, amplified and sequenced in all family members (father, mother, proband, and the unaffected sibling) in both forward and reverse directions. Of these 56 mutations, 48 yielded a PCR product in all four family members and all 48 (100%) were validated as *de novo* mutations.

Category-wide association study (CWAS)

Variant annotation

For gene-defined annotations, we used the Ensembl Variant Effect Predictor (VEP; version 90.4a44397) to assess the most severe consequence for each mutation on GENCODE complete version 27. Variant annotations with multiple predicted effects were assigned in the default order for VEP: coding, UTR, intron, promoter, and intergenic. Promoters were defined as 2kb upstream of the transcription start site (TSS). Additional VEP plugins were used to predict a functional effect of missense variants (PolyPhen2 v2.2.2) and obtain a population allele frequency from the gnomAD database (version 2.0.1.).

Creating annotation categories

Seventy annotation terms in five groups, applied separately and in all possible combinations between groups, were used to define 55,143 non-redundant annotation categories (Table S3 and

Fig. S5).

1. Variant types: We considered *de novo* SNVs versus indels obtained from the high quality call set.

2. Gene lists: We selected 13 gene lists as previously described (Table S8). We selected these gene lists for prior association with ASD, including ASD risk genes (*1*), expression in human brain (*42*), CHD8 binding targets (*8*,*9*), developmental disorders (DDD; Deciphering Developmental Disorders study) genes (*2*, *43*), FMRP binding targets (*30*), postsynaptic density genes (*44*), and genes co-expressed with ASD risk genes during early neurodevelopment (*6*). We also chose five gene lists based on GENCODE biotype to help distinguish coding and noncoding categories: protein coding genes, antisense, lncRNA, pseudogenes, and processed transcripts. Other noncoding transcripts (e.g. miRNAs) had too few mutations to be considered in this cohort.

3. Conservation scores: We used the PhastCons scores from a 46-way vertebrate comparison (score ≥ 0.2), and PhyloP scores from a 46-way vertebrate comparison (score ≥ 2) downloaded from the UCSC Genome Browser, followed by liftover to the hg38 genome coordinates using the CrossMap software (*40*) with 99.74% of loci being successfully converted.

4. GENCODE prediction: Gene definition was based on the VEP result with GENCODEv27, as described above.

5. Functional annotations: We applied the same functional annotations from our previous study (*15*), with the addition of H3K27ac ChIP-seq peaks from the dorsolateral prefrontal cortex and cerebral cortex from the PsychENCODE consortium (*36*). Since these annotations are based on the hg19 genome build, we converted coordinates to hg38 using the

CrossMap software (*40*) with 99.97% of loci being successfully converted.

Correcting *de novo* mutation count for paternal age

As previously described (*15*), we observed that the rate of *de novo* mutations was strongly correlated with paternal age, in contrast with small deviations associated with measures of sequencing quality. Overall, we observed 67.06 autosomal *de novo* mutations per child. To correct for paternal age, we fit a linear regression model to the per sample *de novo* mutation count against paternal age. The intercept and residuals from this model were used to adjust the per sample *de novo* mutation count and 67.06 was added to this number to keep the total number of mutations constant. The ratio between the original and corrected counts was calculated for each sample and this ratio was used to correct mutations counts in subsets of mutations for downstream analyses.

An alternative way to account for paternal age would be to model age as a confounder when assessing each category, with the effect size calculated in each category. To illustrate the problem with this approach, we compute the estimated age effect for each category independently, and then plot the estimated coefficient as a function of category size; i.e., number of mutations per category (Fig. S2). Notice that for large categories the estimated effects are significant, and quite stable. For small categories the estimated effects are not significant and vary dramatically due to sampling error, because there is very little data from which to estimate the effects. Indeed, the estimates are so unreliable that, for many categories, the estimated parameters suggest that mutation rates decrease for older age fathers. For this reason, we elected to use the per sample ratio corrected described above.

Per-category association test

For the 28,844 non-redundant annotation categories with at least one mutation in this sample (Table S3), we tested for ASD association by comparing the adjusted number of case and

control *de novo* mutations within each category, using a two-sided binomial test. We also used a permutation approach to compare case and control variant counts to a null distribution. For the permutation test, we calculated the relative risk per category as the ratio of the number of case variants to control variants (adjusted for paternal age, above). We then performed 10,000 within-sibship, case-control label-swapping permutations and calculated p-values as the proportion of these 10,000 permutations in which the label-swapped data returned a relative risk as or more extreme than the observed data. For categories with empirical $p<0.01$, we ran an additional 90,000 permutations to more precisely define the p-value. The p-values calculated by the binomial exact test were highly correlated with those estimated by permutation (R^2=0.82). The results of these burden tests are shown in Table S3.

Cross-category burden

To test whether a higher than expected proportion of related categories (e.g. all noncoding categories) showed an excess of mutations in cases, we counted the number of annotation categories that reached a nominally significant p-value ($\leq$0.05) from a two-sided binomial test comparing the adjusted number of *de novo* mutations in cases and controls in the observed data. We then calculated cross-category burden p-values as the proportion of the 10,000 case-control permutations, described above, in which the permuted data yielded as many or a greater number of nominally significant (binomial $p\leq0.05$) categories, with the same direction of effect, than the observed data.

Simulated p-values

Several downstream analyses rely on an assessment of the expected overlap between the annotation categories. To make this assessment, we generated random mutations to obtain the correlation structure of annotation categories underlying observed *de novo* mutations. For weighted random selection of nucleotide, we performed simulation using the frequency of all

12 mutational combinations (e.g. C to T) as previously described (*15*). To accurately reflect the distribution of *de novo* mutations across the genome, we estimated the effective length of autosomal chromosomes by excluding gaps and low complexity regions. We generated 255,106 random mutations for each simulation with the same proportion of SNVs and indels as observed variants. For each simulated dataset, we annotated the variants in the same manner as the observed data and identified the number of mutations in cases and controls for all 55,143 non-redundant annotation categories. A one-sided binomial exact test was used to estimate a p-value from the number of variants in cases and controls. We repeated this process 10,000 times to generate 10,000 simulated p-values for each category.

Number of effective tests

Many of the categories assessed for ASD association share substantial numbers of *de novo* mutations. To estimate the effective number of tests to correct for in the CWAS analysis, we used the methods described previously (*15*). Briefly, we considered the 28,844 non-redundant annotation categories with at least one mutation. Eigen decomposition, based on the simulated p-values (section Simulated p-values), was then used to identify the number of tests that explain 99% of the variation. We used the value of 6,711, which is in line with our prior estimates for a cohort of this size (*15*).

De novo risk score

Rare annotation categories

In coding regions, known ASD association is the result of mutations at a small number of critical loci (*28*), such as protein truncating variants (PTVs) in ≈5% of genes. Assuming a similar distribution of ASD risk for *de novo* mutations in the noncoding genome, we limited our analysis to annotation categories with few variants. Our aim was to construct a *de novo* risk score to predict ASD status, similar in concept to a polygenic risk score for common variation.

To select a threshold to define an annotation category with few variants, we reasoned as follows.

Exonic PTVs are not uncommon in the human population, as reflected in the Exome Aggregation Consortium database (5.3% of all exonic variation), and the frequency of PTVs cannot distinguish ASD cases from controls for a sample of size similar to that analyzed here. Even when PTVs are limited to those seen in only one individual and absent in ExAC, this subset remains insufficient to generate significant differences in rates between cases and controls. For reference, ExAC singleton PTVs occurred at a rate of $\approx 10^{-4}$. By limiting the PTVs to those found in highly constrained genes, which are impoverished for PTVs compared to expectation ($< 17\%$ of coding genes), and thus making them even rarer, case-control differences do emerge. We reason that the same principles should apply to *de novo* noncoding variation. Specifically, that a category should not be a notable source of risk unless the *de novo* noncoding variants within it are rare in controls. Thus, to develop a *de novo* risk score, we *a priori* exclude *de novo* events falling in an annotation category if their total frequency, adjusted for paternal age, taken over an annotation category, is ≥ 3 in controls. Given that there are, on average, ≈ 67 mutations per individual and 1,902 controls, 3 mutations per annotation category corresponds to a rate of 2.35×10^{-5}. To evaluate the robustness of this choice, we also computed the results for 1, 5 and 10 mutations in controls per annotation category (Table S5). For completeness, we also assessed the results if we selected a threshold of 3 mutations in cases per annotation category.

One issue that arises in the Lasso analysis is how to adjust for paternal age. Again, we can either pre-adjust mutation counts or estimate the age effect as a confounder for each analysis. With the latter approach, the same issues described for CWAS analysis arise (Fig. S2). The problem is particularly acute for the risk core analysis because the analysis is based on categories with few mutations. For this reason we preadjusted mutation counts to adjust for the paternal age effect.

Predictive R^2 of Lasso model

We built a Lasso prediction model for the rare annotation categories using the 519 previously described families (*15*). In particular, the regularization parameter in Lasso was selected by 5-fold cross validation, with minimized squared error loss. Then the fitted prediction model was applied to the remaining 1,383 new families, and the predicted phenotypes $\hat{y}$ were obtained. Finally, the predictive power of the Lasso model was evaluated by $R^2 = 1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y})^2$, where i is the category, $\hat{y}_i$ is the estimated phenotype for category i, $\bar{y}$ is the mean phenotype, and y_i is the observed phenotype for category i, coded as $y_i = -1$ for a control and $y_i = 1$ for a case. To account for the randomness in cross validation, we repeated this procedure for 10 times and reported the average predictive R^2. The set of categories that consistently had non-zero Lasso coefficients in all repetitions were treated as selected by the model. Using this procedure, we obtained the predictive R^2 and 238 categories with predictive power in all regions, as well as sub-regions including coding, noncoding, promoters, and noncoding regions excluding promoters (Table S5).

P-value of predictive R^2

To further evaluate the significance of the predictive R^2, permutation was used to estimate the p-value. Specifically, the null distribution was estimated by $\mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are the sample mean and variance of R^2, respectively, in 1,000 permutations. In each permutation, we randomly flipped the labels of case and control with probability 0.5 in each family, that is, the case/control labels were flipped in $\approx 50\%$ families. Next, we repeated the full procedure on the permuted data, including:

- Limiting the analysis to the set of rare categories with total *de novo* frequency $< m$ in "controls", where the "control" labels were the ones after random flipping and m is the mutation count threshold to exclude a category (e.g. m = 3).

- Fitting a Lasso model on the 519 previously described families with randomly flipped labels, where the Lasso regularization parameter was selected by 5-fold cross validation.
- Obtaining the predictive R^2 on the 1,383 new families, where the permuted labels were used.

Note that although we used the *de novo* frequency in all 1,902 families, including the 1,383 new ones, to select the set of rare categories, the potential selection bias was properly controlled when evaluating the significance because the same screening step was also applied in permutations. To further verify that the predictive power was not driven by such selection bias, we built a Lasso prediction model for the "rare" categories with total *de novo* frequency < 3 in *cases*, and the resulting predictive R^2 were no longer significant in all regions (Table S5).

Enrichment of annotation terms in the Lasso mode

To assess which of the 62 noncoding annotation terms (Fig. S5) were responsible for the noncoding signal detected by the Lasso model, we considered the frequency of these annotation terms in the 36,828 non-redundant noncoding annotation categories (Table S5-3). The number of categories that included each term was calculated and divided by 36,828 to get the probability of picking each annotation annotation term at random. The frequency of these 62 annotation terms in the 163 categories selected in the noncoding Lasso model was calculated and the difference been expected and observed frequencies was estimated using the binomial test. These p-values were corrected for 62 comparisons, after which promoters were highly significant (45 observed vs. 18 expected; 2.45-fold over expectation; $p = 6 \times 10^{-7}$, corrected), intergenic regions also showed modest enrichment (33 observed vs. 18 expected; 1.79-fold over expectation; $p = 0.04$, corrected).

Repeating this approach for the 17,061 non-redundant noncoding annotation categories with at least one *de novo* mutation (Table S5-3), promoter regions were the only region enriched in

the model (45 observed vs. 29 expected; 1.55-fold over expectation; $p = 0.04$, corrected).

DAWN analysis

Clustering

Having demonstrated that *de novo* variants in promoter regions were associated with ASD, we focused on which promoters and which mutations within promoter regions showed the strongest evidence of ASD association. We estimated the clustering of promoter categories by exploiting the correlation structure in 10,000 CWAS simulated datasets described in the "*Simulated p-values*" subsection. These simulated p-values were uniformly distributed under the null distribution, but for these finite sample sets, they were not ideal for evaluating the correlation structure. Thus we transformed p-values to z-scores, which were normally distributed under the null. Because of zero counts and p-values of one, we added one pseudo-count to both case and control counts and computed the perturbed p-values and z-scores. After obtaining the correlation matrix, C, of the perturbed z-scores, we utilized spectral clustering and performed K-means on the leading 50 eigenvectors of the normalized weight matrix, $D^{-1/2}|C|D^{-1/2}$, where D is a diagonal matrix with $D_{ii} = \sum_j |C_{ij}|$. We ignored the first eigenvector because it mainly accounts for the mean level. We normalized the remaining 49 selected eigenvectors such that each category had unit L_2 norm. To enhance the stability of clustering, we first ran K-means in a lower dimensional space, and used the medoids to initialize the final K-means clustering. Specifically, a 2-dimensional space was obtained using t-SNE (R package, (`https://github.com/jdonaldson/rtsne`) with perplexity parameter 30 and 500 iterations. To select the number of clusters, we visualized the total within-cluster sum of squares and picked the elbow point. Here we chose $K = 70$ (Table S7).

Processing categories

Having partitioned promoter categories into $K = 70$ clusters (Table S7). We converted one-sided permutation p-values, described in the "*Per-category association*" subsection, into z-scores, utilizing the estimated relative risk to determine the directionality of the association. We restricted analysis to categories with at least 20 mutations, including both cases and controls. After this pruning, we removed clusters containing only one category, leaving 47 clusters (Table S7).

Sparse PCA

Equipped with the z-scores for the categories, we computed the z-score for each cluster. This combined Z_j, the vector of z-scores for all categories in the jth cluster, with the estimated correlation matrix C described previously. Specifically, for each cluster j, we identified the correlation submatrix C_j that contains the rows and columns of all categories in said cluster. Because v_j, the leading right-singular vector of C_j, captured how the categories primarily behave under the null distribution, we aimed to compare how aligned v_j and Z_j were against the null distribution. To increase the power of our analysis, we used sparse PCA, which encourages sparsity on v_j but not u_j, the leading left-singular vectors. Specifically, this solves

$$\max_{u_j, v_j} u_j^T C_j v_j \quad \text{subject to } \|v_j\|_1 \leq \lambda, \|v_j\|_2^2 = 1, \|u_j\|_2^2 = 1,$$

where λ is a tuning parameter. We set $\lambda = 5.25$ because this had the desired effect of preserving all categories in smaller clusters and discarding at most half of the categories in larger clusters. We computed the magnitude of the z-score of the cluster by computing the magnitude of Z_j in the direction of v_j, normalized by the expected standard deviation of this quantity under the null distribution. The sign of the z-score of the cluster, s_j was 1 if the majority of values of Z_j when projected onto v_j are positive, and was -1 otherwise. Together,

the z-score of the cluster was formalized as

$$Z_{(j)} = s_j \cdot \sqrt{\frac{(v_j^T Z_j)^2}{v_j^T C_j v_j}}.$$

Using a similar strategy, we computed the adjusted relative risk of each cluster, working in the log scale. We converted the z-score of the cluster into a one-sided p-value by taking the right quantile of a Gaussian distribution evaluated at $Z_{(j)}$, after accounting for whether the adjusted relative risk of the cluster was less or greater than 1 (Table S7).

Correlation graph

We formed the graph G to encode which cluster of categories were related by inspecting submatrices in C. Specifically, we only placed an edge between cluster i and j if the average (mean) value in C between the categories in cluster i and those in cluster j was larger than $\tau = 0.12$. We chose this value since it produced a sparse graph that displayed good scale-free properties. Let A denote the adjacency matrix encoding G, with 0's along the diagonal.

Hidden Markov Random Field

Equipped with the z-scores for clusters, $Z_{(j)}$'s, and the adjacency matrix A, we propagated the signal among neighboring categories to increase the power. This was formalized by assuming the data followed a mixture of Gaussian distributions. Let $I_j = 1$ if the categories in cluster j contained more signal in the case subjects compared to the control subjects, and $I_j = 0$ otherwise. The following mixture model captured this,

$$Z_{(j)} \sim \mathbb{P}(I_j = 0)N(0, \sigma^2) + \mathbb{P}(I_j = 1)N(\mu, \sigma^2),$$

where μ and σ must be estimated, both larger than 0. To encode the graph structure, we modeled I_j as an Ising model,

$$\mathbb{P}(I_j = \eta_i) \propto \exp(b\eta_j + c\eta^T A\eta), \quad \text{for } \eta = (\eta_1, \ldots, \eta_K) \in \{0,1\}^K,$$

where b and c were scalar parameters to be estimated as well. We fitted all parameters above using a Gibbs sampler.

False discovery rate

We computed the posterior probability of $I_j = 1$ after fitting the above model. We selected any cluster of categories as significant using a Bayesian false discovery rate at level $\alpha = 0.01$.

Transcription factor binding site prediction

To predict transcription factor (TF) binding sites, we used a non-redundant collection of position frequency matrices (PFM) from the JASPAR database (*32*). FIMO, from the MEME suite (*45*), was used to predict TF binding sites in the sequence 2 kilobases upstream of the TSS of all transcripts from GENCODE version 27. The analysis was made with default parameters:

```
fimo --text <JASPAR meme file> fasta
```

Predictions with a p-value<0.0001 were selected for downstream analyses.

Ancestry prediction

Based on self-reported ancestry of samples, we knew *a priori* that samples were ancestrally heterogeneous. Thus, we undertook an analysis to determine if this heterogeneity was related to patterns of mutations in subjects. To do so, we first estimated genetic ancestry based on SNPs detected from the WGS data on 7,608 subjects. For genetic ancestry analysis we selected 69,641 SNPs that had the following SNP-based attributes:

- Called in all three batches of WGS data
- Genotype completion rate, over samples, was $> 99.9\%$
- Minor allele frequency > 0.05 using parental data only
- Exact Hardy Weinberg P-value > 0.005 in an ancestrally homogeneous subset of the parental subjects (78% of the sample and corresponding to those of European ancestry, which was genetically determined by analyses similar to those described below)
- FST < 0.0025 across the three batches of data for the ancestrally homogeneous subset of the parental subjects

Next, using the PCA function from PLINK (*46*) for analysis of genetic ancestry and the genotypes from 69,641 SNPs and 7,608 subjects, we decomposed the genetic covariance space into 20 principal components of ancestry. Fig. S3 illustrates some of this structure, which shows features consistent with numerous other analyses of samples as diverse as the ones analyzed here.

To test whether ancestry was a significant predictor of mutation rates, we fit the total count of mutations, per child (case or sibling control), to the child's estimated ancestry (20 PCs) and parental ages using the cv.glmnet function in R (*47*). This function implements the Lasso model selection procedure (*48*). Because our interest is in what predictors are significant, we use lamda-1se for model selection, which is the default option of the package. The analysis supports only two significant predictors to describe mutation counts, namely maternal and paternal age at birth (Fig. S4A). To explore this further, we also drop parental ages from the model and refit the counts. Consistent with the previous results, the Lasso provides no support for any ancestry PCs as significant predictors of mutation counts (Fig. S4B).

Fig. S1.

***De novo* mutation rates across studies.** The rate of *de novo* mutations is shown from previous studies of ASD cohorts (*13-15*, *22*, *33*, *49*) and population cohorts (*50-52*).

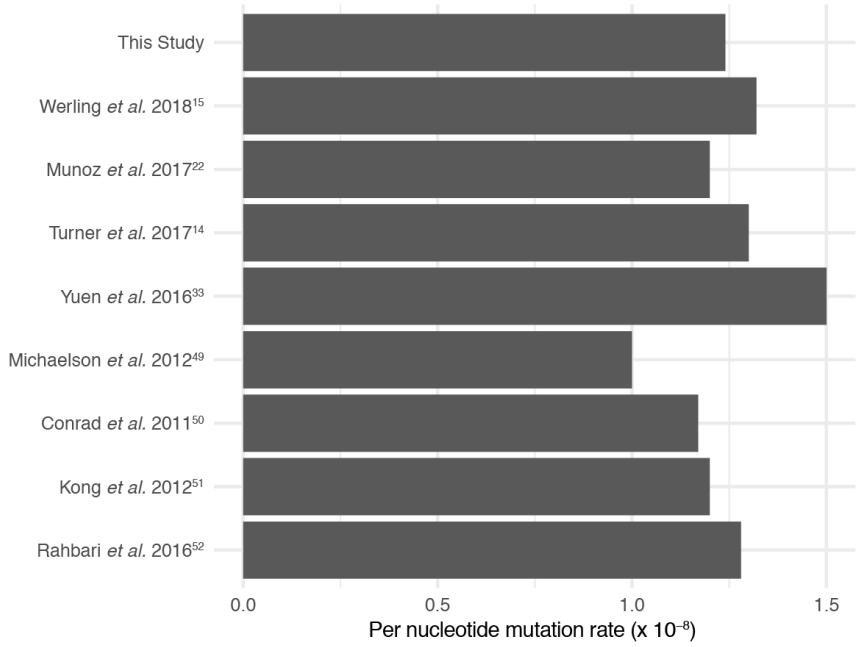


Fig. S2.

Paternal age effect estimated for each category in CWAS. Logistic regression predicts case/control status of each mutation based on paternal age. Category size, on the x-axis, is the total number of mutations observed in a category. For each category, the log(Beta) is shown on the y-axis, while the color represents whether paternal age was a significant predictor of case-control status. Estimates of Beta vary considerably for small categories.

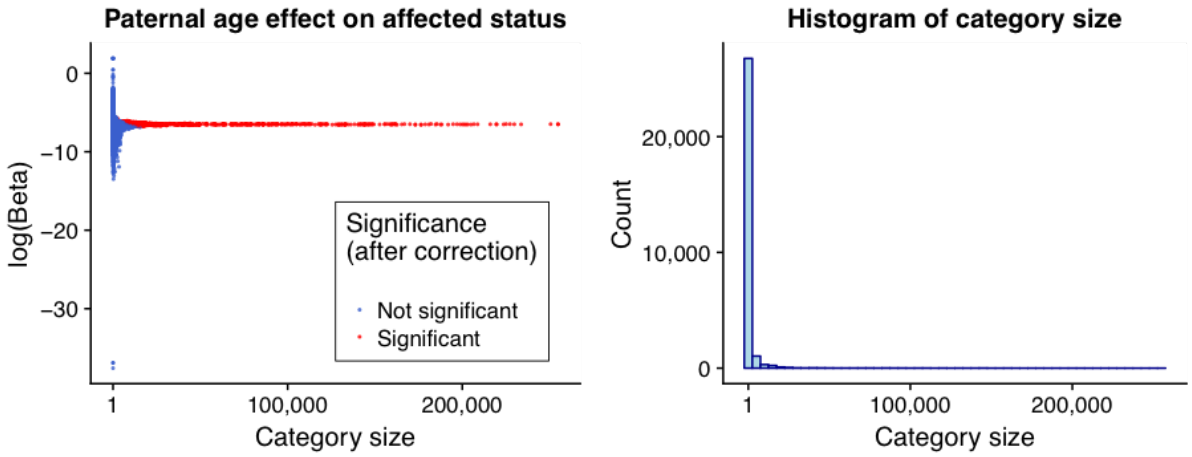


Fig. S3.

Principal Components (PC) of ancestry. The distribution of 7,608 subjects (black points) is shown across the first four principal components (PC1-4) based on 69,641 SNPs.

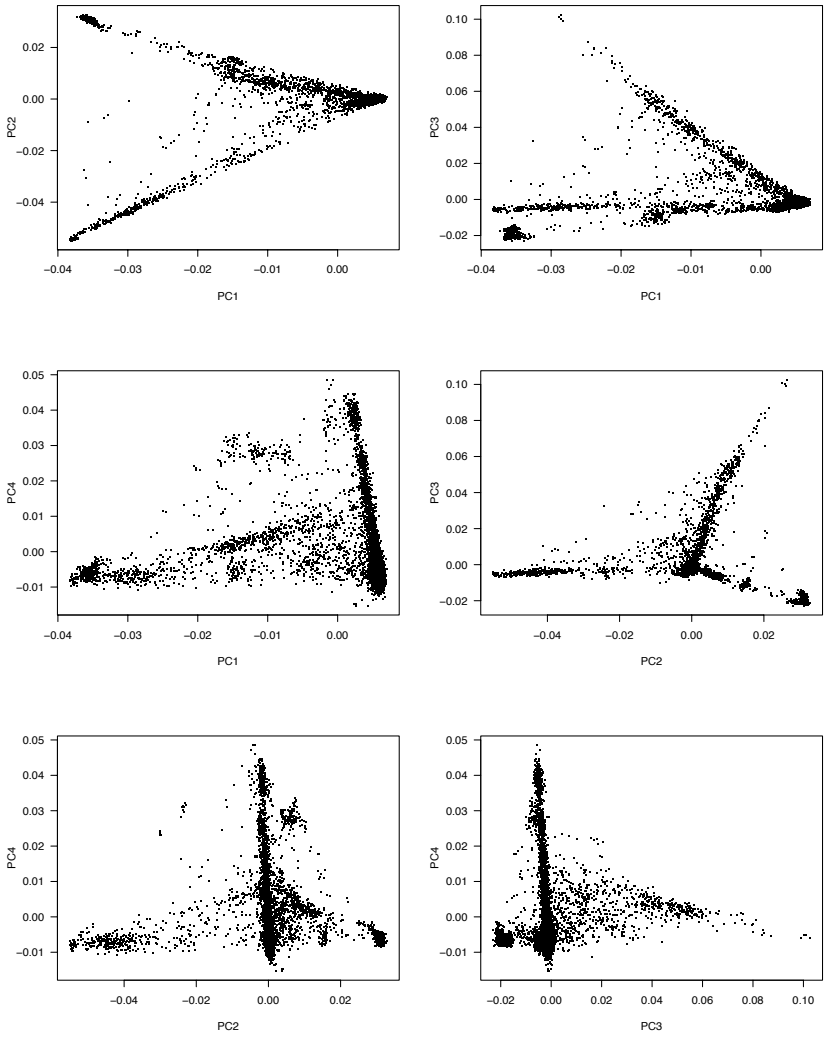


Fig. S4.

Lasso analysis for covariates. Results from the lasso model selection procedure fitting the total count of mutations, per child (cases and sibling controls), to the child's estimated ancestry (20 PCs) and parental ages (top panel) or 20PCs without parental age (bottom panel). Lambda is the penalty parameter of the Lasso, numbers at the top of the graph show the number of covariates in the model for each value of lambda. The rightmost reference line is "lamda-1se", which gives the most regularized model within one standard error of the minimum mean-square error; that is to say, the most parsimonious model that fits the data. For the top panel, the lasso includes parental ages as significant predictors; for the bottom panel, it selects no covariates. For reference, in the panel plotting principal component 2 (PC2) versus PC1, a vertical reference line at zero would separate the bulk of the sample (78%, right), who are of European Ancestry, from those of other ancestries. The concentration of subjects in the top left corner of the panel are of Asian ancestry.

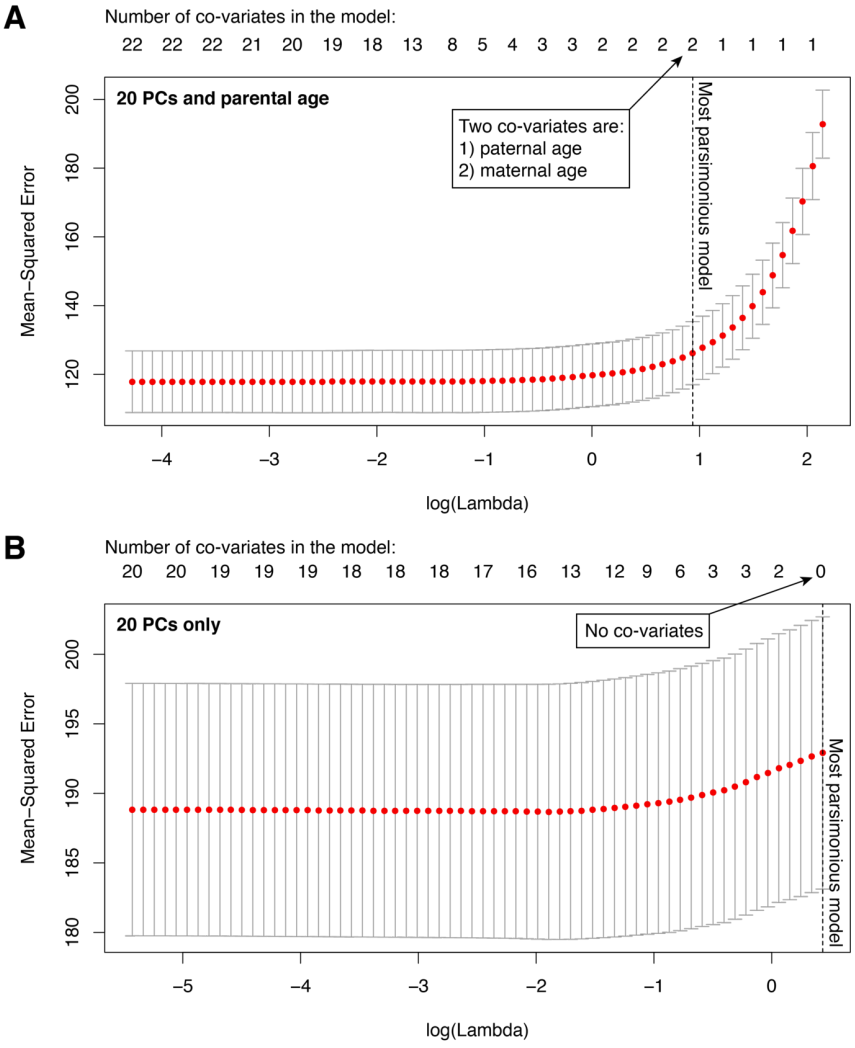

A
Number of co-variates in the model:
22 22 22 21 20 19 18 13 8 5 4 3 3 2 2 2 2 1 1 1 1
20 PCs and parental age
Two co-variates are:
1) paternal age
2) maternal age
Most parsimonious model
Mean−Squared Error
120 140 160 180 200
−4 −3 −2 −1 0 1 2
log(Lambda)
B
Number of co-variates in the model:
20 20 19 19 19 18 18 18 17 16 13 12 9 6 3 3 2 0
20 PCs only
No co-variates
Most parsimonious model
Mean−Squared Error
180 185 190 195 200
−5 −4 −3 −2 −1 0
log(Lambda)

Fig. S5.

Definition of annotation categories. We defined 70 annotation terms in five groups, including 1) GENCODE gene definitions (17 annotation terms); 2) Variant type (3 annotation terms); 3) Conservation across species (3 annotation terms); 4) Gene lists (14 annotation terms); and 5) Functional annotation (33 annotation terms). Picking one annotation term from each group resulted in 70,686 possible combinations, of which 55,143 were non-redundant (Table S3).

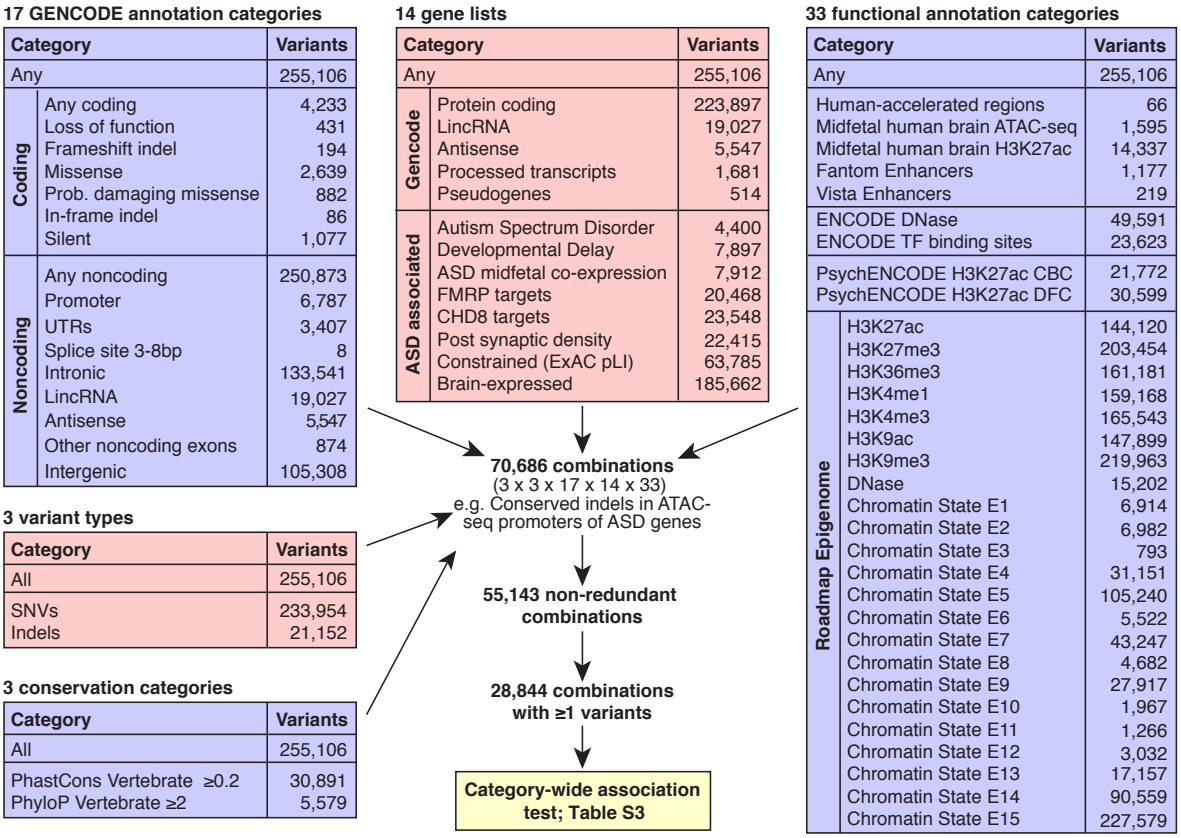


Fig. S6.

Cross-category burden of missense mutations. A) Missense categories that include ASD-associated genes as an annotation term are shown in a CWAS plot. Four categories reach statistical significance after correcting for multiple comparisons across all categories ($p=7.5 \times 10^{-6}$; dotted horizontal line). All categories are enriched in cases and there are more nominally significant categories than expected by chance in cases ($p<0.0001$, permutation test) but not controls ($p=1.00$, permutation test). **B)** Considering missense categories that do not include ASD-associated genes as an annotation term, no single category is significant after correcting for multiple comparisons, however there is an excess of nominally significant categories enriched in cases ($p=0.002$, permutation test), but not controls ($p=0.72$, permutation test).

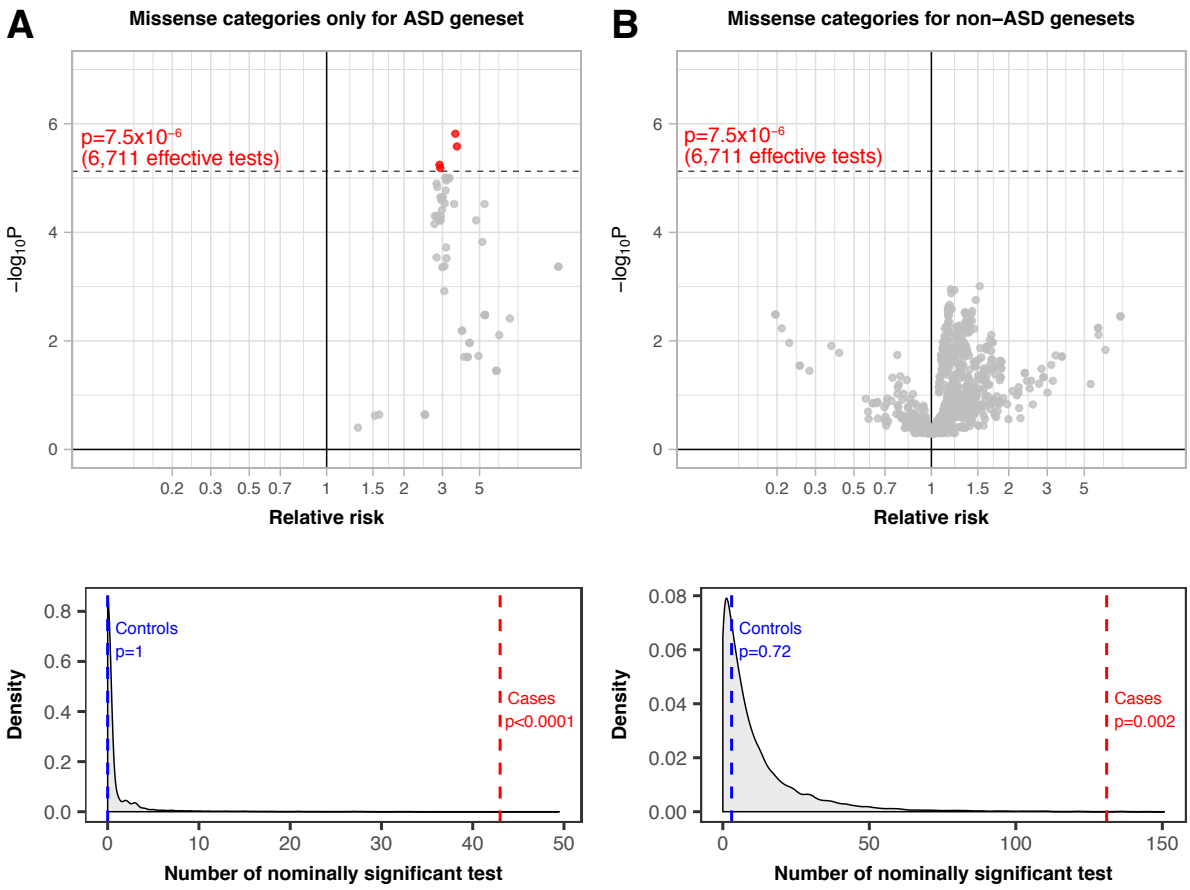

A
Missense categories only for ASD geneset
p=7.5x10⁻⁶
(6,711 effective tests)
−log10P
Relative risk
B
Missense categories for non−ASD genesets
p=7.5x10⁻⁶
(6,711 effective tests)
−log10P
Relative risk
Density
Controls
p=1
Cases
p<0.0001
Number of nominally significant test
Density
Controls
p=0.72
Cases
p=0.002
Number of nominally significant test

Fig. S7.

Burden of conserved mutations binned by cohort. **A)** The degree of enrichment of mutations in 1,902 cases vs. 1,902 controls (red line) is shown in relation to permuted expectation (grey distributions). The mean number of mutations per child is shown in parentheses on the left. Uncorrected P-values are shown in red, calculated by permutation test. **B)** The plot in 'A' is repeated showing the subset of 1,383 new families for this study. **C)** The plot in 'A' is repeated showing the subset of 519 families previously described in Werling *et al.* 2018 (*15*).

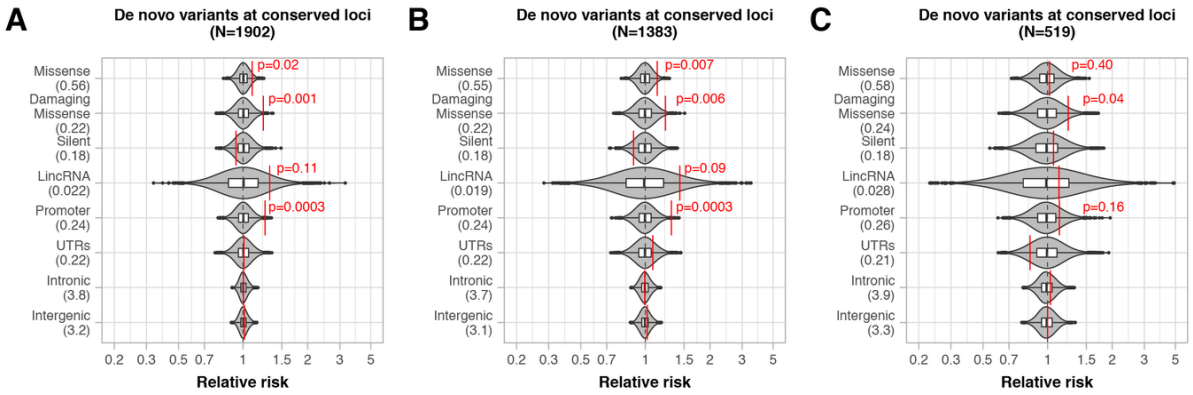


Fig. S8.

Burden of conserved mutations binned by the presence or absence of known ASD-associated disruptive coding mutations. A) The degree of enrichment of mutations in 1,902 cases vs. 1,902 controls (red line) is shown in relation to permuted expectation (grey distributions). The mean number of mutations per child is shown in parentheses on the left. Uncorrected P-values are shown in red, calculated by permutation test. **B)** Of the 1,902 families, there are 143 families in which the ASD case has a *de novo* PTV or *de novo* CNV at a locus previously associated with ASD (FDR ≤ 0.1)(*1*) and 1,759 families without such mutations (Table S1). The plot in 'A' is repeated showing the subset of 1,759 families without previously identified ASD-associated risk loci. A significant excess of conserved promoter mutations in cases is observed in families without prior ASD-associated risk variants (484 in cases vs. 378 in controls; cRR=1.26; p=0.005, permutation test). **C)** The plot in 'A' is repeated showing the subset of 143 families with previously identified ASD-associated risk loci. There is no significant excess of conserved promoter mutations (38 in cases vs. 31 in controls; cRR=1.20; p=0.24, permutation test), however there is no difference between the burden of conserved promoter mutations in these 143 families and the other 1,759 families (p=0.61; Fisher's exact test). Similarly, there is no significant difference between these two groups for conserved *de novo* missense mutations (p=0.20, Fisher's exact test).

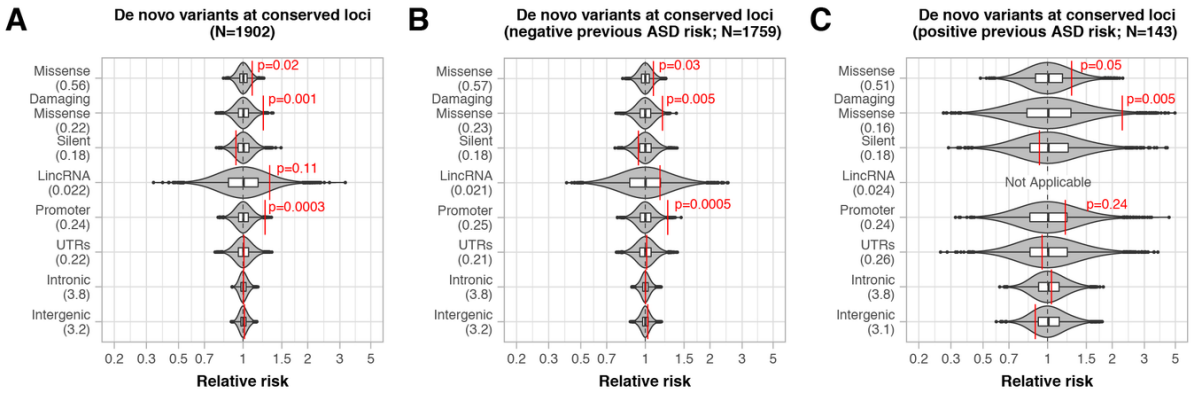
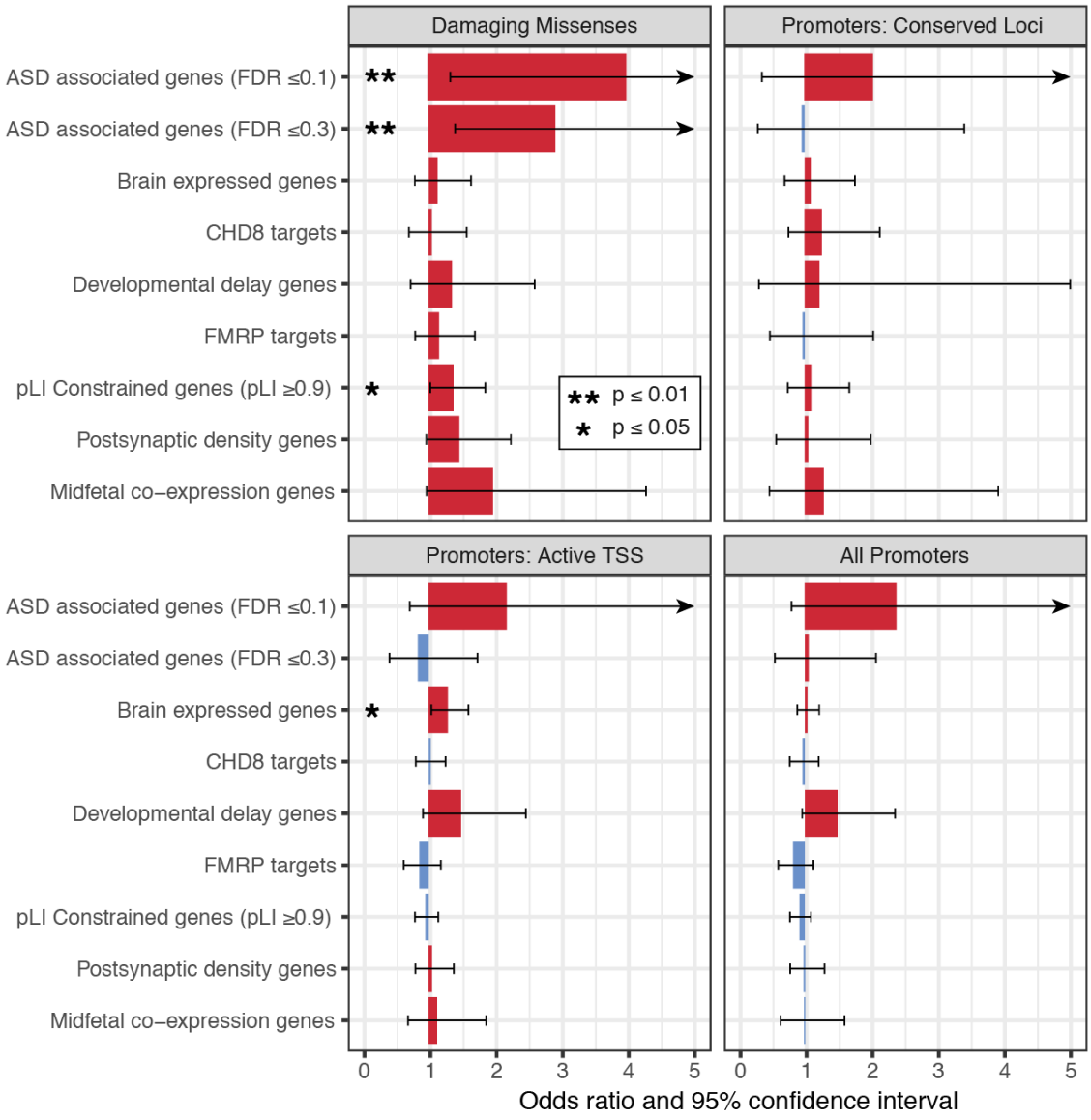

A
De novo variants at conserved loci
(N=1902)
Missense
(0.56)
Damaging
Missense
(0.22)
Silent
(0.18)
LincRNA
(0.022)
Promoter
(0.24)
UTRs
(0.22)
Intronic
(3.8)
Intergenic
(3.2)
p=0.02
p=0.001
p=0.11
p=0.0003
0.2 0.3 0.5 0.7 1 1.5 2 3 5
Relative risk
B
De novo variants at conserved loci
(negative previous ASD risk; N=1759)
Missense
(0.57)
Damaging
Missense
(0.23)
Silent
(0.18)
LincRNA
(0.021)
Promoter
(0.25)
UTRs
(0.21)
Intronic
(3.8)
Intergenic
(3.2)
p=0.03
p=0.005
p=0.0005
0.2 0.3 0.5 0.7 1 1.5 2 3 5
Relative risk
C
De novo variants at conserved loci
(positive previous ASD risk; N=143)
Missense
(0.51)
Damaging
Missense
(0.16)
Silent
(0.18)
LincRNA
(0.024)
Promoter
(0.24)
UTRs
(0.26)
Intronic
(3.8)
Intergenic
(3.1)
p=0.05
p=0.005
Not Applicable
p=0.24
0.2 0.3 0.5 0.7 1 1.5 2 3 5
Relative risk

Fig. S9.

Gene set enrichment for *de novo* missense and promoter mutations. The results of gene set enrichment tests are shown for four groups of mutations: probably damaging missense mutations (defined by PolyPhen2 (`https://genetics.bwh.harvard.edu/pph2/`)), conserved promoter mutations, active TSS promoter mutations, and all promoter mutations (Table S8). The gene sets considered are: ASD-associated genes with FDR ≤ 0.1 and FDR ≤ 0.3 (*1*), genes expressed in the human brain (*42*), CHD8 binding targets (*8*, *9*), genes implicated in developmental delay (*2*, *43*), FMRP binding targets (*30*), post-synaptic density genes (*44*), and genes co-expressed with ASD-associated genes during midfetal development in the prefrontal cortex (*6*). P-values calculated with Fisher's exact test without correction; * $P < 0.05$, ** $P < 0.01$.


Damaging Missenses
Promoters: Conserved Loci
Promoters: Active TSS
All Promoters
ASD associated genes (FDR ≤0.1)
ASD associated genes (FDR ≤0.3)
Brain expressed genes
CHD8 targets
Developmental delay genes
FMRP targets
pLI Constrained genes (pLI ≥0.9)
Postsynaptic density genes
Midfetal co−expression genes
** p ≤ 0.01
* p ≤ 0.05
0
1
2
3
4
5
Odds ratio and 95% confidence interval

Additional Data Table S1 (separate file)

Sample list and WGS quality metrics. This file provides details of the 7,608 samples assayed by WGS, their family relationships, sex, affected status, prior findings at ASD-associated loci (PTVs/CNVs), nonverbal IQ, and parental ages at birth. For each sample, the WGS quality metrics relating to sequencing coverage are provided.

Additional Data Table S2 (separate file)

List of *de novo* mutations. This large file (˜83Mb) contains all 255,106 high confidence *de novo* mutations identified in the cases and controls from 1,902 families, including genomic position (hg38), variant prediction, and annotation categories. It is available online at XXXX insert Dryad link here XXXX.

Additional Data Table S3 (separate file)

Category-wide association study results. This file contains three sheets: 1) An overview of the other two sheets; 2) 3-1 The burden results of category-wide association tests for 28,844 categories with $\geq$1 mutations; 3) 3-2 The 55,143 non-redundant annotation categories, see also Fig. S5.

Additional Data Table S4 (separate file)

Prior hypotheses of noncoding association. This file assesses 13 prior hypotheses of noncoding association against the data in our 1,902 families. Overlap between the discovery cohorts and our replication cohort are highlighted.

Additional Data Table S5 (separate file)

***De novo* risk score using a Lasso model.** This file contains six sheets: 1) An overview of

the other four sheets; 2) 5-1 The 238 categories selected in the complete Lasso model, with the Lasso coefficient, number of mutations, relative risk, and observed p-value. 3) 5-2 The 163 noncoding categories that are a subset of the full Lasso model, with the Lasso coefficient, number of mutations, relative risk, and observed p-value. 4) 5-3 Enrichment of 62 noncoding annotation terms in the categories selected in the noncoding Lasso model. 5) 5-4 The 45 promoter categories that are a subset of the full Lasso model, with the Lasso coefficient, number of mutations, relative risk, and observed p-value. 6) 5-5 The results of the Lasso model in 1,383 families, assessing the full, noncoding, promoter-only, and noncoding without promoters subsets with the predictive R^2 and p-value. For each submodel, we focus on the rare categories with $< m$ adjusted mutations, and we present results for our *a priori* analysis in which $m = 3$ in controls, for the robustness testing with $m = 1, 5, 10$ in controls, and $m = 3$ in cases. The p-values are estimated with 1,000 permutations.

Additional Data Table S6 (separate file)

List of promoter variants. This file contains six sheets: 1) An overview of the other five sheets; 2) 6-1 A list of the 6,787 promoter mutations, along with their inclusion in the Active TSS group, the Conserved Loci group, the Lasso model, the DAWN clusters, specific epigenetic or conservation annotations, and the gene. 3) 6-2 Distribution of the GENCODE-defined transcript biotype across the 931 Conserved Loci promoter mutations. 4) 6-3 The burden results of promoter variants after removing variants that are also annotated to non-synonymous coding variants in transcripts with lower GENCODE ranking. 5) 6-4 Gene Ontology analysis of promoter variants with the Conserved Loci Group. 6) 6-5 The enrichment of conserved promoter variants within JASPAR-defined transcription factor binding sites.

Additional Data Table S7 (separate file)

DAWN clustering analysis. This file contains three sheets: 1) An overview of the other two sheets; 2) 7-1 A list of 3,038 promoter categories with $\geq$20 mutations that were considered by the DAWN model, including number of mutations, inclusion in DAWN model (Final) with reason if excluded, and relative risk/p-value of the category and cluster. 3) 7-2 A list of 47 DAWN clusters, including relative risk, p-value, and a guide to the annotation categories included within the cluster. The nine clusters with FDR $\leq$ 0.1 are indicated by bold text (see Table 1 of main manuscript).

Additional Data Table S8 (separate file)

Gene set enrichment. This file contains three sheets: 1) An overview of the other two sheets; 2) 8-1 A list of 56,622 GENCODE genes and whether they are included in ASD-related gene lists (Fig. S5). 3) 8-2 The results of gene set enrichment rests for the gene lists on the first sheet for damaging missense, promoter, Active TSS promoter, and Conserved Loci promoter mutations. These enrichment test results are shown in Fig. S9.

Additional Data Table S9 (separate file)

Conserved promoter sites across the genome. This BED file lists all genomic regions (hg38) that are conserved promoters, as defined by our analysis (2,000bp upstream of the TSS using VEP and GENCODEv27 with PhyloP score $\geq$2 and/or PhastCons score $\geq$0.2, both conservation scores estimated from an analysis of 46 vertebrate species). This file offers a simple method to test the results observed in our analysis on future cohorts.