

## RESEARCH ARTICLE SUMMARY

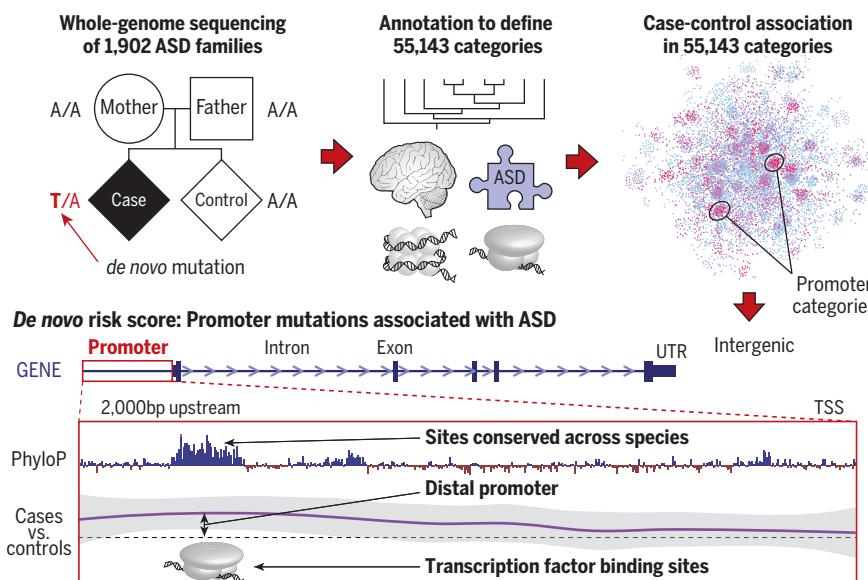
## PSYCHIATRIC GENOMICS

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

Joon-Yong An\*, Kevin Lin\*, Lingxue Zhu\*, Donna M. Werling\*, Shan Dong, Harrison Brand, Harold Z. Wang, Xuefang Zhao, Grace B. Schwartz, Ryan L. Collins, Benjamin B. Currall, Claudia Dastmalchi, Jeanselle Dea, Clif Duhn, Michael C. Gilson, Lambertus Klei, Lindsay Liang, Irene Markenscoff-Papadimitriou, Sirisha Pochareddy, Nadav Ahituv, Joseph D. Buxbaum, Hilary Coon, Mark J. Daly, Young Shin Kim, Gabor T. Marth, Benjamin M. Neale, Aaron R. Quinlan, John L. Rubenstein, Nenad Sestan, Matthew W. State, A. Jeremy Willsey, Michael E. Talkowski†, Bernie Devlin†, Kathryn Roeder†, Stephan J. Sanders†

**INTRODUCTION:** The DNA of protein-coding genes is transcribed into mRNA, which is translated into proteins. The “coding genome” describes the DNA that contains the information to make these proteins and represents ~1.5% of the human genome. Newly arising de novo mutations (variants observed in a child but not in either parent) in the coding genome contribute to numerous childhood developmental disorders, including autism spectrum disorder (ASD). Discovery of these effects is aided by the triplet code that enables the functional impact of many mutations to be readily deciphered.

In contrast, the “noncoding genome” covers the remaining ~98.5% and includes elements that regulate when, where, and to what degree protein-coding genes are transcribed. Understanding this noncoding sequence could provide insights into human disorders and refined control of emerging genetic therapies. Yet little is known about the role of mutations in noncoding regions, including whether they contribute to childhood developmental disorders, which noncoding elements are most vulnerable to disruption, and the manner in which information is encoded in the noncoding genome.



**Promoter regions in autism.** De novo mutations from 1902 quartet families are assigned to 55,143 annotation categories, which are each assessed for autism spectrum disorder (ASD) association by comparing mutation counts in cases and sibling controls. A de novo risk score demonstrated a noncoding contribution to ASD driven by promoter mutations, especially at sites conserved across species, in the distal promoter or targeted by transcription factors.

**RATIONALE:** Whole-genome sequencing (WGS) provides the opportunity to identify the majority of genetic variation in each individual. By performing WGS on 1902 quartet families including a child affected with ASD, one unaffected sibling control, and their parents, we identified ~67 de novo mutations across each child’s genome. To characterize the functional role of these mutations, we integrated multiple datasets relating to gene function, genes implicated in neurodevel-

opmental disorders, conservation across species, and epigenetic markers, thereby combinatorially defining 55,143 categories. The scope of the problem—testing for an excess of de novo mutations in cases relative to controls for each category—is challenging because there are more categories than families.

**RESULTS:** Comparing cases to controls, we observed an excess of de novo mutations in cases in individual categories in the coding genome but not in the noncoding genome. To overcome the challenge of detecting noncoding association, we used machine learning tools to develop a de novo risk score to look for an excess of de novo mutations across multiple categories. This score demonstrated a contribution to ASD risk from coding mutations and a weaker, but significant, contribution from noncoding mutations. This noncoding signal was driven by mutations in the promoter region, defined as the 2000 nucleotides upstream of the transcription start site (TSS) where mRNA synthesis starts. The strongest promoter signals were defined by conservation across species and transcription factor binding sites. Well-defined promoter elements (e.g., TATA-box) are usually observed within 80 nucleotides of the TSS; however, the strongest ASD association was observed distally, 750 to 2000 nucleotides upstream of the TSS.

**CONCLUSION:** We conclude that de novo mutations in the noncoding genome contribute to ASD. The clearest evidence of noncoding ASD association came from mutations at evolutionarily conserved nucleotides in the promoter region. The enrichment for transcription factor binding sites, primarily in the distal promoter, suggests that these mutations may disrupt gene transcription via their interaction with enhancer elements in the promoter region, rather than interfering with transcriptional initiation directly. ■

The list of author affiliations is available in the full article online.

\*These authors contributed equally to this work.

†Corresponding author. Email: talkowski@chgr.mgh.harvard.edu (M.E.T.); devlinbj@upmc.edu (B.D.); roeder@andrew.cmu.edu (K.R.); stephan.sanders@ucsf.edu (S.J.S.)

Cite this article as J.-Y. An et al., *Science* **362**, eaat6576 (2018). DOI: 10.1126/science.aat6576

## RESEARCH ARTICLE

## PSYCHIATRIC GENOMICS

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

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

Whole-genome sequencing (WGS) has facilitated the first genome-wide evaluations of the contribution of de novo noncoding mutations to complex disorders. Using WGS, we identified 255,106 de novo mutations among sample genomes from members of 1902 quartet families in which one child, but not a sibling or their parents, was affected by autism spectrum disorder (ASD). In contrast to coding mutations, no noncoding functional annotation category, analyzed in isolation, was significantly associated with ASD. Casting noncoding variation in the context of a de novo risk score across multiple annotation categories, however, did demonstrate association with mutations localized to promoter regions. We found that the strongest driver of this promoter signal emanates from evolutionarily conserved transcription factor binding sites distal to the transcription start site. These data suggest that de novo mutations in promoter regions, characterized by evolutionary and functional signatures, contribute to ASD.

**D**e novo mutations play an important role in human disorders that impair reproductive fitness, including autism spectrum disorder (ASD) (1), severe developmental delay (2), epileptic encephalopathy (3), and a spectrum of congenital anomalies (4, 5). Analysis of de novo mutations in the 1.5% of the genome that encodes proteins has identified numerous genes associated with ASD (1), and these findings have provided a foundation from which to interrogate ASD etiology (6–9). The contribution of de novo variation in the 98.5% of sequence that constitutes the noncoding genome remains largely unknown (10, 11).

Identifying noncoding variants that regulate gene function could provide important insights into when, where, and in which cell type ASD pathology occurs. Such knowledge could have broad implications for targeted therapeutics (10).

Targeted sequencing of highly evolutionarily conserved loci in 7930 families with a child affected by severe developmental delay identified a modest contribution from de novo mutations at loci that are active in the fetal brain (12). Whole-genome sequencing (WGS) represents the next critical step in such explorations, enabling the contribution of noncoding de novo mutations to be evaluated systematically across the

genome; however, the multiplicity of hypotheses that can be tested in an unbiased screen requires careful consideration of statistical interpretation. To date, WGS analyses of as many as 519 families with a child affected by ASD have yet to identify a significant noncoding contribution from de novo mutations, after appropriate correction for the multiple comparisons necessary in genome-wide analyses (13–16).

WGS analyses are complicated by the sheer scale of the noncoding genome and by limited methods to predict functional regions and disruptive variants. The category-wide association study (CWAS) framework applies multiple annotation methods to define thousands of annotation categories, each of which is tested for association with ASD. This CWAS approach is similar to that used in a genome-wide association study, with single-nucleotide polymorphisms (SNPs) substituted for annotation categories, and uses similar correction for multiple comparisons (15, 17). The CWAS-defined categories can also be used to build a de novo risk score, akin to a polygenic risk score, by selecting multiple annotation categories in a training cohort for assessment in a testing cohort (15). This model is generated once, so it does not incur a multiple testing penalty. In the present study, our results demonstrate an association between de novo noncoding mutations and ASD that is driven by mutations in conserved promoter regions.

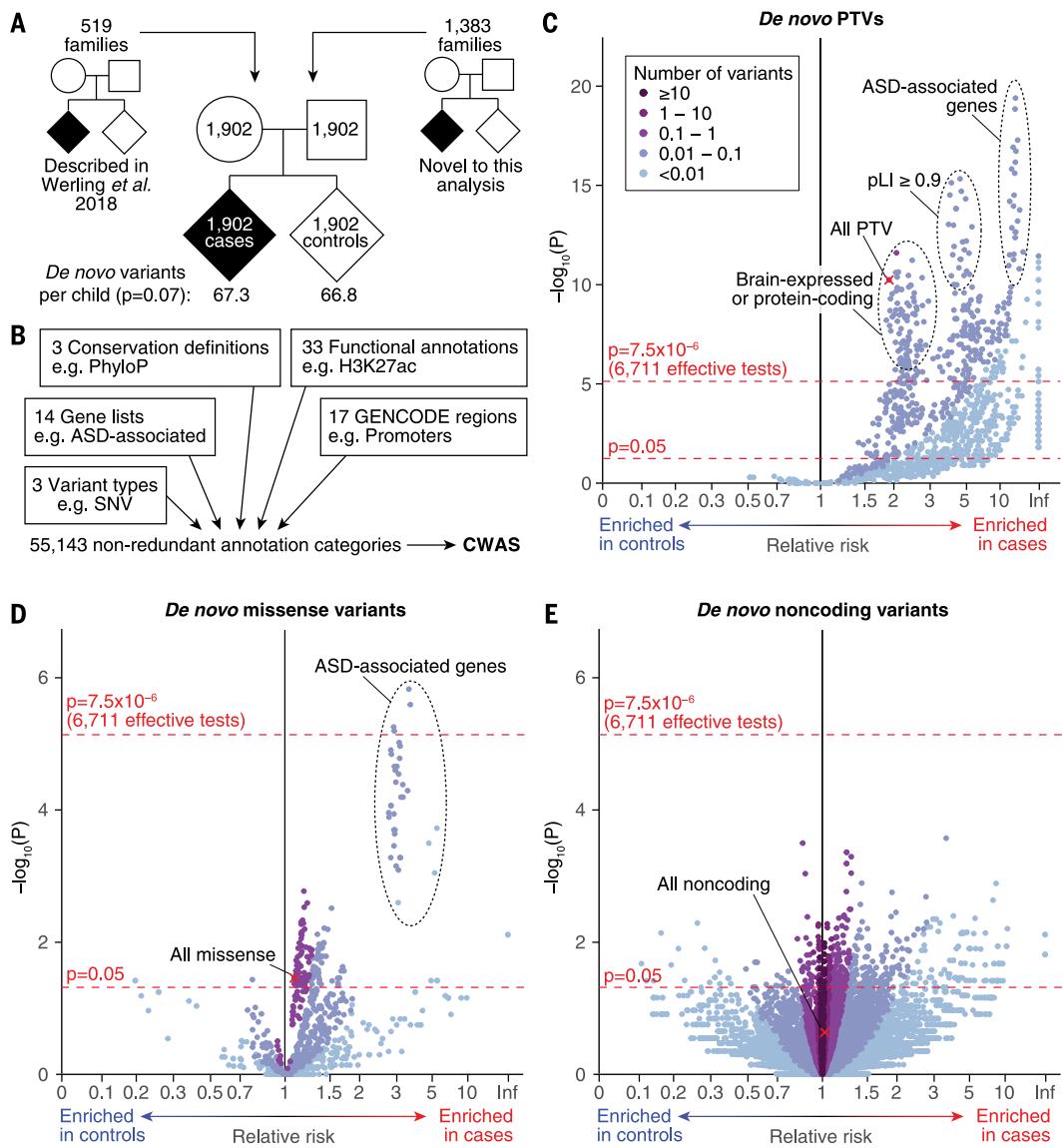
## Identification of de novo mutations in 1902 families

We analyzed the results of WGS in 7608 samples from 1902 quartet families from the Simons Simplex Collection (18), each composed of a mother and father, a child affected by ASD, and an unaffected sibling (table S1). This family-based design enables the detection of newly arising de novo mutations that are rare but can have drastic effects, and allows a direct comparison between ASD cases and their unaffected siblings as controls. By comparing each affected and unaffected child to their parents, we identified 255,106 de novo mutations in 1902 families (Fig. 1A and table S2), with 61.5 de novo single-nucleotide variants (SNVs) and 5.6 de novo insertions or deletions [indels; <50 base pairs (bp)] per child, using a high-quality variant filter defined in our previous study (15). These mutation rates are similar to those reported previously (fig. S1). Independent experimental validation confirmed

<sup>1</sup>Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California, San Francisco, CA, USA. <sup>2</sup>Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. <sup>3</sup>Center for Genomic Medicine and Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. <sup>4</sup>Department of Neurology, Harvard Medical School, Boston, MA, USA. <sup>5</sup>Program in Medical and Population Genetics and the Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA. <sup>6</sup>Program in Bioinformatics and Integrative Genomics, Division of Medical Sciences, Harvard Medical School, Boston, MA, USA. <sup>7</sup>Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA. <sup>8</sup>Department of Neuroscience and Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT 06510, USA. <sup>9</sup>Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA. <sup>10</sup>Institute for Human Genetics, University of California, San Francisco, CA, USA. <sup>11</sup>Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. <sup>12</sup>Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. <sup>13</sup>Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>14</sup>Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>15</sup>Department of Psychiatry, University of Utah School of Medicine, Salt Lake City, UT, USA. <sup>16</sup>Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, USA. <sup>17</sup>Analytical and Translational Genetics Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. <sup>18</sup>Department of Medicine, Harvard Medical School, Boston, MA, USA. <sup>19</sup>Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT, USA. <sup>20</sup>USTAR Center for Genetic Discovery, University of Utah School of Medicine, Salt Lake City, UT, USA. <sup>21</sup>Institute for Neurodegenerative Diseases, UCSF Weill Institute for Neurosciences, University of California, San Francisco, CA, USA. <sup>22</sup>Quantitative Biosciences Institute, University of California, San Francisco, CA, USA. <sup>23</sup>Departments of Pathology and Psychiatry, Massachusetts General Hospital, Boston, MA, USA. <sup>24</sup>Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

\*These authors contributed equally to this work.

†Corresponding author. Email: talkowski@chgr.mgh.harvard.edu (M.E.T.); devlinbj@upmc.edu (B.D.); roeder@andrew.cmu.edu (K.R.); stephan.sanders@ucsf.edu (S.J.S.)

**Fig. 1. Category-wide association study on 1902 ASD families.**

(A) De novo mutations were identified in 7608 samples from 1902 quartet families, each including an ASD case and an unaffected sibling control. The mean genome-wide mutation rate, corrected for paternal age, is shown for cases and controls.

(B) Each mutation was annotated against 70 annotation terms in five groups, combinations of which defined 55,143 annotation categories (table S3 and fig. S5). (C) A category-wide association study (CWAS) shows the degree to which de novo protein-truncating variants (PTVs) in each category (points) are enriched in cases (right x axis) or controls (left x axis) against the statistical evidence for this enrichment (y axis). Red lines show the threshold for nominal significance ( $P = 0.05$ ) and significance after correction for 6711 effective tests (19).

The red X shows the category of all PTVs without other annotations. (D and E) The equivalent CWAS is shown for de novo missense (D) and de novo noncoding (E) variants. Statistical tests: binomial exact test, two-tailed [(C) to (E)].

Downloaded from https://www.science.org at University of Pennsylvania on November 29, 2022

97.1% of SNVs (238/245) and 82.7% of indels (148/179) (19). No difference in noncoding de novo rate was observed between cases and controls after correcting for the established correlation between parental age and de novo frequency (20) [corrected relative risk (cRR) = 1.005;  $P = 0.15$  by permutation of case-control labels; table S3 and fig. S2]. Ancestry was not a significant predictor of de novo mutation rate; thus, it was not included in this correction (figs. S3 and S4).

### Only protein-coding categories show genome-wide enrichment in cases

In coding regions, ASD-associated mutations are found at a small number of critical loci—for example, protein-truncating variants (PTVs) in ~5% of genes (21). In the absence of an equivalent definition for critical noncoding loci, we annotated the mutations against gene definitions, ASD-associated gene lists, species conservation, types of mutation, and functional annotations (e.g., ChIP-seq, ATAC-seq, DNase-seq) to define

55,143 annotation categories (Fig. 1B, fig. S5, and table S3). Considering each category separately in a CWAS, 579 categories reached our correction threshold of  $7.5 \times 10^{-6}$ , generated by Eigen decomposition of 10,000 simulated datasets (15). All 579 categories were enriched in cases rather than controls; 575 of these included de novo PTV mutations (cRR = 1.92;  $P = 2.9 \times 10^{-11}$ , binomial; Fig. 1C), and the remaining four categories were subsets of missense mutations in genes previously associated with ASD (cRR = 2.90;  $P = 5.7 \times 10^{-6}$ ; Fig. 1D and fig. S6). No noncoding categories reached the correction threshold (Fig. 1E). We note that many of the ASD-associated genes were identified by de novo PTVs, and to a lesser extent de novo missense mutations, in these same cases (1). To focus on classes of variation with more subtle impacts on ASD risk, we excluded all annotation categories that included PTVs from further analysis.

Previous analyses have used WGS data to screen the genome, but those analyses were re-

stricted to “candidate” noncoding categories selected on the basis of assumptions about functional impact as opposed to unbiased genome-wide analyses, in cohorts ranging from 39 to 516 ASD families (13, 14, 22). Although these candidate categories were enriched at nominal significance in ASD cases in those initial discovery cohorts, no candidate categories reached nominal significance in this larger cohort, despite similar mutation rates (table S4). Similarly, we did not observe enrichment of mutations in ASD cases in the conserved noncoding elements described with targeted sequencing of 6239 families with severe developmental delay (12), although we note that our replication cohort is substantially smaller than the discovery cohort and of a different phenotype.

### Analysis across multiple noncoding categories highlights the role of promoters

No single noncoding annotation category passed our threshold of significance (Fig. 1E), so we

further explored the data by building a de novo risk score (15) to identify groups of categories in an unsupervised genome-wide analysis. To generate the score, we first restricted the analysis to annotation categories with a relatively small number of de novo mutations (19). This thresholding step is critical because the presence of numerous de novo mutations in an annotation category could represent false negatives in parents (i.e., apparent de novo mutations that were actually inherited variants), highly mutable regions, regions with limited impact on natural selection, or categories covering large swaths of the genome; none of these possibilities are likely to enrich for ASD risk at a small number of critical loci. Next, to select annotations likely to be important for risk from the remaining annotations, we generated a risk score using a Lasso regression from 519 families, described in (15), to identify annotation categories with rates of mutations that distinguish cases from controls. The resulting risk score was composed of 238 annotation categories, each with a coefficient reflecting the contribution of the category to the score (table S5). Applying the risk score to 1383 new families revealed it to be a significant

predictor of case status ( $R^2 = 1.67\%$ ,  $P = 5 \times 10^{-12}$ ; Fig. 2A). Of the 238 annotation categories, 75 were in coding regions ( $R^2 = 1.08\%$ ,  $P = 4 \times 10^{-9}$ ; table S5) and 163 were noncoding ( $R^2 = 0.54\%$ ,  $P = 0.02$ ; table S5); this finding demonstrates a noncoding contribution of de novo mutations to ASD risk.

To understand the nature of this noncoding contribution, we assessed the relative frequencies of the individual annotation terms from which the 163 noncoding categories are composed. The three annotation terms most frequently selected were PhastCons-defined (23) evolutionarily conserved regions (68 of 163 categories), PhyloP-defined (24) evolutionarily conserved nucleotides (49 of 163 categories), and promoter regions, defined as 2 kb upstream of the transcription start site (TSS) (45 of 163 categories). The inclusion of 45 promoter categories in the model is enriched by a factor of 2.45 over expectation ( $P = 6 \times 10^{-7}$  after correcting for 62 noncoding annotation terms; Fig. 2A and table S5). The risk score remained a significant predictor of case status with only these promoter categories included and accounted for the majority of the noncoding signal ( $R^2 = 0.50\%$ ,  $P =$

0.01; Fig. 2A and table S5). In contrast, the remaining 118 noncoding categories, without promoters, were not significant predictors of case status ( $R^2 = 0.22\%$ ,  $P = 0.25$ ; Fig. 2A). The 45 promoter categories selected in the risk score encompassed 150 independent mutations, 112 in cases and 38 in controls (table S6).

To examine whether this promoter signal was detectable beyond these 150 mutations, we considered the pattern of de novo mutation enrichment across all 1855 promoter-defined annotation categories with  $\geq 7$  mutations. Of these, 112 were enriched in cases at nominal significance, which is more than expected (cross-category burden  $P = 0.03$ ; Fig. 2, B and C), unlike the six categories enriched at nominal significance in controls (cross-category burden  $P = 0.94$ ; Fig. 2, B and C). Ten of the 112 case-enriched categories were also selected for inclusion in the de novo risk score; no control-enriched categories were selected.

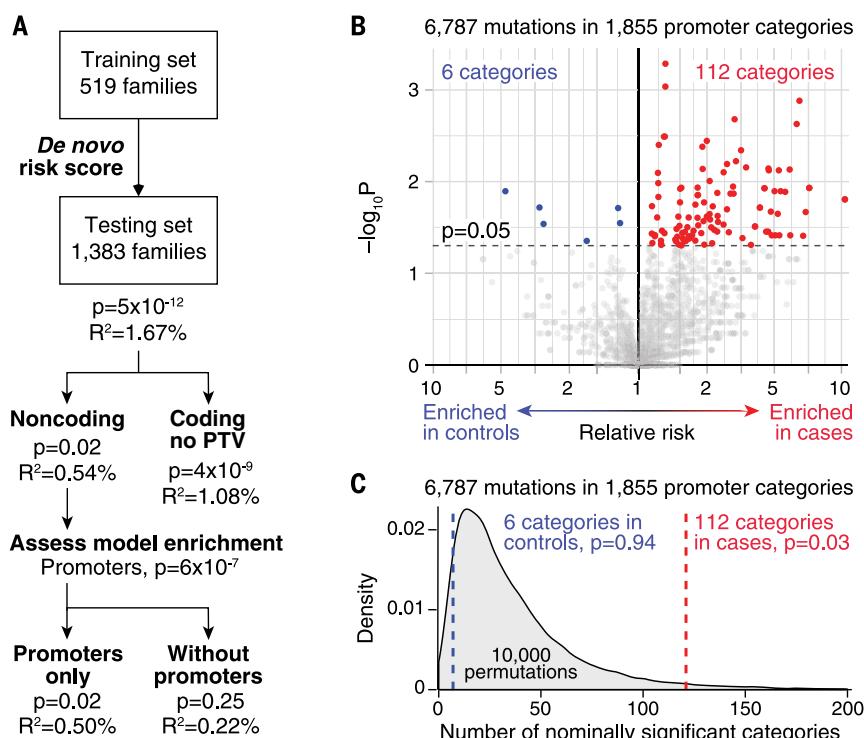
### Promoter association is driven by evolutionary conservation

To understand the types of variants and genes that account for this association between promoter mutations and ASD, we performed an exploratory analysis of the 6787 promoter region mutations and the 1310 promoter annotation categories with at least 20 mutations. Considering the correlation of  $P$  values across annotation categories, on the basis of 10,000 simulations (19), we identified 47 clusters, each composed of multiple highly correlated categories (Fig. 3A and table S7). Using the DAWN hidden Markov random field model (25) to refine the evidence for association based on the strength of association in neighboring clusters, nine of the 47 clusters were identified at a Bayesian false discovery rate of 0.01 (Fig. 3A and Table 1).

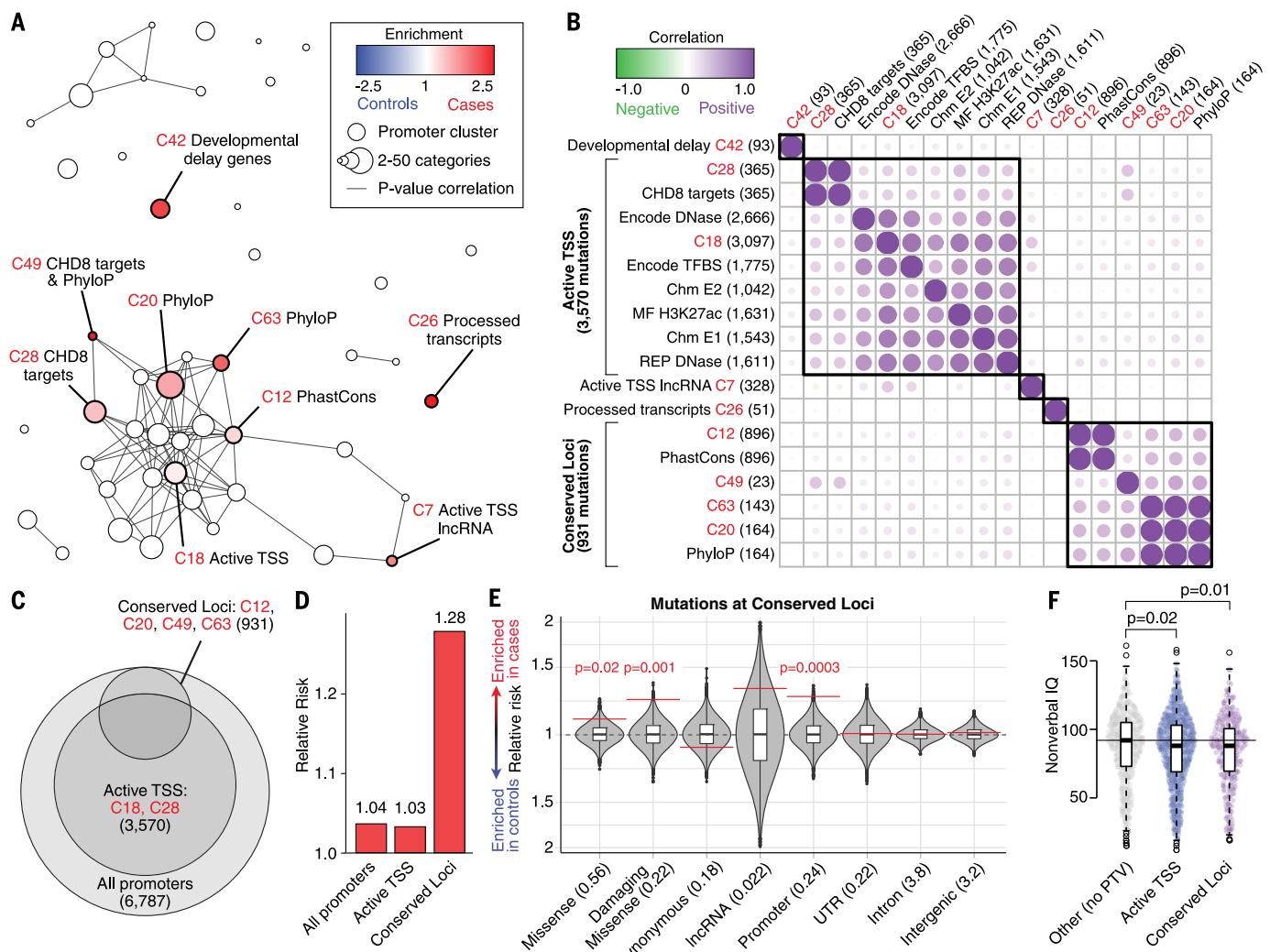
Assessment of the overlap of mutations between clusters and annotation terms identified two large groups of promoter mutations (Fig. 3, B and C): an “Active Transcription Start Site (TSS)” group ( $RR = 1.03$ ;  $P = 0.32$ , binomial test; Fig. 3D), distinguished by correlated epigenetic markers (C18 and C28; Fig. 3B), and a “Conserved Loci” group ( $RR = 1.28$ ;  $P = 0.0002$ , binomial test; Fig. 3D), distinguished by PhastCons and/or PhyloP scores (C12, C20, C49, C63; Fig. 3B). Of the 931 de novo mutations in the Conserved Loci group, 557 (60%) are also in the Active TSS group (Fig. 3C) and removing these conserved loci from the Active TSS group removes almost all of the signal ( $RR = 1.00$ ).

The three remaining small clusters show limited overlap with the Active TSS and Conserved Loci groups (Fig. 3B and Table 1): C7, defined by long noncoding RNAs (lncRNAs) at active TSSs ( $RR = 1.19$ ); C42, defined by developmental delay genes (2) ( $RR = 1.51$ ); and C26, defined by processed transcripts ( $RR = 2.00$ ).

When we consider all mutations in promoters as a single category, we see a nonsignificant trend toward weak enrichment in cases (3458 in cases versus 3329 in controls; cRR = 1.03;  $P = 0.16$ , permutation test). Because the cluster analysis highlighted the role of evolutionary conservation



**Fig. 2. Enrichment of conserved promoters in cases.** (A) After excluding categories with PTVs, which are known to have a strong contribution to ASD, a de novo risk score was generated using Lasso regression to distinguish cases and controls in the first 519 families and tested on 1383 new families. The same risk score was tested considering 163 noncoding categories only and, based on the enrichment of promoter categories in the risk score, for 45 promoter categories and 118 noncoding categories without promoters (table S5). (B) Considering 1855 promoter annotation categories with  $\geq 7$  mutations, 112 reached nominal significance, 112 of which had an excess of mutations in cases. (C) The observation of 112 nominally significant case-enriched categories (red line) and six control-enriched categories (blue line) in (B) is compared to permuted expectation (gray distribution). Statistical tests: Lasso regression with permutation testing (A); binomial, two-sided (B); permutation testing (C).



**Fig. 3. Mapping ASD association within promoter regions by annotation terms.** (A) DAWN uses *P*-value correlations between 1310 promoter categories with  $\geq 20$  mutations to define 47 clusters (nodes, with size representing the number of categories in the cluster). Evidence for ASD association is evaluated in the context of the local *P*-value correlation network (edges) to estimate false discovery rate (FDR). Enrichment is shown by color for the nine clusters with  $\text{FDR} \leq 0.01$  (Table 1). (B) The number of de novo mutations shared between these nine clusters and the annotation terms enriched in these clusters is shown as a correlation with hierarchical clustering. The black boxes show the first five divisions based on hierarchical clustering with two large groups: Active TSS and Conserved Loci. The numbers of de novo mutations in each group are shown in parentheses. (C) The size and relationship of the groups of promoter mutations identified in (A) and (B), based on de novo mutation counts. The

number of mutations in each group is shown in parentheses. (D) Estimates of relative risk based on the number of de novo mutations in cases and controls within each group. (E) Considering mutations at Conserved Loci, the degree of enrichment of mutations in cases versus controls (red line) is shown in relation to permuted expectation (gray distributions). The mean number of mutations per child is shown in parentheses. Nominally significant uncorrected *P* values are shown in red. (F) Distribution of nonverbal IQ in cases with mutations at Active TSS (blue) and Conserved Loci (purple) promoters versus cases with neither (gray). Cases with de novo PTVs were excluded from all groups. Statistical tests: DAWN (A); permutation testing (E); Wilcoxon signed rank, two-sided (F). Box plot in (E) and (F) shows the median (black line), interquartile range (white box), and a further 1.5 times the interquartile range (whiskers). DD, developmental delay; MF, midfetal; REP, Roadmap Epigenome; UTR, untranslated region.

(Fig. 3D), we assessed case-control burden for all 30,891 conserved mutations, split by GENCODE-defined (26) genic regions (Fig. 3E). We observed an excess of mutations in cases at conserved loci in promoters (522 versus 409; cRR = 1.26;  $P = 0.0003$ , permutation test), but not for mutations in other noncoding regions (Fig. 3E and fig. S7). In coding regions, de novo mutations that are not observed in the general population according to the Genome Aggregation Database (gnomAD) (27) are more likely to be associated with ASD

(28). Similarly, we observe stronger ASD association at promoter regions if mutations seen in gnomAD are excluded (470 versus 350; cRR = 1.34;  $P = 3 \times 10^{-5}$ , permutation test). Given the rarity and high effect sizes of protein-disrupting de novo mutations, we might expect a marginally higher rate of risk-mediating mutations in the 1759 ASD cases without previously identified ASD-associated mutations (1) relative to the 143 families with prior findings (table S1). However, no such difference was observed be-

tween these two groups in conserved promoters ( $P = 0.61$ , permutation test; fig. S8) or for conserved missense mutations ( $P = 0.20$ , permutation test; fig. S8).

#### Gene set enrichment and phenotype in the Conserved Loci group

The Conserved Loci group includes the promoters of 886 unique genes, of which 53% are protein-coding, 15% are processed pseudogenes, and 14% are lncRNAs (table S6) with similar

**Table 1.** Groups and clusters of categories within promoter regions.

Cluster	Description	Active TSS	Conserved loci	CHD8 binding targets	Total mutations (case/control)	Absolute RR	Binomial P value	DAWN RR	DAWN P value
C7	Active TSS lncRNAs	98%	18%	0%	328 (178/150)	1.19	0.14	1.66	0.03
C12	PhastCons	59%	100%	8%	896 (495/401)	1.23	0.002	1.22	0.003
C18	Active TSS	100%	16%	10%	3097 (1600/1497)	1.07	0.07	1.1	0.03
C20	PhyloP	82%	100%	14%	164 (100/64)	1.56	0.006	1.48	0.03
C26	Processed transcripts	57%	20%	0%	51 (34/17)	2	0.02	2.39	0.009
C28	CHD8 targets	100%	21%	100%	365 (183/182)	1.01	1	1.34	0.03
C42	Developmental delay genes	77%	11%	10%	93 (56/37)	1.51	0.06	2.06	0.02
C49	CHD8 targets and PhyloP	100%	100%	100%	23 (16/7)	2.29	0.09	2.43	0.01
C63	PhyloP	79%	100%	12%	143 (91/52)	1.75	0.001	1.87	0.03
All promoters		53%	14%	5%	6787 (3458/3329)	1.04	0.12	—	—
Active TSS group		100%	16%	10%	3570 (1815/1755)	1.03	0.32	—	—
Conserved Loci group		60%	100%	8%	931 (522/409)	1.28	0.0002	—	—

distributions in cases and controls except for processed transcripts (17 in cases, 0 in controls). In cases, genes with promoter mutations in the Conserved Loci group are enriched for “regulation of cell differentiation” (GO:0045595, FDR = 0.02), “transcription, DNA-templated” (GO:0006351, FDR = 0.04), and “regulation of transcription by RNA polymerase II” (GO:0006357, FDR = 0.04), whereas no biological processes are enriched in controls (table S8). Comparing cases to controls, there are nonsignificant trends toward enrichment in cases for ASD-associated genes (5 in cases, 2 in controls) and several ASD-related gene lists: brain-expressed (29), constrained (27), or CHD8 targets (8, 9, 30) (fig. S9 and table S8).

In coding regions, ASD-associated genes can be identified by the presence of multiple independent PTVs in different cases disrupting the same gene (1). In the WGS data, this approach did not yield specific promoters, because similar numbers of promoters had multiple Conserved Loci mutations in cases and controls (11 promoters in cases versus 7 in controls;  $P = 0.81$ , Fisher exact test). An equivalent analysis of damaging missense mutations, split into 2000-bp blocks to simulate promoters, suggests that we lack the power to detect specific promoters in a cohort of this size (22 in cases, 17 in controls;  $P = 1.00$ ).

Prior analyses of coding mutations have found large comorbid effects on nonverbal IQ, with ASD cases that carry ASD-associated mutations having a lower nonverbal IQ, on average (1). Excluding cases with de novo PTVs, we observed a 4-point reduction in median nonverbal IQ for cases with mutations in either the Active TSS [ $P = 0.02$ , Wilcoxon signed-rank test (WSRT)] and/or Conserved Loci ( $P = 0.01$ , WSRT) groups, relative to cases without such mutations (Fig. 3F). Furthermore, individuals with Conserved Loci promoter mutations show a trend toward a higher rate of mutations in female ASD cases ( $OR = 1.13$ ; 95% CI = 0.74 to 1.73;  $P = 0.31$ , Fisher exact test) and increased incidence of nonfebrile seizures ( $OR = 1.46$ ; 95% CI = 0.90 to

2.36;  $P = 0.07$ , Fisher exact test); both trends are consistent with results seen in coding mutations.

### The distal promoter shows the strongest evidence of association, especially at transcription factor binding sites

Because promoters are defined by their relationship to the TSS (31), we considered how ASD association varied by TSS distance, with the expectation that association would diminish with distance from the TSS. We first examined four bins: the core promoter ( $\leq 80$  bp), which we would expect to contain the TATA box, initiator element, and/or downstream promoter element; the proximal promoter (81 to 250 bp); and two divisions of distal promoters (251 to 1000 bp, 1001 to 2000 bp). In contrast to this expectation, mutations in the Conserved Loci group are most strongly enriched in the distal region (RR = 1.32;  $P = 0.005$ , binomial test; Fig. 4A). This distal association prompted us to consider only mutations at experimentally defined transcription factor binding sites (JASPAR CORE) (32), which enhanced the association (RR = 2.05;  $P = 0.0003$ , binomial test; Fig. 4B). Although a trend toward enrichment in cases is observed in the core promoter (Fig. 4, A and B), we do not see enrichment for motifs associated with RNA polymerase II (e.g., TATA; table S6). Looking at the enrichment in cases across the promoter in 200-bp sliding windows (Fig. 4, C and D), the strongest enrichment is observed between 750 and 2000 bp.

### Discussion

These analyses used WGS from 7608 individuals with an unbiased genome-wide association framework to demonstrate that de novo noncoding mutations alter risk for a complex neurodevelopmental disorder (Fig. 2). In a recent study (15), we highlighted the importance of genome-wide analyses with appropriate correction for multiple testing to identify noncoding regions robustly associated with ASD. Following this principle, no single noncoding annotation category was significant after conservative correction for multiple testing (Fig. 1E). Similarly,

we could not replicate candidate noncoding hypotheses described in previous analyses of ASD and developmental delay cohorts (table S4) (12–14, 22, 33). However, a “de novo risk score,” developed from a genome-wide Lasso analysis of multiple noncoding annotation categories, was a significant predictor of ASD risk (Fig. 2A). Such scores are routinely used in genomic analyses, including polygenic risk scores of common variants and, recently, a rare variant risk score for coding mutations in schizophrenia (34). Consistent with expectations, the magnitude of the contribution from noncoding mutations is smaller than that of the coding region, even having excluded de novo PTVs (Fig. 2A). Yet this early iteration of a de novo risk score could underestimate the true risk conferred by all noncoding mutations, as has been seen for polygenic risk score from common variants in successively larger cohorts (35).

Enrichment of annotation terms in the de novo risk score reveals that it is mutations in promoter regions (defined as 2000 bp upstream of the TSS) that underlie this noncoding association with ASD (Fig. 2A); the risk score continues to demonstrate ASD association when considering only promoter categories (45 of 163 categories; Fig. 2A). A consistent association signal can be observed across all 1855 promoter categories (Fig. 2B) and for 931 mutations at conserved loci (Fig. 3E). Notably, ASD cases with conserved promoter mutations have lower nonverbal IQ scores than ASD cases without these mutations (Fig. 3F)—an effect also observed in children with ASD-associated PTV mutations and missense mutations (1). Within promoters, the most robust association is observed for promoter mutations at Conserved Loci (Table 1), particularly at known transcription factor binding sites (Fig. 4B) (32). At Conserved Loci, the relative risk is similar to that observed for de novo damaging missense mutations (Fig. 3E). It is possible that the true relative risk is somewhat smaller, a phenomenon seen many times when the genome is searched for loci of relatively small effect and often called the winner’s curse. Surprisingly, the

strongest signal was not at the TSS and core promoter, but rather in the distal promoter, 750 to 2000 bp away from the TSS (Fig. 4). As expected for the distal promoter, the mutations in cases are frequently at experimentally defined transcription factor binding sites (Fig. 4D).

A key question is whether the de novo variation found in promoter regions is targeting the same set of genes implicated in ASD by de novo variants in protein-coding regions or a distinct set of genes not yet known to play a role in ASD. We favor the former possibility, although we cannot definitively exclude the latter, on the basis of (i) the enrichment for GO terms relating to transcriptional regulation and cell differentiation in the genes targeted by Conserved Loci

mutations, terms that are also enriched in ASD-associated genes (1); (ii) the trend toward enrichment for ASD-associated genes and several other gene sets previously implicated in ASD (fig. S9); and (iii) the detection of clusters defined by developmental delay genes and CHD8 binding targets (Fig. 3A and Table 1), both of which are enriched for ASD risk genes.

Our analysis establishes a specific hypothesis that can be tested for replication in future ASD cohorts and assessed in developmental and neuropsychiatric disorder cohorts: De novo mutations at conserved loci (46 vertebrate species PhastCons  $\geq 0.2$  and/or 46 vertebrate species PhyloP  $\geq 2$ ) in promoter regions (2000 bp upstream of the TSS based on GENCODEv27 annotation with VEP) are

associated with risk. To facilitate such analyses by others, we have generated a file of loci that meet these criteria (table S9). Despite these promising insights, we cannot yet identify which of the 522 conserved promoter mutations in cases truly confer risk, nor can we be confident which of the remaining 126,031 noncoding case mutations do not. Instead, our results demonstrate that elucidation of the contribution of de novo noncoding mutations to human disorders is feasible, and that the yields are likely to improve substantially with increases in cohort size (10, 15).

That conserved loci are one of the major factors underlying the promoter association could be interpreted to mean that nonhuman models can be used to assay noncoding function in humans, although parallel work in humans will be required to show that the specific regulatory effects are also conserved. Enrichment at transcription factor binding sites is also promising. If ASD association can be detected for specific transcription factors or loci, it raises the prospect of high-resolution neurobiological insights into spatiotemporal development, especially when, where, and in which cell type typical development is disrupted in ASD. Such insights will require detailed functional data on transcription factors and how they relate to mutations found in ASD.

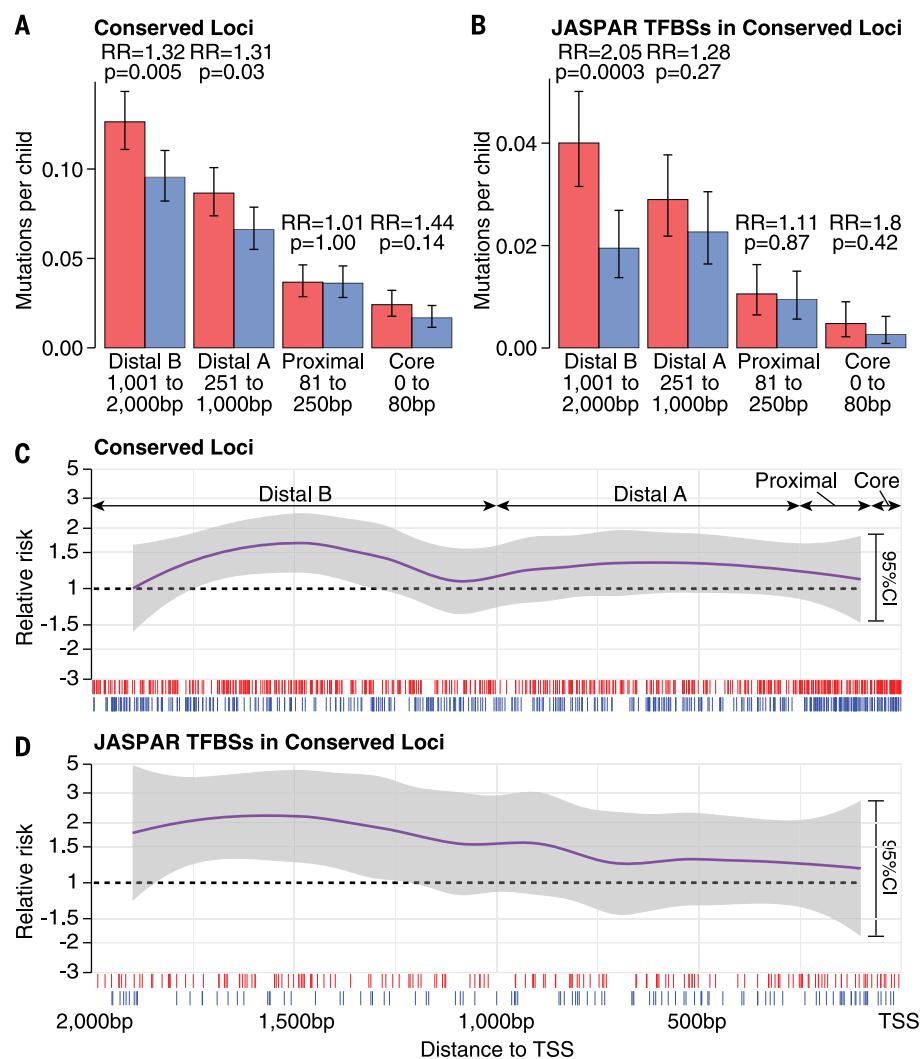
The association that we observe from these data represents the integration of work from multiple fields, including human cohort collections (2, 18), gene definitions (26), comparative genomics (23, 24), and functional genomics (32, 36). Methods and infrastructure are being developed to replicate and refine this association, identify specific loci, or extend beyond promoters. These include larger cohorts with consistently analyzed WGS data [e.g., the WGSPD consortium (10)], refined annotation of noncoding regions in the human brain [e.g., the PsychENCODE consortium (36)], WGS-tailored analytical methods (15, 25), and large-scale functional assays [e.g., massively parallel reporter assays (37)]. The evolving results from these fields provide a path to improving diagnosis and novel therapeutic strategies that could benefit a wide range of human disorders.

## Materials and methods

See (19) for additional details.

## Detection and annotation of de novo mutations

WGS data were generated by the New York Genome Center with a mean coverage of 35.5 in 1902 ASD quartet families. Previously described variant filtering criteria were applied (15) to identify 255,106 high-quality de novo mutations. These mutations were annotated using the Ensembl Variant Effect Predictor (VEP; version 90.4a44397) with GENCODE v27 gene definitions. Nucleotide sequence conservation across 46 vertebrate species (PhyloP, PhastCons), and regulatory regions (e.g., transcription factor binding sites, chromatin states) were annotated using VEP. In addition to 424 previously validated loci, 45 de novo mutations in promoter regions with two or more mutations in different samples



**Fig. 4. Relationship of conserved promoter mutations to the TSS.** (A) Frequency of Conserved Loci promoter mutations in cases and controls across the promoter region. (B) Frequency of Conserved Loci promoter mutations in cases and controls at JASPAR transcription factor binding sites (TFBSs) across the promoter region. (C) Enrichment of Conserved Loci promoter mutations in cases, shown as relative risk, in sliding windows of 200 bp across the promoter region. The purple line is the generalized additive model fit for relative risk and the 95% confidence interval is in gray. Ticks under the plot show individual mutations in cases (red) and controls (blue). (D) The plot in (C) is repeated for Conserved Loci promoter mutations at JASPAR TFBS. Statistical tests: binomial, two-sided [(A) and (B)]. Error bars show the 95% confidence interval (95% CI).

were validated as de novo by analyzing all four members of each family with PCR and Sanger sequencing.

### Category-wide association study (CWAS)

To assess multiple hypotheses, we implemented the CWAS method, described in (15). Considering 70 annotation terms from five groups in combination defined 55,143 nonredundant categories for downstream analysis. ASD association was tested for each category by comparing the burden of case and control mutations with a two-sided binomial test, having corrected the rate of de novo mutations for paternal age. To estimate the penalty of multiple comparisons, the number of effective tests was estimated using Eigen decomposition of *P* values in 10,000 simulated datasets. Each simulated dataset contained 255,106 random variants and maintained the GC bias and proportion of SNVs to indels observed in the original data.

### De novo risk score analysis

To build a de novo risk score, we excluded all categories that could contain de novo PTVs, then selected 8418 rare annotation categories with  $\leq 3$  mutations in controls. From the training dataset of 519 families described previously (15), we used a Lasso regression with five-fold cross-validation to estimate the regularization parameter, and then applied this fitted prediction model to the remaining 1383 new families to estimate the predictive power of the risk score. The significance of the prediction was calculated from 1000 permutations with case-control status swapped in 50% of families selected at random. The frequency of the 62 noncoding annotation terms was compared between the 36,828 nonredundant noncoding categories and the 163 noncoding categories in the de novo risk score. A binomial test was used to assess the enrichment of these terms, corrected for 62 comparisons.

### DAWN clustering analysis of promoter categories

The DAWN hidden Markov random field model (25) was used to assess the risk factors underlying ASD association of promoters. Clusters of individual promoter categories were defined by *K*-means ( $K = 70$ ) based on the *P*-value correlation network generated from 10,000 simulated datasets. Of these 70 clusters, 47 had at least 20 mutations and 2 categories and were considered further. Observed *P* values were transformed to z-scores and sparse PCA analysis was used to estimate the *P* value and relative risk per cluster. Using a hidden Markov random field model, these estimates were modified to yield a posterior probability based on enrichment in neighboring clusters in the simulated *P*-value correlation network.

### REFERENCES AND NOTES

- S. J. Sanders *et al.*, Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015). doi: [10.1016/j.neuron.2015.09.016](https://doi.org/10.1016/j.neuron.2015.09.016)
- Deciphering Developmental Disorders Study, Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017). doi: [10.1038/nature21062](https://doi.org/10.1038/nature21062)
- H. O. Heyne *et al.*, De novo variants in neurodevelopmental disorders with epilepsy. *Nat. Genet.* **50**, 1048–1053 (2018). doi: [10.1038/s41588-018-0143-7](https://doi.org/10.1038/s41588-018-0143-7)
- S. C. Jin *et al.*, Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017). doi: [10.1038/ng.3970](https://doi.org/10.1038/ng.3970)
- C. Redin *et al.*, The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.* **49**, 36–45 (2017). doi: [10.1038/ng.3720](https://doi.org/10.1038/ng.3720)
- A. J. Willsey *et al.*, Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997–1007 (2013). doi: [10.1016/j.cell.2013.10.020](https://doi.org/10.1016/j.cell.2013.10.020)
- R. Ben-Shalom *et al.*, Opposing effects on NaV1.2 function underlie differences between SCN2A variants observed in individuals with autism spectrum disorder or infantile seizures. *Biol. Psychiatry* **82**, 224–232 (2017). doi: [10.1016/j.biopsych.2017.01.009](https://doi.org/10.1016/j.biopsych.2017.01.009)
- A. Sugathan *et al.*, CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E4468–E4477 (2014). doi: [10.1073/pnas.1405266111](https://doi.org/10.1073/pnas.1405266111)
- J. Cotney *et al.*, The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat. Commun.* **6**, 6404 (2015). doi: [10.1038/ncomms7404](https://doi.org/10.1038/ncomms7404)
- S. J. Sanders *et al.*, Whole Genome Sequencing in Psychiatric Disorders: The WGSPD Consortium. *Nat. Neurosci.* **20**, 1661–1668 (2017). doi: [10.1038/s41593-017-0017-9](https://doi.org/10.1038/s41593-017-0017-9)
- F. Zhang, J. R. Lupski, Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015). doi: [10.1093/hmg/ddv259](https://doi.org/10.1093/hmg/ddv259)
- P. J. Short *et al.*, De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018). doi: [10.1038/nature25983](https://doi.org/10.1038/nature25983)
- T. N. N. Turner *et al.*, Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am. J. Hum. Genet.* **98**, 58–74 (2016). doi: [10.1016/j.ajhg.2015.11.023](https://doi.org/10.1016/j.ajhg.2015.11.023)
- T. N. Turner *et al.*, Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**, 710–722.e12 (2017). doi: [10.1016/j.cell.2017.08.047](https://doi.org/10.1016/j.cell.2017.08.047)
- D. M. Werling *et al.*, An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018). doi: [10.1038/s41588-018-0107-y](https://doi.org/10.1038/s41588-018-0107-y)
- W. M. Brandler *et al.*, Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018). doi: [10.1126/science.aan2261](https://doi.org/10.1126/science.aan2261)
- F. Dudbridge, A. Gusnanto, Estimation of significance thresholds for genome-wide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008). doi: [10.1002/gepi.20297](https://doi.org/10.1002/gepi.20297)
- G. D. Fischbach, C. Lord, The Simons Simplex Collection: A resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010). doi: [10.1016/j.neuron.2010.10.006](https://doi.org/10.1016/j.neuron.2010.10.006)
- See supplementary materials.
- H. Jónsson *et al.*, Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017). doi: [10.1038/nature24018](https://doi.org/10.1038/nature24018)
- S. De Rubeis *et al.*, Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014). doi: [10.1038/nature13772](https://doi.org/10.1038/nature13772)
- A. Munoz *et al.*, De novo indels within introns contribute to ASD incidence. *BioRxiv* **137471** [Preprint], 24 May 2017.
- A. Siepel *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005). doi: [10.1101/gr.3715005](https://doi.org/10.1101/gr.3715005)
- K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010). doi: [10.1101/gr.097857.109](https://doi.org/10.1101/gr.097857.109)
- L. Liu *et al.*, DAWN: A framework to identify autism genes and subnetworks using gene expression and genetics. *Mol. Autism* **5**, 22 (2014). doi: [10.1186/2040-2392-5-22](https://doi.org/10.1186/2040-2392-5-22)
- J. Harrow *et al.*, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012). doi: [10.1101/gr.135350.111](https://doi.org/10.1101/gr.135350.111)
- M. Lek *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016). doi: [10.1038/nature19057](https://doi.org/10.1038/nature19057)
- J. A. Kosmicki *et al.*, Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* **49**, 504–510 (2017). doi: [10.1038/ng.3789](https://doi.org/10.1038/ng.3789)
- H. J. Kang *et al.*, Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011). doi: [10.1038/nature10523](https://doi.org/10.1038/nature10523)
- J. C. Darnell *et al.*, FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011). doi: [10.1016/j.cell.2011.06.013](https://doi.org/10.1016/j.cell.2011.06.013)
- H. Kwak, N. J. Fuda, L. J. Core, J. T. Lis, Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339**, 950–953 (2013). doi: [10.1126/science.1229386](https://doi.org/10.1126/science.1229386)
- A. Khan *et al.*, JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018). doi: [10.1093/nar/gkx126](https://doi.org/10.1093/nar/gkx126)
- R. K. C. Yuen *et al.*, Genome-wide characteristics of de novo mutations in autism. *npj Genomic Med.* **1**, 16027 (2016). doi: [10.1038/npjgenmed.2016.27](https://doi.org/10.1038/npjgenmed.2016.27)
- S. M. Purcell *et al.*, A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014). doi: [10.1038/nature12975](https://doi.org/10.1038/nature12975)
- N. Chatterjee *et al.*, Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–405 (2013). doi: [10.1038/ng.2579](https://doi.org/10.1038/ng.2579)
- S. Akbarian *et al.*, The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015). doi: [10.1038/nn.4156](https://doi.org/10.1038/nn.4156)
- F. Inoue *et al.*, A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017). doi: [10.1101/gr.21092.116](https://doi.org/10.1101/gr.21092.116)

### ACKNOWLEDGMENTS

We are grateful to all the families participating in this research, including the Simons Foundation Autism Research Initiative (SFARI) Simplicity Collection (SSC) and Korean cohort. We thank the SSC principal investigators (A. L. Beaudet, R. Bernier, J. Constantino, E. H. Cook Jr., E. Fombonne, D. Geschwind, D. E. Grice, A. Klin, D. H. Ledbetter, C. Lord, C. L. Martin, D. M. Martin, R. Maxim, J. Miles, O. Ousley, B. Peterson, J. Piggot, C. Saulnier, M. W. State, W. Stone, J. S. Sutcliffe, C. A. Walsh, and E. Wijnsma) and the coordinators and staff at the SSC clinical sites; the SFARI staff, in particular N. Volkovsky; D. B. Goldstein and K. C. Eggan for contributing to the experimental design; the Rutgers University Cell and DNA repository for accessing biomaterials; and the New York Genome Center for generating the WGS data. Annotation data was generated as part of the PsychENCODE Consortium, supported by NIH grants U01MH103392, U01MH103365, U01MH103346, U01MH103340, U01MH103339, R21MH109956, R21MH105881, R21MH105853, R21MH103877, R21MH102791, R01MH11721, R01MH110928, R01MH110927, R01MH110926, R01MH110921, R01MH110920, R01MH110905, R01MH109715, R01MH109677, R01MH105898, R01MH105898, R01MH094714, R01MH109901, P50MH106934, 5R24HD000836 and SFARI #307705 awarded to S. Akbarian (Icahn School of Medicine at Mount Sinai), G. Crawford (Duke University), S. Dracheva (Icahn School of Medicine at Mount Sinai), P. Farnham (University of Southern California), M. Gerstein (Yale University), D. Geschwind (University of California, Los Angeles), I. Glass (Washington University), F. Goes (Johns Hopkins University), T. M. Hyde (Lieber Institute for Brain Development), A. Jaffe (Lieber Institute for Brain Development), J. A. Knowles (University of Southern California), C. Liu (SUNY Upstate Medical University), D. Pinto (Icahn School of Medicine at Mount Sinai), P. Roussos (Icahn School of Medicine at Mount Sinai), P. Sklar (Icahn School of Medicine at Mount Sinai), P. Sullivan (University of North Carolina), F. Vaccarino (Yale University), D. Weinberger (Lieber Institute for Brain Development), S. Weissman (Yale University), K. White (University of Chicago), P. Zandi (Johns Hopkins University), S.J.S., N.S., M.W.S., and A.J.W. **Funding:** Supported by Simons Foundation for Autism Research Initiative (SFARI) grants 402281 (S.J.S., M.W.S., B.D., and K.R.), 385110 (S.J.S., M.W.S., A.J.W., and N.S.), 574598 (S.J.S.), 385027 (M.E.T., B.D., K.R., J.B., and M.J.D.), 346042 (M.E.T.), 575097 (B.D. and K.R.), 573206 (M.E.T.), 513631 (G.T.M.), and 388196 (H.C. and G.T.M.). NIH grants U01 MH105575 (M.W.S.), U01 MH100239-03S1 (M.W.S., S.J.S., and A.J.W.), R01 MH110928 (S.J.S., M.W.S., and A.J.W.), R01 MH109901 (S.J.S., M.W.S., and A.J.W.), U01 MH11662 (S.J.S. and M.W.S.), U01 MH11658 (B.D.), U01 MH11660 (M.J.D.), U01 MH11661 (J.D.B.), R37 MH057881 (B.D. and K.R.), R01 HD081256 (M.E.T.), R01 MH115957 (M.E.T.), R01 MH049428 (J.L.R.), R01 MH107649-03 (B.M.N.), and R01 MH094400 (H.C.), and the Stanley Center for Psychiatric Genetics.

**Author contributions:** Experimental design, J.-Y.A., K.L., L.Z., D.M.W., H.B., N.A., J.D.B., H.C., M.J.D., Y.S.K., G.T.M., B.M.N., A.R.Q., J.L.R., N.S., M.W.S., A.J.W., M.E.T., B.D., K.R., and S.J.S.; data generation, G.B.S., J.D., C.D., Y.S.K., and S.J.S.; data processing, J.-Y.A., D.M.W., S.D., C.D., M.C.G., L.L., and S.J.S.; annotation of functional regions, J.-Y.A., D.M.W., E.M.-P., S.P., J.L.R., N.S., and S.J.S.; data analysis, J.-Y.A., K.L., L.Z., D.M.W., S.D., H.B., H.Z.W., X.Z., G.B.S., R.L.C., B.B.C., L.K., M.E.T., B.D., K.R., and S.J.S.; statistical analysis, J.-Y.A., K.L., L.Z., D.M.W., L.K., M.E.T., B.D., K.R., and S.J.S.; manuscript preparation, J.-Y.A., K.L., L.Z., D.M.W., S.D., N.A., H.C., G.T.M., M.E.T., B.D., K.R., and S.J.S. **Competing interests:** G.T.M. is co-founder and chief scientific officer of Frameshift Labs Inc. B.M.N. is a member of the

Deep Genomics scientific advisory board and a consultant for Camp4 Therapeutics Corporation, Merck & Co., and Avanir Pharmaceuticals. M.W.S. is on the scientific advisory boards for ArRett Pharmaceuticals and BlackThorn Therapeutics and holds stock options in ArRett Pharmaceuticals. **Data and code availability:** All sequencing and phenotype data are hosted by the Simons Foundation for Autism Research Initiative (SFARI) and are available for approved researchers at SFARIbase (<https://base.sfari.org/>), accession ID: SFARI\_SSC\_WGS\_p, SFARI\_SSC\_WGS\_1, and SFARI\_SSC\_WGS\_2). The code for running a category-wide association study can be accessed at <http://doi.org/10.5281/zenodo.1489239>. For methods for estimating the de novo risk score, clustering of annotation

categories, and estimating the significance of clusters of annotation categories, see <http://doi.org/10.5281/zenodo.1489250>.

## SUPPLEMENTARY MATERIALS

[www.sciencemag.org/content/362/6420/eaat6576/suppl/DC1](http://www.sciencemag.org/content/362/6420/eaat6576/suppl/DC1)  
Figs. S1 to S9  
Tables S1 to S9  
References (38–52)

21 March 2018; accepted 15 November 2018  
10.1126/science.aat6576

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

Joon-Yong AnKevin LinLingxue ZhuDonna M. WerlingShan DongHarrison BrandHarold Z. WangXuefang ZhaoGrace B. SchwartzRyan L. CollinsBenjamin B. CurrallClaudia DastmalchiJeanselle DeaClif DuhnMichael C. GilsonLambertus KleiLindsay LiangEirene Markenscoff-PapadimitriouSirisha PochareddyNadav AhituvJoseph D. BuxbaumHilary CoonMark J. DalyYoung Shin KimGabor T. MarthBenjamin M. NealeAaron R. QuinlanJohn L. RubensteinNenad SestanMatthew W. StateA. Jeremy WillseyMichael E. TalkowskiBernie DevlinKathryn RoederStephan J. Sanders

Science, 362 (6420), eaat6576. • DOI: 10.1126/science.aat6576

### View the article online

<https://www.science.org/doi/10.1126/science.aat6576>

### Permissions

<https://www.science.org/help/reprints-and-permissions>