A statistical framework for mapping risk genes from *de novo* mutations in whole-genome sequencing studies

Yuwen Liu,[1] A. Ercument Cicek,[2,3] Yanyu Liang,[3] Jinchen Li,[4,6,10] Rebecca Muhle,[7,8,9], Martina Krenzer,[9] Yue Mei,[4] Yan Wang,[4] Nicholas Knoblauch,[5] Jean Morrison,[1] Yi Jiang,[6,10] Evan Geller,[7,9] Zhongshan Li,[6] Iuliana Ionita-Laza,[11] Jinyu Wu,[4,6] Kun Xia,[10] James Noonan,[7,9] Zhong Sheng Sun[4,6*] & Xin He[1*]

[1]Department of Human Genetics, The University of Chicago, Chicago, IL 60637, USA. [2]Computer Engineering Department, Bilkent University, Ankara 06800, Turkey. [3]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15123, USA. [4]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100000, China. [5]Committee on Genetics, Genomics and Systems Biology, The University of Chicago, Chicago, IL 60637, USA. [6]Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou, Zhejiang 325000, China. [7]Department of Genetics, Yale School of Medicine, New Haven, Connecticut 06520, USA. [8]Child Study Center, Yale Medicine, New Haven, Connecticut 06520, USA. [9]Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, Connecticut 06520, USA. [10]The State Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan 410000, China. [11]Department of Biostatistics, Columbia University, New York, New York 10027, USA. Correspondence should be addressed to S.Z. (sunzs@mail.biols.ac.cn) or H.X. (xinhe@uchicago.edu).

Abstract

Analysis of *de novo* mutations (DNMs) from sequencing data of nuclear families has identified risk genes for many complex diseases, including multiple neurodevelopmental and psychiatric disorders. Most of these efforts have focused on mutations in protein-coding sequences. Evidence from genome-wide association studies (GWAS) strongly suggests that variants important to human diseases often lie in non-coding regions. Extending DNM-based approaches to non-coding sequences is, however, challenging because the functional significance of non-coding mutations is difficult to predict. We propose a new statistical framework for analyzing DNMs from whole-genome sequencing (WGS) data. This method, TADA-Annotations (TADA-A), is a major advance of the TADA method we developed earlier for DNM analysis in coding regions. TADA-A is able to incorporate many functional annotations such as conservation and enhancer marks, learn from data which annotations are informative of pathogenic mutations and combine both coding and non-coding mutations at the gene level to detect risk genes. It also supports meta-analysis of multiple DNM studies, while adjusting for study-specific technical effects. We applied TADA-A to WGS data of ~300 autism family trios across five studies, and discovered several new autism risk genes. The software is freely available for all research uses.

s

Introduction

De novo mutations (DNMs) arise spontaneously in offspring, and are often detected by sequencing patient families, usually trios of parents and affected children ("trio-sequencing"). Researchers can identify risk genes using DNMs by searching for genes that harbor more *de novo* mutations in affected offspring than expected by chance. This approach has been highly successful in studying a range of developmental and psychiatric disorders including autism, intellectual disability, schizophrenia, epilepsy and congenital heart disease.[1–6] DNMs tend to have larger effects than standing variants because they have not yet been acted on by natural selection. The DNM approach may be particularly helpful for early-onset diseases because standing risk variants for these phenotypes are rare and hard to identify with GWAS.

Most existing work on DNMs focuses on mutations in protein-coding regions. Even when whole genome sequence (WGS) data are available, researchers often analyze only the coding portion of the genome due to a lack of analytic tools for non-coding mutations.[7–9] However, the majority of disease-associated variants identified by GWAS are located in non-coding sequences, potentially affecting gene regulation rather than protein function. This suggests that non-coding DNMs represent a large, currently unexplored, source of genetic variation that can aid gene discovery. The knowledge of non-coding disease variants will provide additional benefits. As the activity of regulatory elements tend to be cell type-specific, the analysis of DNM could offer clues as to which cell types are most relevant to disease etiology. A key research challenge is thus to provide an analytic framework for DNM data that incorporates non-coding mutations from WGS studies of disease families.

Current tools for DNM analysis perform some kind of “burden test” which evaluates whether the number of mutations in a gene is larger than would be expected by chance. He et.al. (2013) propose the method TADA for DNM analysis which effectively performs a weighted Bayesian burden analysis.[10] TADA divides all mutations into categories, such as nonsense and missense mutations. Mutations in each category are weighted according to how damaging they are expected to be, with the weights for each category learned from the data. Another method, FitDNM, similarly performs weighted burden analysis, however, the weights are assumed to be known (from external source) instead of being estimated from data.[11]

Unlike protein-coding sequences, there is no simple genetic code for researchers to predict the functional effects of non-coding mutations, and thus difficult to assign them to simple categories. Instead, we can describe non-coding mutations using a number of overlapping genomic annotations such as tissue-specific epigenomic marks and cross-species conservation. Which annotations are relevant to a particular disease is not known *a priori.* Additionally, each annotation may only be weakly informative of pathogenic variants so we may need to combine multiple annotations. Existing tests developed for *de novo* coding mutations can not handle such complications: TADA can only handle disjoint categories of DNMs and has only been used with a small number of mutational categories; FitDNM is designed for exome sequencing data, assuming that the probability of a variant affecting protein function is known (from PolyPhen2).[11]

In this work, we present a novel statistical framework for analysis of DNMs, which we call TADA-Annotations (or TADA-A). TADA-A uses a probabilistic model of mutation counts for each position in the genome (Figure 1). Specifically, we model the mutation counts as following Poisson distribution, and the background mutation rates depend on covariates such as types of nucleotide changes and local GC content. We expect the mutation rates in positions assigned to a disease gene to be elevated compared with background rates, and the fold increases depend on functional annotations in a log-linear model. TADA-A offers several features important for WGS-based DNM studies. First, the model can take an arbitrary number of possibly overlapping annotations, and learn from the data which annotations are enriched for causal mutations. No arbitrary weighting scheme or variant filtering is needed. Secondly, the method predicts risk genes by combining information in both coding and non-coding regions. In addition, the information from coding mutations may come from an independent study, allowing a WGS study to borrow strength from published Whole Exome Sequencing (WES) studies. Finally, TADA-A supports meta-analysis of multiple WGS studies. It adjusts for possible difference in technical factors across studies by fitting a different background mutation model for each study.

We apply TADA-A to study the contribution of non-coding sequences in autism spectrum disorder (ASD). WES studies using DNMs in autistic families have identified more than 70 ASD risk genes, highlighting the importance of DNMs in the study of autism.[1,12–14] Recently, efforts have been expanded to whole genome sequencing of ASD families. Two studies reported modest enrichment of functional non-coding DNMs near known ASD genes in autistic children, comparing with controls or unaffected siblings.[8,15] However, none of the published work has utilized non-coding DNMs to map specific risk genes or functional elements. We use TADA-A

to analyze a collection of five whole-genome DNM datasets, leveraging a number of genomic annotations. We find that brain enhancers marked by H3K27ac, and regions predicted to affect splicing have increased rates of DNMs in ASD cases. Our conservative estimates suggest that regulatory non-coding mutations contribute to about a third of *de novo* autism risk (i.e. autism risk attributable to all DNMs). Using the DNMs from WGS data as well a published WES study, we were able to identify five new ASD risk genes at a FDR < 0.1. Multiple lines of evidence support the possible roles of these genes in ASD.

Results

Overview of TADA-Annotations (TADA-A)

TADA-A consists of two steps. In the first step, study-specific, background mutation rates are estimated at each position. We use an initial estimate from the literature, and then adjust for genomic and technical covariates such as sequencing depth and local GC content. In the second and main step, we use DNMs and annotation information to identify risk genes. We assume that we can assign each DNM to one gene, however, we could also analyze at the level of genomic regions and the DNM-to-gene assignment is not strictly necessary (see Discussions). TADA-A takes as input the number of DNMs at each genomic position, summing over all affected subjects, and a set of possibly overlapping, genomic annotations (Figure 1, upper panel). Genomic annotations might include cell specific histone modifications or evolutionary conservation. TADA-A produces two main outputs: (1) the annotations that are informative of causal mutations and their effect sizes and (2) the predictions of specific susceptibility genes of the

disease of interest (Figure 1, bottom panel). We measure the effect of an annotation by its relative risk, i.e. the fold increase of disease risk for a variant carrying that annotation vs. a variant without that annotation (assuming the annotation is binary). The model of TADA-A is general enough that it can analyze either coding or non-coding mutations. If both types of mutations are analyzed, the results could be easily combined by multiplying the resulting Bayes factors (BFs).

The intuition behind TADA-A is that, in affected individuals, disease-causing mutations should appear at higher rates than expected from the baseline mutation rate. Our model can be written as $y_i \sim Poisson(2N\mu_i\gamma_i)$, where y_i is the observed number of *de novo* mutations at position i (usually 0 or 1), μ_i is the expected background mutation rate estimated in the first step, N the sample size, and γ_i is the relative risk of a DNM at position i (greater than 1 for risk mutations). To model the relative risk, γ_i, we define a binary (unobserved) variable for each gene indicating whether it is a risk gene or not. For a non-risk gene, all its positions have relative risk equal to 1. For risk genes, we model $\log(\gamma_i)$ as a linear function of the genomic annotations. Each gene has a prior probability of being risk gene. TADA-A offers the option of using informative prior probabilities, e.g. a likely risk gene from previous WES studies would have high prior probability. Intuitively, this allows us to put more emphasis on highly plausible risk genes to estimate the parameters of annotations, while discounting the unlikely disease genes. This is important when statistical signals in the annotations are weak.

We estimate model parameters (mainly the relative risk of each annotation) using maximum likelihood. Since annotations could be partially redundant (e.g. an enhancer may be associated

with multiple annotations such as open chromatin and H3K27ac) and not all annotations are informative, we implement a stepwise regression procedure to select a subset of most informative annotations. Once we have estimates of all the parameters, we predict whether a gene is a risk gene or not using Bayes factor (BF), combining information in all its associated DNMs. Similar to the original TADA method, we test each gene separately, and contrast the null model where the relative risk is always 1 with the alternative model described above for risk genes (relative risks dependent on annotations). To use TADA-A in a meta-analysis setting that combines multiple studies with possibly different rates and patterns of DNMs (e.g. in relation to GC content), we fit a different background mutation model for each study, but estimates a common set of parameters related to functional annotations.

TADA-A can be used to answer several questions about genetics of a complex disease, ASD in our case. What annotations are associated with causative mutations? Based on this knowledge, can we learn about the genetic architecture of the disease, especially about the relative contribution of coding vs. non-coding DNMs to the disease liability? Finally, can we identify specific disease genes? We present below our results in answering these questions for ASD.

Whole genome sequencing data of ASD and mutation rate calibration

We analyzed DNM data from five WGS studies of ASD trios or quartets, with a total of 314 affected subjects (Table S1). Mutation data is limited to *de novo* SNVs. The validation rate of *de novo* SNVs based on Sanger sequencing ranges from 85% to 94% in the five studies[16–18]. The number of DNMs per subject ranges from 57 to 63. Additionally, we collected the control data

from a cohort of 700 non-ASD subjects. While TADA-A does not need control data, we use this additional dataset to perform burden analysis often employed in DNM studies, comparing the rate of DNMs in affected subjects with the rate in controls (see Methods). Our main results are limited to sequences close to genes, including protein-coding sequences, non-coding sequences within $\pm$10 kb of TSSs, and potential splicing-regulatory regions that are not covered by these two categories. In a later section, we present results based on distal sequences.

To account for technical difference among studies that may affect observed mutation rates, we use published mutation rates from Samocha et al. as an initial estimate,[19] and adjust for covariates using a Poisson regression model, separately for each of the five datasets (see Methods). We perform analysis at the level of 50bp windows and consider three covariates for estimating baseline mutation rates: (1) whether a window is in coding regions (transcribed sequences may have lower mutation rates because of transcription-coupled repair); (2) whether the window is in promoter regions (CpG (de)methylation in promoters might affect mutation rates); (3) the percent GC in the window, which may correlate with sequencing depth and hence DNM detectability.[8] The effects of these covariates are summarized in Table S2. In all of the five studies, the intercept in the regression model is significantly different from 0 ($P < 0.05$), suggesting systematic departure of average mutation rates from the published rates. In three of the five ASD studies, GC content has a negative effect on the observed DNM rate ($P < 0.05$ for three). The mutation features representing whether a sequence is in coding or promoter region were found to have a relatively large effect in specific studies (e.g. the coding feature in Jiang et al. was significant at $P = 0.003$). The results from our mutation rate modeling thus support the importance of accounting for difference in studies in meta-analysis of DNM datasets.

Risk-increasing mutations in ASD are associated with active enhancer mark and damaging effects on splicing

We first assessed the quality of data using coding DNMs. We performed a simple burden analysis of protein-coding sequences in probands vs. controls. We adjusted for the difference in baseline mutation rates in the ASD studies and controls using synonymous mutations (whose true mutation rates should be the same across studies, see Methods). As expected, we found that the average rate of non-synonymous mutations per subject is about 1.2 fold higher in ASD subjects vs. controls (Figure 2A), in line with previous estimates.[2,12,20] We also observed an increased rate of non-synonymous mutations in gene sets enriched with ASD risk genes, including known ASD genes, genes likely involved in neuropsychiatric disorders (dubbed "neuropsychiatric genes"), genes intolerant to mutations, and genes highly expressed in the brain (Fig 2A). Only the burden in mutation intolerant genes is statistically significant ($P < 0.01$).

For TADA-A analysis, we use a total of seven functional annotations (Table S3). Some annotations measure the regulatory function of variants, including fetal and adult brain H3K27ac,[21,22] fetal[22] and adult brain[23] DNase Hypersensitive Sites (DHS). H3K27ac is a mark of active enhancers and DHS is a mark of open chromatin, often suggestive of regulatory functions. We also use two conservation-based annotations, PhyloP,[24] and GERP,[25] and an aggregate variant score CADD.[26] Splicing has been shown to be important for many human diseases, so we include the splicing effects predicted by SPIDEX.[27] We limit our analysis to sequences within 10 kb of transcription start sites (TSSs) of protein-coding genes (not including UTRs) and

potential splicing-regulating regions (which could be far away from TSSs). To increase the power of TADA-A to detect predictive annotations, we take advantage of existing WES studies. For each gene, we summarize the findings of previous WES studies as the probability of being an ASD risk gene[2], which is then used as the prior probability of being a risk gene in the TADA-A model. This step allows us to put large weights on known ASD risk genes, whose probabilities are close to 1, comparing with average genes (about 0.05).

In our initial analysis using TADA-A, we found that among seven annotations, only brain H3K27Ac and the SPIDEX score make marginally significant contributions ($p < 0.05$, Table S3). We therefore retrain the model using only these two features and estimate their relative risks at 1.8 and 3.5 respectively (Table 1 and Table S3). In the following analysis, we refer to *de novo* SNVs in H3K27 regions within 10kb of genes as regulatory SNVs, and *de novo* SNVs predicted by SPIDEX to affect splicing as splicing SNVs.

We sought to replicate the TADA-A results using burden analysis with independent control samples (note that TADA-A analysis was performed only on ASD data). Burden analysis on all regulatory SNVs did not reveal any enrichment of mutations in ASD cases (Figure S1, the numbers of synonymous SNVs in ASD and controls were used as background). This is not entirely unexpected because TADA-A incorporates prior information about plausible ASD genes, while burden analysis does not. Similarly, the commonly used annotations for non-coding variants, GERP++[25], PhyloP[24] and CADD[26] were not significantly associated with mutational burden (Figure S1). We found a 2.5-fold enrichment of regulatory SNVs near known ASD genes ($P = 0.029$) and a 1.2-fold enrichment near a set of genes likely involved in neuropsychiatric

phenotypes, called "neuropsychiatric genes" (P = 0.047) (Figure 2B, see Methods for definitions of gene groups). In contrast, we observed no enrichment of regulatory SNVs near genes that are unlikely to be ASD risk genes (non-ASD genes). We also did not observe enrichment of regulatory SNVs that are farther away (10-50kb) from TSSs in ASD genes or neuropsychiatric genes

For splicing SNVs, we found a significant burden in neuropsychiatric genes (Figure 2C, P = 0.033) as well as in genes intolerant to mutations (Figure 2C, P = 0.033). In contrast, no significant burden was found in SNVs not predicted to affect splicing ("control mutations" in Figure 2C). Similarly, we found that in genes less likely to be ASD genes, such as non-ASD genes and genes tolerant to mutations, there is no enrichment of splicing SNVs. Different from regulatory SNVs, we did not observe enrichment of splicing SNVs in known ASD genes. We think this is due to small sample sizes and a small gene set: there are very few splicing mutations falling into these ASD genes (0 in ASD and 2 in control).

Enhancer and splicing mutations make substantial contributions to the *de novo* risk of autism

A fundamental question in genetics is how the risk variants are distributed among various functional classes, such as protein-coding sequences, enhancer sequences, non-coding RNAs, etc. This question has been studied recently using common variants.[28,29] It was found that even though variants in protein coding regions are highly enriched with risk variants, they explain only a small fraction of total disease risk. The results from the previous TADA-A analysis allows us to address this "risk partition" problem from a different angle, using DNMs. Based on TADA-

A results, we considered two types of non-coding *de novo* SNVs: SNVs in brain enhancers near TSSs (regulatory SNVs) and splicing SNVs, in addition to two classes of coding mutations: LoF and probably damaging missense (predicted by PolyPhen-2, denoted as Mis3). We quantify the contribution of a mutation type as Liability Variance Explained (LVE), taking into account both the frequencies of this mutation type and its average relative risk (see Methods). For coding mutations (LoF and mis3), the relative risks were obtained from published TADA estimates in WES studies.[10] For non-coding SNVs, we used the relative risks estimated by TADA-A.

The relative risks of regulatory SNVs and splicing SNVs are significantly lower than those of coding SNVs (Table 1). Despite a lower risk per variant, regulatory SNVs are much more frequent than other classes of mutations, making the total contribution of regulatory SNVs comparable to LoF or missense coding mutations. Each class of mutation explains only a small fraction of estimated total ASD genetic risk (Table 1), consistent with the conclusion of an earlier study.[30] Considering only the risk due to *de novo* mutations, we found that non-coding SNVs (including regulatory SNVs and splicing SNVs) explain 37% of the *de novo* risk (Figure 2D). This estimate, however, is likely very conservative (Discussion).

<u>TADA-A identifies novel ASD risk genes by combining coding and noncoding mutations</u>

A recent WES study (~3,500 samples) identified 58 ASD risk genes at FDR < 0.1[2] using *de novo* SNVs. Applying TADA-A on the WGS data and combining them with the WES results, we discovered five new ASD risk genes at FDR < 0.1, and 14 at FDR < 0.3. Each of the five genes at FDR < 0.1 (Table 2) has at least one LoF or Mis3 mutation, and the evidence for these genes is

strengthened by the presence of regulatory or splicing SNVs. We found extensive evidence supporting the plausibility of these genes as ASD risk genes. *APBB1* and *NRXN1* are the targets of neuronal-RNA binding protein FMRP, whose loss of function causes Fragile X syndrome and autistic features (Table 2, hypergeometric test, P = 0.017). Four of the five genes have been identified as haploinsufficient genes[31,32] (P = 0.00034) and are highly expressed in the brain (top 25% of all genes, P = 0.015). The genes also tend to be evolutionarily constrained as measured by either RIVS (RVIS top 25% genes, P = 0.010) or another metric based on tolerance of LoF variants in ExAC (ExAC top 25% genes, $P = 0.087$) (Table 2). Evolutionary constraint in the human population has been shown to be a strong predictor of autism genes[19].

We performed network analyses to further establish the link of new candidate genes to autism. DAWN[33,34] is a recently developed method that predicts autism risk genes by virtue of the genes' association with known ASD genes in co-expression networks of early developing brain. A gene receives a high DAWN score if it is highly connected with other likely ASD genes. We found that three out of the five new genes (*JUP*, *ARHGAP5,* and *NRXN1*) have DAWN q-value < 0.05 in at least one of the two critical spatial-temporal developmental windows for ASD[35] (Table 2, enrichment P = 0.00049). Using GeneMania,[36] we found our candidate genes were highly connected to high-confidence ASD genes in the gene co-expression network constructed from multi-tissue gene expression data (112 co-expression links between the two gene sets, P = 0.04, Figure 3A). These analyses, using different analytic tools and genomic data, thus support that our new genes are functionally related to known ASD genes.

Expanding the analysis to the 14 new genes at FDR < 0.3, we observed significant enrichment of multiple gene annotations (Table S4), including haploinsufficient genes (P = 0.00010), FMRP targets (P = 0.021), constrained genes (P = 0.00089 using RVIS and P = 0.0066 using variant frequency in ExAC), high expression level in the early developing brain from BrainSpan (P =0.035) and genes significantly co-expressed with known ASD genes from DAWN analysis ($P = 8.84\times10^{-9}$). The 14 new genes are also significantly enriched in genes predicted to be ASD risk genes (FDR < 0.05) by a recently developed machine learning approach that utilizes a brain-specific functional gene network (P = 0.037).[37] Literature inspection provides further support of the roles of most of these genes in ASD (Table S5).

We next explore the mechanisms by which non-coding mutations contribute to ASD. We found an interesting example in *PLXNB1* (a new risk gene with q = 0.22). This gene has important functions in neurodevelopment (see Discussions). In the promoter region of *PLXNB1*, an SNV disrupts the binding motif of the transcription factor RFX2 (Figure 3B). RFX2 is a transcription factor highly expressed in the tegmental region of the the fetal mid-brain, a region linked to cognitive abilities[38] and social reward processing.[39,40] An intronic variant in *RFX2* has been associated with schizophrenia and bipolar disorder.[41] Additionally, the RFX2 binding motif is enriched in regulatory regions of several co-expression gene modules (Modules 1, 5, 7, and 15)[21] that are active in the early developing brain. Taken together, our analyses offer a model that the mutation in *PLXNB1* promoter changes the affinity of RFX2 binding, and as a result, perturbs the expression of *PLXNB1* and increases the risk of ASD.

<u>Distal enhancers and TADs with multiple *de novo* mutations implicate additional risk genes</u>

Our TADA-A analyses were performed at the gene level, and only considered enhancers within 10 kb of TSSs. We use a different approach in this section to identify distal enhancers with roles in autism, based on the notion that the probability of multiple DNMs occurring in a single enhancer is very low. We found 25 H3K27ac enhancers with two or more SNVs in ASD cases, significantly higher than random expectation based on simulations (Figure 3C, $P = 0.0025$). We predicted the likely target genes of recurrent enhancers based on cross-tissue correlation between enhancer activity and gene expression from Roadmap Epigenomics (Table S6). We found a recurrent enhancer putatively targeting *ZMIZ1*, more than 250 kb away (Figure 3D). The region contains two other DNMs in two enhancers, one of which also has correlated activities with the ZIMZ1 promoter. A target of FMRP, *ZMIZ1* is highly expressed in the brain and interacts with neuron-specific chromatin remodeling complex (nBAF) which is important in regulating synaptic functions.[42,43] Several nBAF members have been linked to autism, such as ARID1B and BCL11A.[44] The pathogenic potential of *ZMIZ1* is further supported by the observation of a *de novo* gene-disrupting translocation in an individual with intellectual disability.[45] These results strongly support the role of ZMIZ1 in autism, and also highlight the mechanism that DNMs may increase ASD risk by disrupting distal regulatory elements.

We applied the similar idea of recurrent DNM analysis at the level of Topologically Associating Domains (TADs).[46] These are megabase–sized chromatin interaction domains that are stable across cell types and have been proposed to demarcate transcriptional regulatory units.[46] Based on estimated mutation rates, we found two TADs with a significant (at FDR < 0.125) number of regulatory SNVs (Figure S2 and Table S7). In both TADs, there are only two or three genes, and

we conjecture that SRBD1 and MRSA are likely the underlying ASD genes in the two TADs, respectively (see Discussion).

Discussion

Analyzing DNMs from exome sequencing data has been shown to be a powerful new paradigm for mapping risk genes of developmental and psychiatric disorders. Extending this to the non-coding genome is the natural next step and has the potential to transform our understanding of these complex disorders. In this work, we present a comprehensive statistical framework to support such analysis. The new method, TADA-A, is able to leverage multiple functional genomic annotations to better detect and prioritize risk-predisposing mutations. More importantly, TADA-A is able to combine information of all DNMs of a gene, in both coding and non-coding regions, to maximize the power to detect risk genes. The results of our meta-analysis of autism WGS datasets demonstrate the effectiveness of TADA-A. We show that *de novo* non-coding mutations make substantial contributions to the risk of autism (comparable to *de novo* LoF or missense mutations), and identified several promising new ASD risk genes. We note that the method can be applied to any units, other than genes, such as regulatory elements or sequence windows, though analysis at the gene level has the benefit of permitting us to borrow external information from previous WES studies.

A common strategy for analyzing DNM data is the burden analysis, which contrasts the rates of DNMs in patients, often limited to likely functional mutations, with the expected rates due to chance alone. When researchers have no matched sibling or control data, the burden analysis can

be confounded by technical factors such as sequencing depth. The burden analysis in non-coding regions is even more challenging because the statistical signal is considerably weaker than the coding signal, as reported by recent publications as well as our own study (Table 1). TADA-A greatly improves the standard burden analysis in several ways. Our mutation model, based on Poisson regression, incorporates covariates known to influence background mutation rates. In our ASD analysis, while we do not have access to genome-wide sequencing depth information, we used GC content as a proxy. Incorporating prior information of which genes are likely risk genes is critical for estimating parameters of annotations, as the genome-wide burden of genomic annotations is very close to 1 (Figure S1).

Previous knowledge of the role of non-coding variants in diseases comes mostly from GWAS. The challenge with GWAS is that regulatory elements are much shorter (~1kb) than regions of linkage disequilibrium (LD, hundreds of kb on average). It is thus not straightforward to assess the contribution of non-coding variants, or identify specific regulatory elements from GWAS.[28] Indeed, the estimated contribution of DHS sites to heritability of complex diseases ranges widely from 79% to 25% in literature,[28,47] largely because of LD. By using DNMs, our work provides independent estimation of the contribution of both coding and non-coding variants to the risk of complex diseases.[28] We estimated a modest average relative risk of about 2 for *de novo* mutations in brain H3K27ac enhancers, compared to 4-5 for missense and 20 for LoF mutations. We were not able to detect signal in evolutionarily conserved sequences (GERP and PhyloP) or putative deleterious variants (CADD, trained mostly from non-brain tissues). These results suggest that evolutionary constraint is only weakly correlated with pathogenicity in ASD,[48] and that regulatory variants of ASD probably act in a tissue and time-specific manner.[49] Compared to

a previous study,[30] our estimate of ASD risk attributable to coding mutations is somewhat higher (1.9% vs. 1.1%), mainly due to a significant contribution from missense DNMs (0.83% vs. the previous estimate of 0.04%). We believe this difference is due to our different modeling assumptions: we treated all mutations in a category as a mixture of causal and non-causal mutations, whereas the previous study treated all mutations in a category equally (see Methods).[30] We estimated that *de* novo coding (1.9%), non-coding (1.11%) and copy number variants (1.46%, estimated by a previous study[30]) together contribute 4.5% of ASD risk. We think that we significantly under-estimated the contribution of non-coding mutations to ASD risk for several reasons. First, we considered only enhancers within 10 kb of TSSs, which constitute about 36% of all enhancers in our data. Second, our dataset contains only regulatory sequences active early in development (5, 7, and 12 weeks post-conception) or in the adult brain. Third, *de novo* indels and copy number variants (CNVs) were not considered in this study.

We identified five new ASD risk genes, four of which are strongly supported by other evidence. JUP is a member of the catenin/cadherin superfamily which has important roles in neuron connections and interactions.[50] It is strongly expressed in the primate prefrontal cortex and hippocampus.[51] APBB1 is an adaptor protein localized in the nucleus. It is down-regulated in ASD cerebellum than control,[52], and its microexons are mis-regulated in the brains of ASD individuals.[53] ARHGAP5 is required for normal neural development, and is implicated in axonal branching and synapse formation.[54,55] NRXN1 belongs to a group of presynaptic cell adhesion molecules which controls synapse development.[56] It has been implicated as a top candidate gene for neurodevelopmental and neuropsychiatric conditions.[57] Interestingly, alternative splicing of

Nrxn1 has been reported to cause defects of synaptic formation in the hippocampus region in a mouse model (the gene is supported by a splicing SNV, Table 2).[58]

Most of the new ASD genes at FDR < 0.3 are supported by functional or association studies (summarized in Table S5). PLXNB1 is a receptor in the GTPase signaling pathway that regulates the development of hippocampal neurons,[59,60] remodeling of dendrites,[61] and the plasticity of axons.[62,63] *PLXNB1* expression has been found to be altered in the prefrontal cortex of schizophrenia patients.[64] *LRRC4* is a member of the synaptic cell adhesion molecule family, of which some are known to be associated with ASD.[65] In the mouse hippocampus, LRCC4 is critical for axon differentiation, pathway-specific synapse development, and functional integration of distinct inputs.[66,67] Recently, a 1.9 Mb *de novo* deletion spanning *LRRC4* has been found in an individual with severe intellectual disability.[68] *PPM1D* has recently been identified as a novel risk gene for intellectual disability.[69] *MSL2*, *CBX4* and GAPVD1 are predicted as autism risk genes by a recently developed machine learning approach that utilizes a brain-specific functional interaction network[37] (q-values 0.038, 0.01 and 0.08 respectively).

We also found that the two TAD regions with excess regulatory SNVs in ASD are supported by CNV studies. In one TAD, recurrent, rare CNVs (Chr2: 45455651-45984915) spanning the entire *SRBD1* gene (the only protein coding gene disrupted by the CNVs within this TAD) were reported in ASD subjects.[70] In a later independent study, CNVs in this TAD region were found to be enriched in ASD cases versus controls.[70] These results suggest that *SRBD1* is likely the risk gene in this TAD. In the other TAD region, ASD-associated duplication of 8p23.1-8p23.2 introduces a breakpoint between MSRA and RP1L1.[71] MSRA is a member of the methionine-

sulfoxide reductase system whose function is to alleviate oxidative stress. Increased exposure to oxidative stress plays an important role in the pathogenesis of ASD.[72] In addition, GWAS studies have established associations of MSRA with schizophrenia[73] and bipolar disorder.[74]

Methods

TADA-A model

TADA-A works in two stages: first, it calibrates the background mutation rates (mutation model); second, it learns which functional annotations are predictive of causal mutations and infer the risk genes (functional model). In the mutation model step, we assume that we have un-calibrated base-level mutation rates from external data, e.g. from human-chimp comparison. We used the estimates from Samocha et al. as our baseline rates.[19] For a particular study, observed mutation rates may differ from these un-calibrated rates as a result of study-specific technical factors. For example, lower sequencing depth reduces the number of called DNMs. Mutation rates may also depend on local genomic features such as GC content. To account for this variability, we calibrate the background mutation rates for each study. To simplify the computation, TADA-A collapses DNMs in a 50-bp genomic window into a single count. We model these mutation counts as Poisson Generalized Linear Model (GLM):

$$Y_i \sim Poisson(2N\tilde{\mu}_i \cdot \exp\,(\alpha_0 + \sum_k \alpha_k U_{ik}))$$

where N is the number of individuals, $\tilde{\mu}_i$ is the un-calibrated baseline mutation rate of window i (summing up the mutation rates of bases in window i), and the exponential term represents the deviation of the actual mutation rate of window i from $\tilde{\mu}_i$. The variable U_{ik} is the k-th mutation-related feature of window i and α_k represents the effect of mutation feature k. The mutation features may include GC content, whether a sequence is transcribed, etc. TADA-A uses glm() in R to estimate the coefficients of the genomic features. After fitting the model, the calibrated mutation rate at each base is the un-calibrated rate multiplied by a factor, $\exp(\alpha_0 + \sum_k \alpha_k U_{ik})$, which is calculated for the window containing that base.

In the functional model, we model the dependency of mutation rates on the gene status (risk gene or not) and functional genomic annotations. As described in the main text, we assume all mutations have been uniquely assigned to genes. Let Z_g be a binary indicator of whether gene g is a risk gene or not. When gene g is a risk gene ($Z_g = 1$), the number of DNMs at base i aggregated from affected individuals follows a Poisson distribution

$$Y_i | Z_{g(i)} = 1 \sim Poisson(2N\mu_i \exp\left(\sum_k \beta_k X_{ik}\right))$$

where $g(i)$ is the gene that base i belongs to, μ_i is the calibrated mutation rate of base i from the previous step, X_{ik} is the k-th genomic annotation of base i, and β_k is the effect of the k-th annotation. Note that we consider annotations related to function at this step, such as conservation and enhancer activity. Since the annotations are binary, e^{β_k} is the relative risk of the k-th annotation, i.e. the fold increase of mutation rates in positions with that annotation vs.

those without the annotation. If gene g is a non-risk gene ($Z_g = 0$), the DNM number at base i simply follows $Y_i|Z_{g(i)} = 0 \sim Poisson(2N\mu_i)$.

Let Y_g be the number of DNMs assigned to gene g. Assuming that DNM events are independent, the likelihood of Y_g given Z_g, $P(Y_g|Z_g)$, is simply the product of the probabilities of DNM counts at all bases, according to the equations above. Let π_g be the prior probability of gene g being a risk gene, we have the full likelihood over all genes:

$$P(Y|\beta,\pi) = \prod_g [\pi_g \, P(Y_g|Z_g = 1, \beta) + (1 - \pi_g) P(Y_g|Z_g = 0)]$$

TADA-A implements two options for estimating parameters, both based on maximum likelihood. In the first option, π_g is the same for every gene, and we estimate its value by maximum likelihood jointly with β. In the second option, we use informative priors for π_g of all genes from external data, and we only estimate β. The confidence intervals of the parameters are based on standard asymptotic approximations using Fisher information matrix. When we have multiple annotations in the model, TADA-A uses a procedure similar to backward feature selection to choose annotations. Specifically, it first fits a model with all annotations. The annotations whose coefficients are not significantly different from 0 (at 95% confidence interval) will be removed, and we refit the model with the remaining features.

Once we estimate the parameters, we compute the Bayes factor (BF) of a gene, as

$$P(Y_g|Z_g = 1, \hat{\beta}) / P(Y_g|Z_g = 0),$$

where the probabilities are evaluated at the MLE of parameter values. In our ASD analysis, we further multiply the BFs from non-coding analysis with the BFs from previous results based on coding mutations to obtain final BFs for all genes. We control for multiple testing using the Bayesian FDR control procedure.[75]

Because our likelihood is defined over all bases, including those possessing no DNM events, naïve parameter estimation is computationally expensive. To alleviate this computational burden, we reformulate the likelihood function by collapsing mutations over all bases sharing the same set of annotations, assuming all annotations are discrete. This strategy greatly reduces the computation time (Supplemental Methods). TADA-A software is available at https://github.com/TADA-A/TADA-A.

DNMs from whole genome sequencing data

The detailed information for each WGS dataset is summarized in Table S1. To remove erroneously called *de novo* SNVs, we excluded 8 individuals with more than 140 (2 times more than the median of ASD DNMs per individual) DNMs and removed all recurrent DNMs (i.e. exactly the same mutation in multiple individuals). Our unpublished (Wu et al.) DNM data are from WGS of 32 ASD trios of Han Chinese ancestry (Supplemental Methods, data are available at http://wwwdev.ebi.ac.uk/eva/?eva-study=PRJEB14713).

Non-coding annotations used in analyzing ASD data

Histone modifications: We used H3K27ac sites in fetal and adult brains to define cis-regulatory regions. Fetal brain sites from human cortex at embryonic stages 5, 7 and 12 p.c.w. were obtained from a recent study.[21] For each stage, only peak regions consistent between two biological replicates were selected. Adult brain H3K27ac sites were obtained from Roadmap Epigenomics Project.[22] They include regions from human angular gyrus, anterior caudate, cingulate gyrus, middle hippocampus, inferior temporal lobe, mid-frontal lobe and substantia nigra. We used MACS2 to call peaks from raw data, and only kept peak regions consistent between two biological replicates for each brain region. We used BEDtools[76] to merge H3K27ac sites from fetal and adult brain.

DNase I hypersensitivity sites: Fetal brain DNase I sites were downloaded from Roadmap Epigenomics (male and female fetal brain) and Adult brain DHS data downloaded from ENCODE (Cerebrum_frontal_OC, Frontal_cortex_OC and Cerebellum_ OC).

Conservation scores: We used ANNOVAR to obtain GERP++ scores for all mutations.[25,77] We also downloaded PhyloP scores from the 100-way PhyloP wiggle file from UCSC (/goldenpath/hg19/phyloP100way/). We binarized Gerp (a base is considered to be conserved if Gerp is greater than 2), and PhyloP scores (a base is considered to be conserved if its PhyloP score is greater than 2).

CADD: We downloaded CADD scores from http://cadd.gs.washington.edu/score, (default parameters, v1.3). We binarized CADD scores (deleterious if one allele has a CADD score greater than 10).

Splicing score: We used results from SPIDEX, a deep learning-based approach to annotate variants that may affect splicing.[27] An SNV is considering a splicing SNV if its delta-psi score is less than -1.416, which is the 10% quantile of all positions with SPIDEX scores.

Meta-analysis strategy and Applying TADA-A to ASD

In the ASD study, we first calibrate mutation rates of each study separately using the mutation model of TADA-A (Poisson regression), as described in Results. We then fit the functional model of TADA-A with non-coding annotations listed in the previous section. Since the calibrated mutation rates for any base may differ between studies, we calculated the likelihood of all genes for each study separately, and then multiplied the likelihoods over all studies to get a total likelihood. Note that the coefficients of the functional annotations are shared among multiple studies. We then estimate parameters via maximum likelihood. We take advantage of a previous Autism study to set the prior probabilities of risk genes π_g.[2] Specifically, we convert the Bayes factors reported in that study to posterior probabilities (assuming each gene has 6% chance of being ASD gene), and use these probabilities as π_g. For computational reasons, we used the top 1000 genes, ranked by π_g, in estimating the annotation parameters, since these genes are the most informative of parameters. After the first round of model fitting using all seven annotations, only brain H3K27ac and splicing effects had significant effect sizes, so we refit the model with only these features.

To identify specific ASD risk genes, we first derived the noncoding BF of each gene using data from each WGS dataset, then multiplied these BFs to get a total noncoding BF for each gene. This BF for a gene is multiplied with the published coding BF based on WES studies[2]. We then estimated q-values as mentioned previously.[75] We used $\pi = 0.06$ as the fraction of ASD risk genes in this step.[10,78] We call a gene "novel ASD gene" (at a particular q-value cutoff) if its final q-value falls below the cutoff and its coding q-value from the previous study is above the cutoff. A gene that has no noncoding DNM will not be considered a new finding (even if a gene has no evidence from non-coding mutations, its q-value could still change).

Definition of gene lists used in the analysis

Known ASD genes (194 genes) include genes with q-value < 0.3 from a combined analysis of CNVs, indels, and WES data using TADA[2], SFARI category I (high confidence) and II (strong candidate) genes (https://sfari.org). Neuropsychiatric genes (2601 genes) are a larger set of genes likely involved in neuropsychiatric disorders, including genes with TADA q-value < 0.5,[2] SFARI genes (Category I high confidence, Category II strong evidence) (https://sfari.org), AutismKB genes,[79] ASD risk genes summarized in a previous study,[80] intellectual disability genes,[81] the union of gene sets enriched with SCZ *de novo* coding mutations,[82], high confidence postsynaptic density genes[83] and FMRP targets.[84] The set of nonASD genes are the 1000 genes with the highest TADA q-values.[2] Intolerant genes include genes with top 5% RVIS[48] and haploinsufficient genes obtained from two sources, one using copy number variations (genes with predicted haploinsufficient probability greater than 0.95),[31] and the other using estimated mutation rate.[32] To define tolerant gene, we started with genes with RVIS scores in the bottom

10%,[48] genes with haploinsufficient probability smaller than 0.1,[31] and genes that were used as control genes for LoF deficient genes.[32] We then removed any genes that were in the intolerant gene set. To define gene groups based on their expression levels, we used the average expression level for each gene across all developing brain tissues from BrainSpan data.

Burden analysis of different types of *de novo* mutations

In our burden analysis, we accounted for the difference in mutation rates between ASD subjects (~60/individual) and controls (~39/individual) using "background sequences" (sequences/mutations not expected to have function). Specifically, to assess the burden of nonsynonymous DNMs in ASD vs. controls, we used the numbers of synonymous SNVs in ASD and controls used as background. We tested if nonsynonymous DNMs were enriched in ASD vs. controls using Fisher's exact test, and the burden was defined as the odds ratio (OR) from the 2 by 2 test.

For the burden of noncoding SNVs, we followed the same strategy but used different background mutations. To assess the enrichment of regulatory SNVs (SNVs in brain H3K27ac regions within 10kb of TSS) in a given gene set, we used all regulatory SNVs in the genome as the background. To assess the enrichment of splicing SNVs in a given geneset, we used all SNVs with SPIDEX scores as background mutations.

Assessing contribution of DNMs to ASD risk

We treated ASD liability (risk) as a continuous trait, and estimated the percentage of variance in ASD liability explained by four types of mutations. The variation of ASD liability explained by the *j-th* type of mutations is expressed as : $V_j = \beta_j^2 * p_j * (1 - p_j)$, where β_j is the effect size of the *j-th* type of mutations at the liability scale and p_j is the probability that an individual carries a mutation of type j (see Supplemental Methods). Note that only causal mutations contribute to ASD liability, so both β_j and p_j are defined for mutations affecting causal genes. We calculated β_j from the relative risk of j-th type of mutation using standard quantitative genetic calculations. To obtain p_j, we calculated the total mutation rate of type j mutations, and then multiply this by 0.06 (fraction of ASD risk genes) to obtain the rate of causal mutations of type j.

Network analysis of candidate ASD genes

We used two tools, DAWN and GeneMania to analyze the connectivity pattern of our candidate ASD genes in gene networks. DAWN (Detecting Association With Networks) algorithm[33,34] is a guilt-by-association-based gene prediction algorithm. Its fundamental assumption is that risk genes tend to be functionally related with each other, and thus tightly connected in gene networks. A gene has a high posterior risk probability if it has a high prior risk probability, interacts in a network with other risk genes, or both. The prior risk probabilities came from published WES results.[78] For the underlying network, we constructed partial co-expression networks for two spatial-temporal windows: the mid-fetal prefrontal cortex (PFC) and the infancy mediodorsal cerebellar cortex (MD-CBC), which are indicated as high risk windows for ASD.[35] BrainSpan microarray dataset is used as the source for spatial-temporal gene expression data (http://developinghumanbrain.org). DAWN was run separately for each above-mentioned

network. We used regularization parameter (lambda) = 0.12, p-value cutoff = 0.1 and correlation thresholds 0.7 for PFC and 0.85 for MD-CBC, respectively. In Table 2, posterior risk scores (q-values) are shown for the candidate genes. A dash means that the corresponding gene is not co-expressed with other risk genes in any of the spatial-temporal windows.

GeneMania[36] is a tool for studying interactions among genes in a network using various types of information, such as gene co-expression and protein-protein interactions (PPIs). We studied the connection between our candidate genes with high-confidence ASD genes (genes with coding TADA FDR < 0.1 and genes in SFARI categories I and II, https://sfari.org) using co-expression data. The significance of the number of connections is assessed by randomly sampling gene sets of the same sizes as the candidate genes.

Enhancers with recurrent *de novo* SNVs

We used all brain H3K27ac regions not overlapping with exons (not limited to sequences within 10kb of TSS). We observed 25 enhancers with at least two *de novo* SNVs in ASD samples, and we performed simulations to assess significance. In each simulation, we randomly re-distributed *de novo* SNVs of all brain enhancers, following a multinomial distribution. The multinomial probability of an enhancer is the ratio between the mutation rate of that enhancer and the sum of mutation rates across all enhancers. We performed simulations 10,000 times and obtained the distribution of the number of enhancers with recurrent SNVs.

TADs with recurrent *de novo* SNVs

For each TAD region, we calculated the regulatory mutation rate as the sum of per-base mutation rates of brain H3K27ac sites within the TAD. Under the null hypothesis, the count of regulatory SNVs follows a Poisson distribution, whose rate is the regulatory mutation rate times twice of the sample size.[10,19] We then calculated the *p*-value of each TAD region using the Poisson test and used the Benjamini-Hochberg procedure to control FDR.

Contributions

Y.L.[Yuwen] and X.H. led the study design and the writing of the manuscript; Y.L.[Yuwen] performed all the major analyses; A.E.C. performed network analysis with DAWN and GeneMania. Y.L.[Yanyu], N.K. and I.I.-L contributed to burden analyses; J.L., Y.M., Y.W., Y.J., Z.L., J.W., K.X., and Z.S.S. generated new WGS data from ASD trios, called DNMs, and calculated mutation rates for non-coding regions; R.M., M.K., E.G., and J.N. provided developing brain enhancer data, and contributed ideas to the burden analysis and TAD analysis. All authors contributed to the writing of the manuscript. This work was supported by National Institutes of Health grant (1R01MH110531) and Simons Foundation award (SFARI Award ID 385027) to X.H.

Competing financial interests

The authors declare no competing financial interests.

References

1. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515,** 209–215 (2014).
2. Sanders, S. J. *et al.* Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87,** 1215–1233 (2015).
3. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19,** 1194–1196 (2016).
4. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506,** 179–184 (2014).
5. Allen, A. S. *et al.* De novo mutations in epileptic encephalopathies. *Nature* **501,** 217–21 (2013).
6. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science (80-.).* **350,** 1262–1266 (2015).
7. Jiang, Y.-H. *et al.* Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93,** 249–63 (2013).
8. Yuen, R. K. *et al.* Genome-wide characteristics of de novo mutations in autism. *Genomic Med.* **1,** 16027 (2016).
9. Yuen, R. K. C. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* 1–13 (2017). doi:10.1038/nn.4524
10. He, X. *et al.* Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* **9,** e1003671 (2013).
11. Jiang, Y. *et al.* Incorporating Functional Information in Tests of Excess De Novo

Mutational Load. *Am. J. Hum. Genet.* **97,** 272–283 (2015).

12. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74,** 285–299 (2012).
13. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485,** 242–245 (2012).
14. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **13,** 216–221 (2014).
15. Turner, T. N. *et al.* Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am. J. Hum. Genet.* **98,** 58–74 (2016).
16. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328,** 636–9 (2010).
17. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).
18. Campbell, C. D. *et al.* Estimating human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44,** 1277–1281 (2013).
19. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).
20. Li, J. *et al.* Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Mol. Psychiatry* **21,** 290–297 (2015).
21. Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347,** 1155–9 (2015).
22. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

23. Elements, D. N. A. An integrated encyclopedia of DNA elements in the human genome. (2012). doi:10.1038/nature11247

24. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).

25. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6,** e1001025 (2010).

26. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).

27. Xiong, H. Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347,** 1254806 (2015).

28. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).

29. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95,** 535–552 (2014).

30. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat. Genet.* **46,** 881–885 (2014).

31. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6,** e1001154 (2010).

32. Petrovski, S. *et al.* The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet.* **11,** e1005492 (2015).

33. Liu, L. *et al.* DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. *Mol. Autism* **5,** 22 (2014).

34. Liu, L., Lei, J. & Roeder, K. Network assisted analysis to reveal the genetic basis of

autism. *Ann. Appl. Stat.* **9,** 1571–1600 (2015).

35. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155,** 997–1007 (2013).

36. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* **9,** S4 (2008).

37. Krishnan, A. *et al.* Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* **19,** 1454–1462 (2016).

38. Robbins, T. W. & Roberts, A. C. Differential regulation of fronto-executive function by the monoamines and acetylcholine. *Cereb. Cortex* **17 Suppl 1,** i151–i160 (2007).

39. Lin, A., Rangel, A. & Adolphs, R. Impaired learning of social compared to monetary rewards in autism. *Front. Neurosci.* **6,** 1–7 (2012).

40. Sepeta, L. *et al.* Abnormal social reward processing in autism as indexed by pupillary responses to happy faces. *J. Neurodev. Disord.* **4,** 17 (2012).

41. Wang, K. S., Liu, X. F. & Aragam, N. A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. *Schizophr. Res.* **124,** 192–199 (2010).

42. Li, X. *et al.* ZMIZ1 preferably enhances the transcriptional activity of androgen receptor with short polyglutamine tract. *PLoS One* **6,** 1–12 (2011).

43. Wu, J. I. *et al.* Regulation of dendritic development by neuron-specific chromatin remodeling complexes. *Neuron* **56,** 94–108 (2007).

44. Vogel-Ciernia, A. & Wood, M. A. Neuron-specific chromatin remodeling: A missing link in epigenetic mechanisms underlying synaptic plasticity, memory, and intellectual disability disorders. *Neuropharmacology* **80,** 18–27 (2014).

45. Córdova-Fletes, C. *et al.* A de novo t(10;19)(q22.3;q13.33) leads to ZMIZ1/PRR12 reciprocal fusion transcripts in a girl with intellectual disability and neuropsychiatric alterations. *Neurogenetics* **16,** 287–298 (2015).

46. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

47. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S. & Balding, D. J. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49,** 986–992 (2017).

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

49. Akbarian, S. *et al.* The PsychENCODE project. *Nat. Neurosci.* **18,** 1707–1712 (2015).

50. Takeichi, M. The cadherin superfamily in neuronal connections and interactions. *Nat. Rev. Neurosci.* **8,** 11–20 (2007).

51. Smith, A., Bourdeau, I., Wang, J. & Bondy, C. A. Expression of Catenin family members CTNNA1, CTNNA2, CTNNB1 and JUP in the primate prefrontal cortex and hippocampus. *Mol. Brain Res.* **135,** 225–231 (2005).

52. Zeidán-Chuliá, F. *et al.* Altered expression of Alzheimer's disease-related genes in the cerebellum of autistic patients: a model for disrupted brain connectome and therapy. *Cell Death Dis.* **5,** e1250 (2014).

53. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159,** 1511–23 (2014).

54. Matheson, S. F. *et al.* Distinct but overlapping functions for the closely related p190 RhoGAPs in neural development. *Dev. Neurosci.* **28,** 538–550 (2006).

55. Rico, B. *et al.* Control of axonal branching and synapse formation by focal adhesion kinase. *Nat. Neurosci.* **7,** 1059–69 (2004).

56. Craig, A. M. & Kang, Y. Neurexin–neuroligin signaling in synapse development. *Curr. Opin. Neurobiol.* **17,** 43–52 (2007).

57. Bruno, D. L. *et al.* Molecular and Clinical Characterization of 25 Individuals With Exonic Deletions of NRXN1 and Comprehensive Review of the Literature. (2013). doi:10.1002/ajmg.b.32148

58. Gomez, A. M., Nguyen, T., Scheiffele, P. & Traunmuller, L. Control of neuronal synapse specification by a highly dedicated alternative splicing program. *Science (80-.).* **352,** 982–986 (2014).

59. Laht, P., Otsus, M., Remm, J. & Veske, A. B-plexins control microtubule dynamics and dendrite morphology of hippocampal neurons. *Exp. Cell Res.* **326,** 174–184 (2014).

60. Oinuma, I., Ito, Y., Katoh, H. & Negishi, M. Semaphorin 4D/Plexin-B1 stimulates PTEN activity through R-Ras GTPase-activating protein activity, inducing growth cone collapse in hippocampal neurons. *J. Biol. Chem.* **285,** 28200–28209 (2010).

61. Tasaka, G.-I., Negishi, M. & Oinuma, I. Semaphorin 4D/Plexin-B1-mediated M-Ras GAP activity regulates actin-based dendrite remodeling through Lamellipodin. *J. Neurosci.* **32,** 8293–8305 (2012).

62. Wolman, M. A., Regnery, A. M., Becker, T., Becker, C. G. & Halloran, M. C. Semaphorin3D regulates axon axon interactions by modulating levels of L1 cell adhesion molecule. *J. Neurosci.* **27,** 9653–9663 (2007).

63. Vodrazka, P. *et al.* The semaphorin 4D-plexin-B signalLing complex regulates dendritic and axonal complexity in developing neurons via diverse pathways. *Eur. J. Neurosci.* **30,**

1193–1208 (2009).

64. Gilabert-Juan, J. *et al.* Semaphorin and plexin gene expression is altered in the prefrontal cortex of schizophrenia patients with and without auditory hallucinations. *Psychiatry Res.* **229,** 850–857 (2015).

65. Chen, J., Yu, S., Fu, Y. & Li, X. Synaptic proteins and receptors defects in autism spectrum disorders SPECTRUM DISORDERS. **8,** 1–13 (2014).

66. DeNardo, L. A., de Wit, J., Otto-Hitt, S. & Ghosh, A. NGL-2 Regulates Input-Specific Synapse Development in CA1 Pyramidal Neurons. *Neuron* **76,** 762–775 (2012).

67. Xu, G. *et al.* NGL-2 Is a New Partner of PAR Complex in Axon Differentiation. *J. Neurosci.* **35,** 7153–7164 (2015).

68. Sangu, N. *et al.* A 7q31.33q32.1 microdeletion including LRRC4 and GRM8 is associated with severe intellectual disability and characteristics of autism. *Hum. Genome Var.* **4,** 17001 (2017).

69. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 novel candidate genes for intellectual disability. *Nat. Publ. Gr.* 5–10 (2016). doi:10.1101/052670

70. Matsunami, N. *et al.* Identification of rare DNA sequence variants in high-risk autism families and their prevalence in a large case/control population. *Mol. Autism* **5,** 5 (2014).

71. Glancy, M. *et al.* Transmitted duplication of 8p23.1–8p23.2 associated with speech delay, autism and learning difficulties. *Eur. J. Hum. Genet.* **17,** 37–43 (2009).

72. Rossignol, D. A. & Frye, R. E. Evidence linking oxidative stress, mitochondrial dysfunction, and inflammation in the brain of individuals with autism. *Front. Physiol.* **5 APR,** 1–15 (2014).

73. Ma, X. *et al.* A genome-wide association study for quantitative traits in schizophrenia in

China. *Genes, Brain Behav.* **10,** 734–739 (2011).

74. Ni, P. *et al*. Methionine sulfoxide reductase A (MsrA) associated with bipolar I disorder and executive functions in A Han Chinese population. *J. Affect. Disord.* **184,** 235–238 (2015).

75. Newton, M. A., Noueiry, A., Sarkar, D. & Ahlquist, P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5,** 155–176 (2004).

76. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

77. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc* **10,** 1556–1566 (2015).

78. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).

79. Xu, L. M. *et al.* AutismKB: An evidence-based knowledgebase of autism genetics. *Nucleic Acids Res.* **40,** 1–7 (2012).

80. Betancur, C. Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380,** 42–77 (2011).

81. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94,** 677–694 (2014).

82. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506,** 185–190 (2014).

83. Bayés, A. *et al*. Comparative study of human and mouse postsynaptic proteomes finds high compositional conservation and abundance differences for key synaptic proteins.

PLoS One **7,** (2012).

84. Darnell, J. C. *et al*. FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. *Cell* **146,** 247–261 (2011).

Figure Legends

Figure 1. Overview of TADA-A. The blue frame illustrates the inputs of the model, including mutation counts, baseline mutation rates and annotations (assumed to be binary). The orange frame shows an example of relative risk estimates of different noncoding annotations by TADA-A. The green frame illustrates our gene mapping strategy. For each gene, we derived its noncoding BF based on the relative risks of its noncoding mutations and calibrated mutations rates, which is then multiplied to the gene's coding BF to get a total BF.

Figure 2. Burdens analyses of different types of *de novo* mutations. The error bars in (a) and (b) represents the 95% confidence intervals of burdens (ORs), based on Fisher's exact tests. We did not show error bars in (c) as they are too large to show. (a) *De novo* nonsynonymous mutations from WGS studies in multiple gene sets. (b) Regulatory SNVs and distal regulatory SNVs (10 kb ~ 50 kb to TSSs) in different gene sets. (c) Burdens of SNVs that are predicted to affect splicing vs. SNVs not predicted to have such effects (control mutations). (d) Partition of *de novo* ASD risk into coding and non-coding mutations.

Figure 3. Predicted risk genes and enhancers of ASD. (a) GeneMania network analysis of the five new ASD genes. Red circles represent new ASD genes while grey circles represent known ones. Two genes are connected if their co-expression across multiple tissues reaches a threshold. Only connections between the two gene sets are shown. (b) A *de novo* SNV in the TSS-proximal regulatory region of *PLXNB1* disrupts a RFX2 binding motif. The gene also has one LoF and one missense mutation. (c) Distribution of the number of enhancers with recurrent (two or more) *de novo* SNVs from 10,000 simulations. The vertical red arrow marks the observed number of enhancers with recurrent *de novo* SNVs. (d) A distal enhancer (marked by a star) of ZMIZ1 with recurrent SNVs. Grey curves represent possible interactions between enhancers and promoters (correlated activities across multiple tissues). Note that the region contains two additional DNMs in other sequences.

Table 1. Mutation frequency (the number of mutations per subject) and the relative risks of different types of *de novo* mutations.

Mutation Class	Mutation frequency	Relative mutational exposure(%)	Relative risk	Variance of ASD liability explained (100%)
Mis3	0.292	12.5	4.70	0.83
Loss-of-function	0.0675	2.90	20.0	1.08
Regulatory SNV	1.67	71.5	1.84	0.57
Splicing SNV	0.305	13.1	3.52	0.54

Table 2. Mutational counts and evidence of five new ASD genes. In the evidence rows, Y means overlap with a gene set and N otherwise.

Gene name	*JUP*	*APBB1*	*NRXN1*	*ARHGAP5*	*PNPLA7*	Enrichment p-value
LoF	1	1	1	1	1	
Mis3	0	0	1	1	1	
Regulatory SNV	4	3	0	1	0	
Splicing SNV	0	0	1	1	1	
HI	Y	Y	Y	Y	N	**3.36×10^{-4}**
RVIS (%)	2.74	19.93	2.25	20.02	64.97	**0.010**
ExAC zscore (%)	30.62	14.25	3.32	9.69	76.73	0.087
FMRP targets	N	Y	Y	N	N	**0.017**
BrainSpan expression (%)	19.29	3.41	12.03	22.03	69.90	**0.015**
DAWN	5.00×10^{-4}	-	0.001	2.00×10^{-4}	-	**4.90×10^{-4}**

Lower RVIS and ExAC zscores percentiles correspond to higher constraint.
Lower BrainSpan percentiles correspond to higher brain expression.
Enrichment p-values were calculated by hypergeometric tests. In RVIS, ExAC zscore, Brainspan, we tested the enrichment of new ASD genes in genes in the lower quartiles. In DAWN analysis, we tested the enrichment of new ASD genes in genes with DAWN q-value < 0.05. The DAWN q-value for each gene in the table is the minimum of the q-values of that gene in two brain regions, mid-fetal prefrontal cortex (PFC) and infancy mediodorsal cerebellar cortex (MD-CBC).

Supplemental Data

Supplemental data include 2 figures and 7 tables.

Figure S1. Burden analysis of *de novo* SNVs in H3K27ac regions within 10 kb of TSSs (Regulatory SNVs) and burden analysis of *de novo* splicing SNVs.

Figure S2. TADs enriched for regulatory SNVs. The heatmaps represent the interaction strength between two genomic loci measured from Hi-C experiments.

Table S1. Summary of studies contributing DNMs.

Table S2. Calibration of mutation rates for each study.

Table S3. Estimates of the relative risks of noncoding annotations

Table S4. Mutational counts and evidence of 14 new ASD genes (FDR < 0.3).

Table S5. Functional relevance and literature support of new ASD genes.

Table S6. Enhancers with recurrent *de novo* SNVs. Genes were assigned to enhancers based on cross-tissue correlation between enhancer activity and gene expression from Roadmap Epigenomics (http://khuranalab.med.cornell.edu/roadmap_stringent_enhancers.txt).

Table S7. TAD-level analysis for SNVs. The table show the top two TADs based on Poisson test (comparing observed and expected number of mutations).

Figure 1

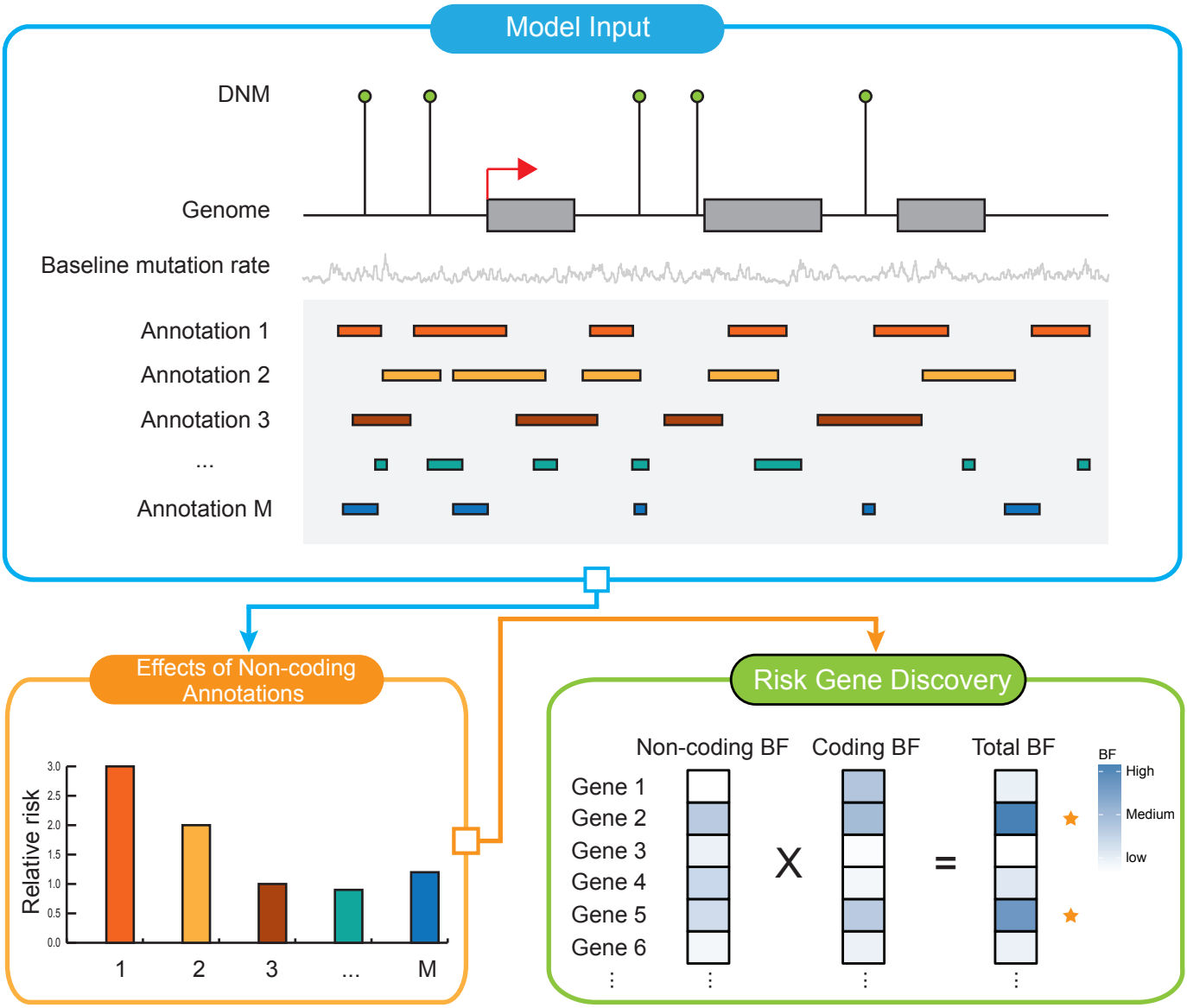

Model Input
DNM
Genome
Baseline mutation rate
Annotation 1
Annotation 2
Annotation 3
...
Annotation M
Effects of Non-coding Annotations
Relative risk
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
...
M
Risk Gene Discovery
Non-coding BF
Coding BF
Total BF
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
X
=
BF
High
Medium
low

Figure 2

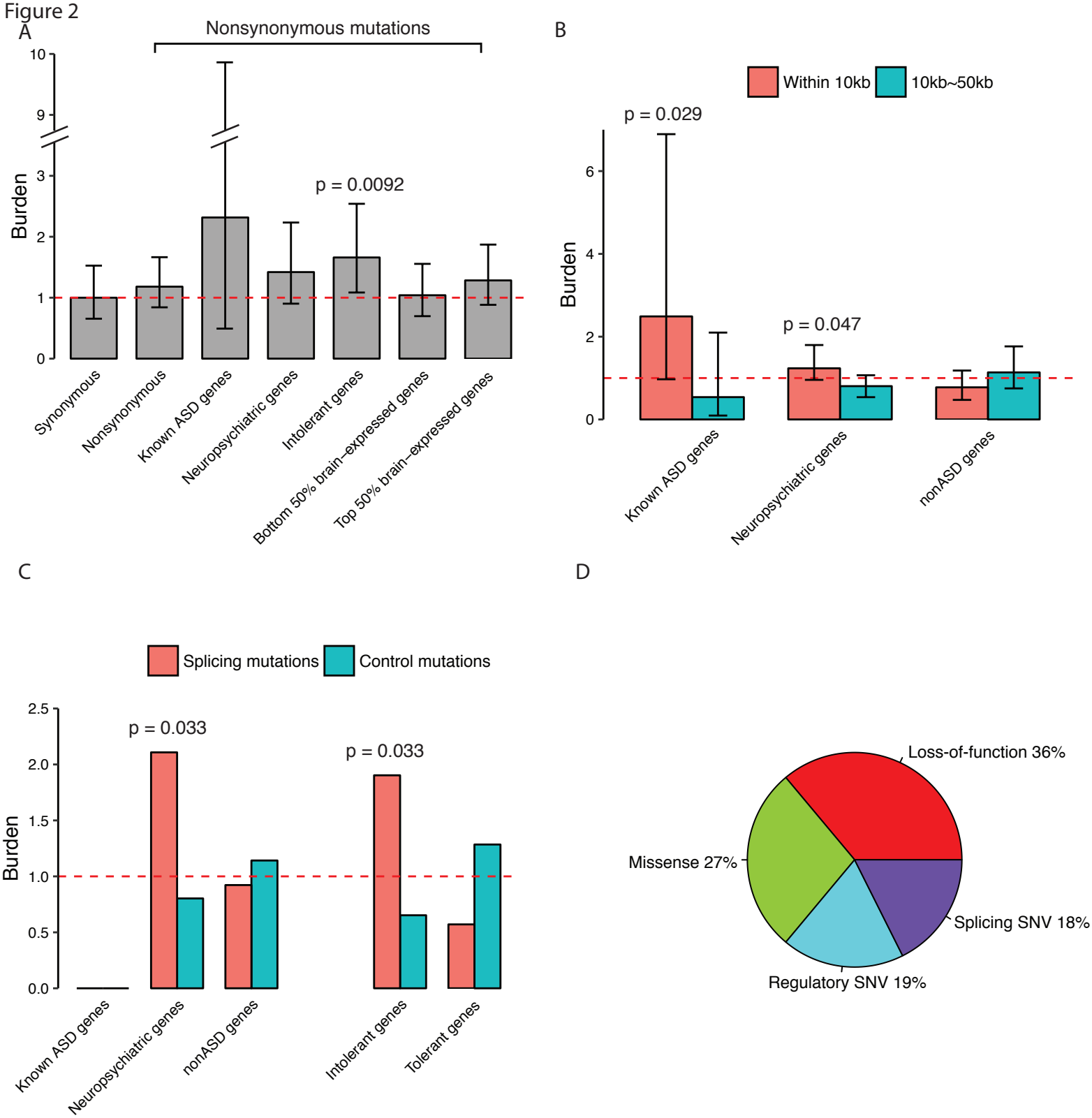
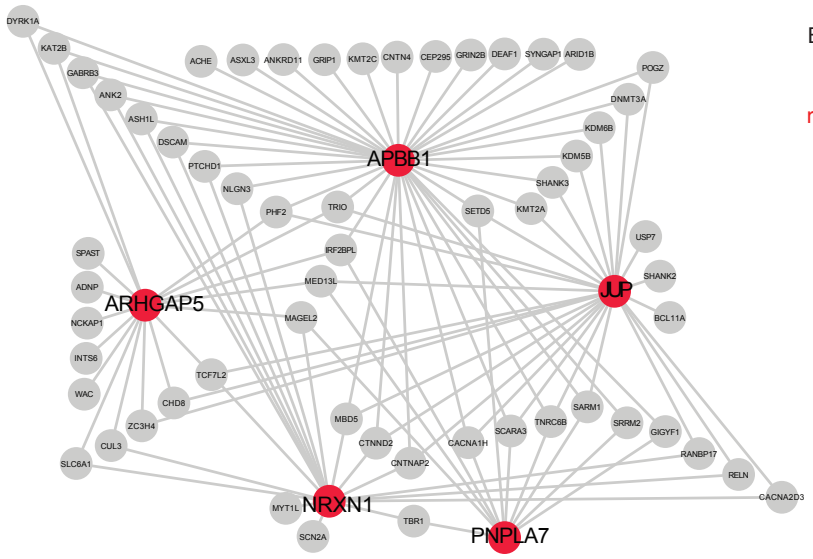
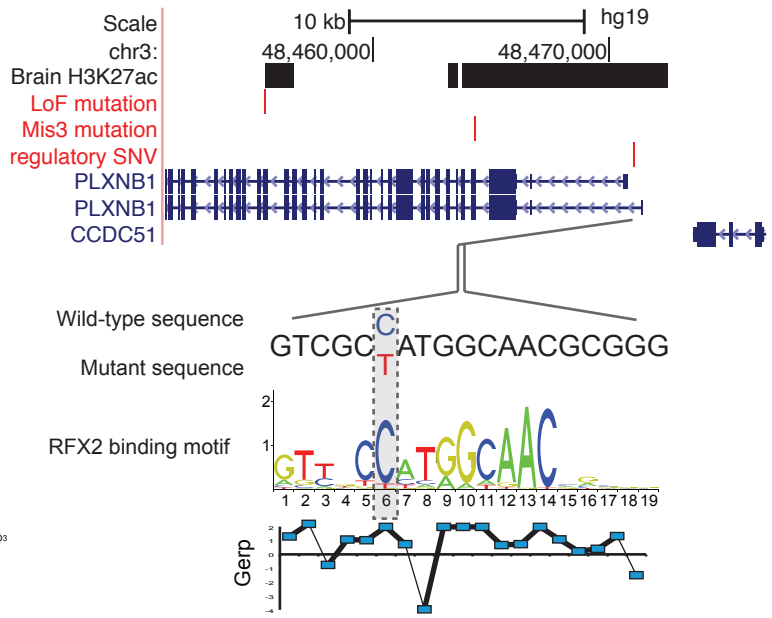

A
Nonsynonymous mutations
Burden
p = 0.0092
Synonymous
Nonsynonymous
Known ASD genes
Neuropsychiatric genes
Intolerant genes
Bottom 50% brain−expressed genes
Top 50% brain−expressed genes
B
Within 10kb
10kb~50kb
Burden
p = 0.029
p = 0.047
Known ASD genes
Neuropsychiatric genes
nonASD genes
C
Splicing mutations
Control mutations
Burden
p = 0.033
p = 0.033
Known ASD genes
Neuropsychiatric genes
nonASD genes
Intolerant genes
Tolerant genes
D
Loss-of-function 36%
Missense 27%
Splicing SNV 18%
Regulatory SNV 19%

Figure 3

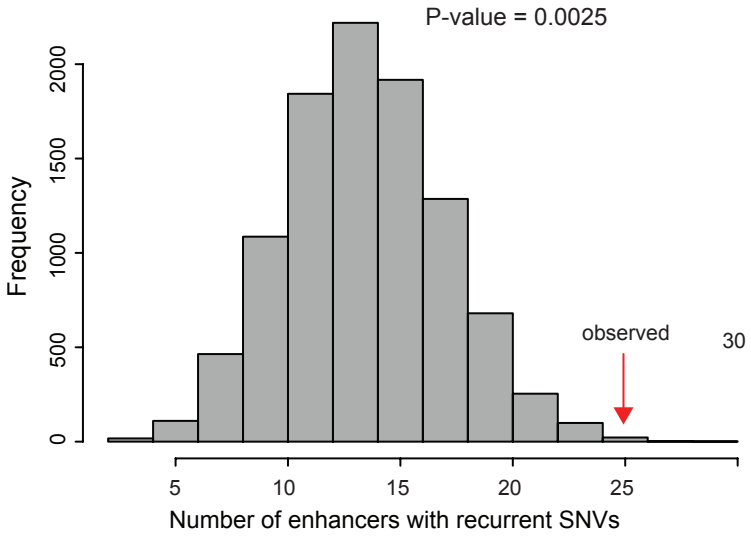
A



B



C



D

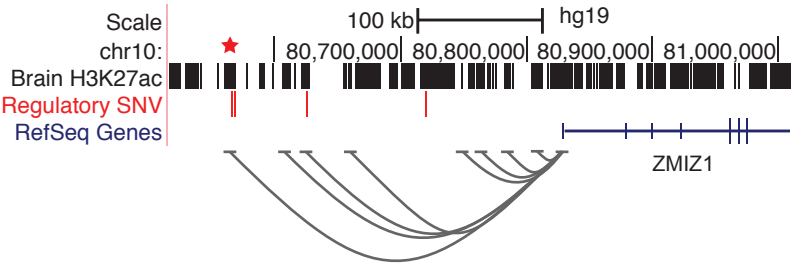


Figure S1

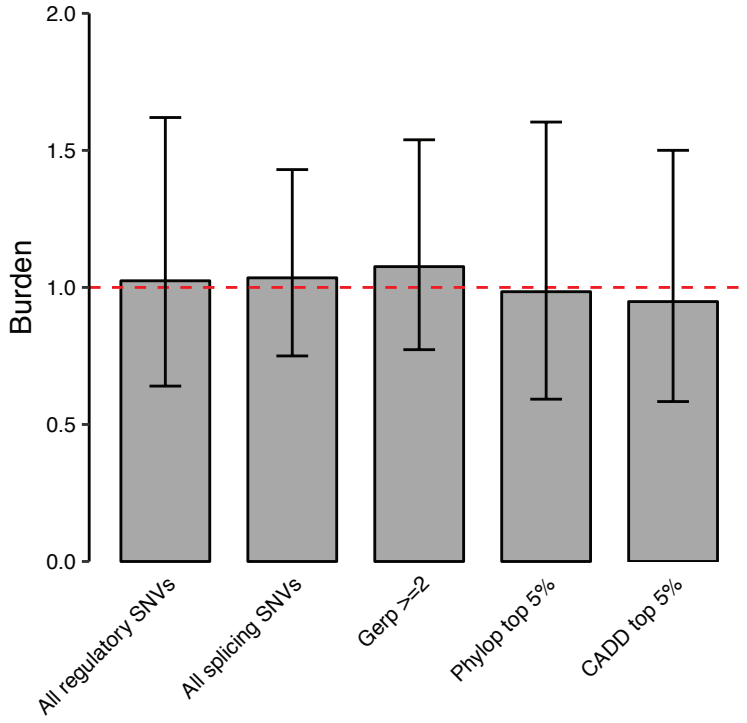
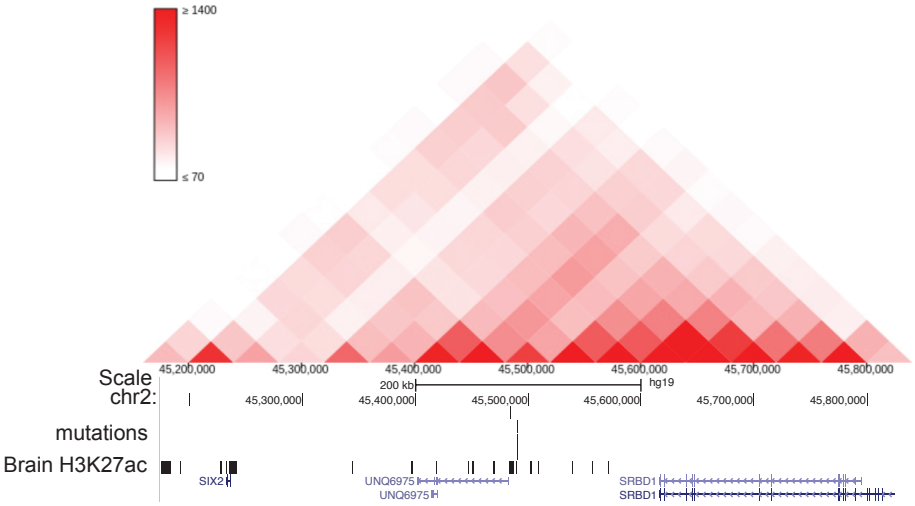
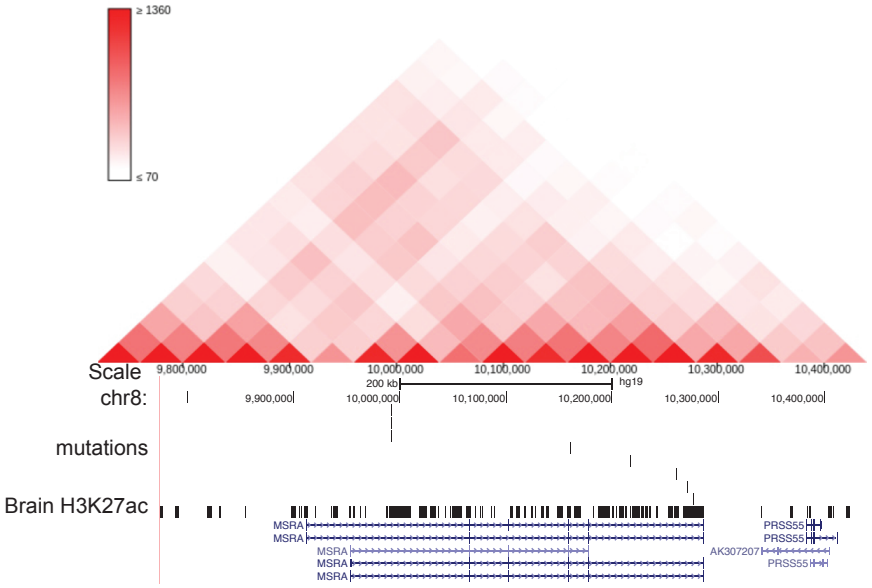

2.0
1.5
1.0
0.5
0.0
Burden
All regulatory SNVs
All splicing SNVs
Gerp >=2
Phylop top 5%
CADD top 5%

Figure S2

A


≥ 1400
≤ 70
45,200,000
45,300,000
45,400,000
45,500,000
45,600,000
45,700,000
45,800,000
Scale
chr2:
200 kb
hg19
mutations
Brain H3K27ac
SIX2
UNQ6975
UNQ6975
SRBD1
SRBD1

B


≥ 1360
≤ 70
9,800,000
9,900,000
10,000,000
10,100,000
10,200,000
10,300,000
10,400,000
Scale
chr8:
200 kb
hg19
mutations
Brain H3K27ac
MSRA
MSRA
MSRA
MSRA
MSRA
PRSS55
PRSS55
AK307207
PRSS55