Contribution of *de novo* non-coding mutations to autism and identification of risk genes from whole-genome sequencing of affected families

Yuwen Liu[1], Ercument Cicek[2,3], Yanyu Liang[3], Jinchen Li[4,6,10], Rebecca Muhle[7,8,9], Nicholas Knoblauch[5], Martina Krenzer[9], Yue Mei[4], Yan Wang[4], 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, USA. [2]Computer Engineering Department, Bilkent University, Ankara, Turkey. [3]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. [4]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing, China. [5]Committee on Genetics, Genomics and Systems Biology, The University of Chicago, Chicago, IL, USA. [6]Institute of Genomic Medicine, Wenzhou Medical University, Wenzhou, China. [7]Department of Genetics, Yale School of Medicine, New Haven, Connecticut, USA. [8]Child Study Center, Yale Medicine, New Haven, Connecticut, USA. [9]Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, Connecticut, USA. [10]The State Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan, China. [11]Department of Biostatistics, Columbia University, New York, New York, USA. *Correspondence should be addressed to S.Z. (sunzs@mail.biols.ac.cn) or H.X. (xinhe@uchicago.edu).

Abstract

Analyzing de novo mutations (DNMs) in protein-coding genes from whole-exome sequencing (WES) data has emerged as a powerful tool for mapping risk genes of autism spectrum disorder (ASD). The impact of non-coding mutations in ASD, however, has been largely unknown. This represents a large gap in our understanding of the genetics of ASD, as the majority of GWAS hits for a range of disorders fall into non-coding regions. To address this question, we performed a meta-analysis of DNMs using whole-genome sequencing (WGS) data from more than 300 individuals with ASD. We found that DNMs are enriched within brain transcriptional regulatory elements near genes involved in neuropsychiatric disorders. In these genes and in evolutionarily constrained genes, we also found an excess of DNMs that are predicted to affect pre-mRNA splicing. Collectively, we estimate that non-coding mutations explain at least one third of the ASD genetic risk attributable to DNMs. By combining information of non-coding DNMs with published WES data, we identified three new ASD risk genes at a false discovery rate (FDR) < 0.1, and 11 at a FDR < 0.3. A number of these genes are known to regulate critical processes in neural development and have been associated with other neuropsychiatric disorders. Taken together, our results demonstrate the pathogenic contribution of non-coding DNMs in ASD etiology and highlight some promising ASD risk genes. The analytic tools we provided in this study, for estimating contribution of non-coding mutations to disease risk and for mapping susceptibility genes using both coding and regulatory mutations, are applicable to any WGS studies on DNMs.

77 # Introduction

78

79 Autism spectrum disorder (ASD) is a neurodevelopmental disorder with onset in early childhood[1,2].
80 Understanding the genetics of ASD would help understand the etiology of the disorder and may
81 lead to the development of more effective diagnostics and therapies[3]. Nonetheless, ascribing ASD
82 genetic risk to specific genomic loci has proved challenging. GWAS have only yielded a handful
83 of small-effect variants that have not replicated between studies[4,5]. In contrast, screening for *de*
84 *novo* mutations (DNMs) with potentially larger effect sizes has emerged as an effective approach
85 for identifying ASD risk loci. Several large whole-exome sequencing (WES) studies have been
86 performed on ASD parent-child trios or quartets[6–9], based on the principle that genes harboring an
87 excess of deleterious DNMs are likely risk genes. These studies have identified approximately 70
88 ASD risk genes.

89

90 Compared with coding mutations, the contribution of noncoding mutations to ASD etiology is
91 largely unknown. For a wide range of diseases and traits, the majority of GWAS hits fall into
92 noncoding regions[10]. Non-coding variants may alter the activity of cis-regulatory elements or
93 affect the splicing of pre-mRNAs, which may in turn alter gene dosage and isoform usage[11–15].
94 Beyond implicating specific disease-associated loci, knowledge of non-coding variants can also
95 offer clues as to which cell types are relevant to disease etiology and how trans-acting genes may
96 contribute to disease risk[12,16]. In the case of autism, disruption of chromatin remodeling, histone
97 modification, and pre-mRNA splicing have been implicated in the disease[8,17–22]. Taken together,
98 these results support the role of transcriptional mis-regulation in ASD etiology. Nevertheless, with
99 few exceptions[23], most existing ASD studies focused on coding sequences, and little is known
100 about cis-regulatory variants predisposing autism risk.

101

102 In this study, we explore the contribution of non-coding mutations to ASD by combining data from
103 multiple whole-genome sequencing (WGS) studies in autism trios. We developed a computational
104 framework to analyze whole-genome DNM data (Fig. 1). To assess the extent non-coding
105 mutations contribute to autism risk, we compared the rates of DNMs within putative regulatory
106 sequences in ASD subjects vs. controls. We employed multiple annotations to identify variants
107 with possible regulatory effects, such as epigenomic modification patterns in brain and predicted

deleteriousness from computational methods[24–26]. Based on these burden analyses, we estimated the effect sizes of different types of regulatory mutations, measured as relative risks. Combining these estimates with the frequencies of mutations, we estimated the contribution of different types of mutations to the variation of autism risk (liability) across individuals. Next, we aimed to identify specific genes and regulatory elements underlying autism. We hypothesize that an ASD risk gene may be disrupted in multiple ways: by coding mutations changing protein function, or by regulatory mutations affecting gene expression level, or by splicing mutations. This motivates a strategy of combining coding and non-coding mutations within/near a gene to assess its role in autism. To optimally combine information from multiple mutational categories, we used a method we developed previously, TADA (*Transmission and de novo association*)[27]. TADA assesses the enrichment of DNMs in a gene relative to random expectation, weighing each mutation based on its likely effect size (Fig. 1).

Using this computational framework, we found modest enrichment of *de novo* single nucleotide variations (SNVs) in individuals with ASD in brain-active enhancers near genes with likely roles in psychiatric disorders, as well as in positions affecting RNA splicing. 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). Although the sample size of WGS is only a tenth of that of WES, we still identified three new ASD risk genes at a FDR < 0.1 and 11 at a FDR < 0.3. Multiple lines of evidence support the possible roles of these genes in ASD.

Results

<u>WGS data of ASD and control families</u>

We combined DNM data from five WGS studies of ASD trios/quartets, including four published[28–31] and one unpublished (Wu et al. See Methods for details), with a total of 300 affected subjects (Supplementary Table 1). As controls, we used recently published whole-genome DNM data from a large cohort of 700 unaffected trios[32]. 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 six studies.

139 Largely in line with published mutation rates[29,33,34], the number of DNMs per subject in ASD
140 studies ranges from 57 to 63. The rate of DNMs in controls is lower (39), which could be due to
141 differences in sequencing depth, SNV calling procedures, or other factors. In subsequent analysis
142 of a particular mutation type (e.g. mutations within enhancers), we accounted for the differences
143 in the total mutation rates between cases and controls by comparing the relative mutation rates
144 with neutral mutations (e.g. synonymous mutations) serving as background (see Methods). Despite
145 the differences in absolute mutation rates, the mutational patterns are similar between cases and
146 controls (see Section 2.1-2.3 of Supplementary Methods, also Supplementary Figs. 1, 2, and 3).
147
148 We were able to replicate the findings of earlier WES studies using the coding mutations. We
149 found that nonsynonymous mutations from WGS data are about 1.2 fold enriched in individuals
150 with ASD versus controls (Fig. 2a), similar to previous numbers[6,35,36]. The burden (i.e. fold of
151 enrichment) is higher for nonsynonymous mutations within known ASD genes, genes likely
152 involved in neuropsychiatric disorders (dubbed "neuropsychiatric genes"), and genes intolerant to
153 mutations (Fig. 2a, See Methods for gene-set definitions). The burden of highly expressed brain
154 genes (from Brainspan: http://www.brainspan.org) is also higher than genes with lower expression
155 (Fig. 2a).
156
157 Enrichment of functional *de novo* non-coding mutations in children with ASD
158
159 To establish the importance of non-coding sequences in autism, we analyzed DNMs in putative
160 regulatory sequences of the brain. Enriched DNMs in these regions comparing with random
161 expectation would suggest that at least some of these DNMs are causative mutations. We used
162 fetal and adult brain H3K27ac regions as markers of regulatory activity[37,38]. We limited our
163 analysis to H3K27ac regions within 10 kb of transcription start sites (TSSs) of protein-coding
164 genes, including promoters and 5'UTR (promoters are defined as sequences within 1 kb upstream
165 of TSSs). We found 1.1 fold enrichment of *de novo* SNVs in promoters with H3K27ac marks
166 (active promoters), compared to de novo synonymous mutations though the results are not
167 statistically significant (Fig. 2b). In contrast, we found no enrichment of *de novo* SNVs in
168 promoter/5'UTR regions without H3K27ac marks (Fig. 2b). We also saw no enrichment of *de*
169 *novo* SNVs across all H3K27ac regions within 10 kb of TSSs (referred to as regulatory SNVs

170 hereafter) (Fig. 2b). Noncoding variants that disrupt or create transcription factor binding motifs
171 have been frequently associated with regulatory effects and complex phenotypes[39,40]. Indeed, we
172 found a significant ASD burden for motif-changing SNVs (see Methods for definition) in
173 H3K27ac enhancers ($P = 0.00019$, one-sided Fisher's exact tests were used in all burden analyses).
174 Notably, the commonly used annotations for non-coding variants, including GERP++[25], PhyloP[41]
175 and CADD[24] and Eigen[42], were not significantly associated with mutational burden (Fig. 2b and
176 Supplementary Table 2).
177
178 We hypothesized that regulatory sequences near ASD risk genes harbor excess noncoding
179 mutations, as suggested by an earlier study[23]. We found a 1.8-fold enrichment of regulatory SNVs
180 near ASD genes ($P = 0.069$) and a 1.2-fold enrichment near neuropsychiatric genes ($P = 0.032$)
181 (Fig. 2c). In contrast, we observed no enrichment of regulatory SNVs near genes that are unlikely
182 to be ASD risk genes (nonASD genes). Additionally, we found that SNVs falling into more distal
183 H3K27ac regions (10 kb - 50 kb to TSSs) are not enriched in known ASD or neuropsychiatric
184 genes (Fig. 2c).
185
186 Because of the concern of potential batch effects between ASD cases and controls, we performed
187 simulation-based alternative burden analyses, without using control data. The procedure is based
188 on reshuffling of DNMs in cases across enhancers according to estimated mutation rates. We found
189 that the burdens of motif-changing SNVs and of regulatory SNVs near neuropsychiatric genes are
190 similar to what we found in Figs. 2b and 2c ($P = 0.025$ and 0.0017, respectively, Supplementary
191 Table 3).
192
193 To show that the enrichment of regulatory SNVs is due to regulatory functions of H3K27ac
194 sequences, rather than to some generic properties of enhancers (such as higher GC content[43]), we
195 compared the burdens of SNVs in enhancers with strong activities in the developing cortex
196 (denoted as "brain-specific enhancers") vs. enhancers with strong activities in other tissues ("non-
197 brain specific enhancers"). The numbers of such enhancers are small, so we extended the distance
198 cutoff to 1 Mb around TSSs. Within brain-specific enhancers, the burdens of regulatory SNVs near
199 known ASD genes, neuropsychiatric genes, and intolerant genes are all higher than baseline

200 (though not significant because of the small mutational counts, Fig. 2d). In contrast, the
201 corresponding burdens are all below 1 for enhancers specifically active in other tissues (Fig. 2d).
202
203 Non-coding mutations may also become pathogenic by affecting pre-mRNA splicing. Using the
204 predictions from a recent paper[26], we found significant burdens of SNVs likely affecting RNA
205 splicing (splicing SNVs) in neuropsychiatric genes (Fig. 2e, $P = 0.011$) and in genes intolerant to
206 mutations (Fig. 2e, $P = 0.018$). In contrast, the SNVs that are not predicted to affect splicing show
207 no such patterns ("non-splicing" mutations in Fig. 2e).
208
209 In summary, we found multiple lines of evidence supporting the roles of regulatory and splicing
210 mutations in autism (see Supplementary Table 4 for detailed results of all burden analyses). The
211 specificity of our findings (e.g. we found burden in brain-specific enhancers but not enhancers
212 specific to other tissues) and the robustness of our results to different ways of performing burden
213 analysis argue that the results are due to true biological signals, rather than batch effects from
214 heterogeneous samples. We summarized the evidence supporting this claim in Section 2 of
215 Supplementary Methods.
216
217 Partition of ASD risk into coding and non-coding DNMs
218
219 We next assessed the relative contribution of non-coding versus coding DNMs to ASD liability.
220 We considered four classes of mutations: loss-of-function (nonsense and frameshift indels in
221 coding sequences), probably damaging missense (predicted by PolyPhen-2, denoted as Mis3),
222 regulatory SNVs, and splicing SNVs. The contribution of any type of mutation is measured as
223 Liability Variance Explained (LVE), taking into account both the mutational target sizes and the
224 average relative risks of mutations (see Methods). Note that our relative risks are defined on
225 mutations of causative genes. We used a strategy previously developed to translate results of
226 previous burden analysis into average relative risks of causal genes.[27]. For coding mutations (LoF
227 and mis3), we used values estimated previously[8]. For regulatory and splicing SNVs, we used the
228 burden results from ASD genes with TADA q-value < 0.5[35] (see Methods and Section 3 of
229 Supplementary Methods for details and possible caveats).
230

231 Our relative risk estimates of regulatory SNVs and splicing SNVs are significantly lower than
232 those of coding SNVs, but regulatory SNVs are much more frequent (Table 1). Each class of
233 mutation explains only a small fraction of estimated total ASD genetic risk (Table 1), which is
234 broadly consistent with the conclusion of an earlier study[44]. Considering only the risk due to *de*
235 *novo* mutations, we found that non-coding SNVs (including regulatory SNVs and splicing SNVs)
236 explain 31% of the *de novo* risk (Fig. 2f). We argue that this estimate is very conservative . First,
237 we considered only enhancers within 10 kb of TSSs, which constitute about 36% of all enhancers
238 in our data. Second, our dataset contains only regulatory sequences active very early in
239 development (5, 7, and 12 post-conception week) or in the adult brain. Many enhancers active in
240 other related developmental periods are not covered in our data. Third, we only considered SNVs
241 in this study. *De novo* indels and CNVs, which are likely to be much more deleterious, were
242 ignored.
243
244 Identification of ASD risk genes using both coding and noncoding mutations
245
246 We performed integrative risk gene mapping using both coding and non-coding mutations. The
247 challenge posed by combining different types of mutations is that these mutations could have very
248 different damaging effects (e.g., nonsense mutations are generally more deleterious than other
249 mutations). Summarizing a gene's mutational burden in a manner agnostic to each mutation's
250 likely consequence would compromise the power to detect risk genes. We adopted TADA[27], to
251 address this issue. Briefly, TADA assesses the evidence of risk on each gene by comparing the
252 observed number of mutations in each mutational class (e.g. LoF or missense) with the expected
253 number. The evidence of each class is weighted by the average relative risk of that class, and the
254 evidence of all classes is combined into a single Bayes factor for the gene, reflecting the strength
255 of evidence (Fig. 1). In earlier studies, we have shown that TADA is much more powerful for
256 mapping risk genes than simple counting of mutation numbers in genes[8,35]. In this study, we
257 considered four classes of DNMs: LoF and mis3 coding mutations, regulatory SNVs within 10 kb
258 of TSS and splicing SNVs. For each class of mutations, we derived its average relative risk from
259 mutational burden analysis (Table 1), and estimated the mutation rate of each class for each gene
260 using a genome-wide mutation rate model[45]. Once we computed the results of all genes from
261 TADA, we corrected for multiple testing using Bayesian FDR[46].

262

263 TADA analysis with only coding mutations from a previous WES study (~3,500 samples)
264 identified 58 ASD risk genes at FDR < 0.1[35]. Adding regulatory SNVs and splicing SNVs from
265 our WGS data added three new ASD risk genes at FDR < 0.1, and 11 new genes at FDR < 0.3.
266 These predictions are robust to TADA parameters within a reasonable range (Supplementary Table
267 5). Each of the three genes at FDR < 0.1 (Table 2) has at least one LoF or Mis3 mutation, and the
268 evidence for these genes is strengthened by the presence of regulatory SNVs. We found extensive
269 evidence supporting the plausibility of these genes as ASD risk genes. *APBB1* is the target of
270 neuronal-RNA binding protein FMRP, whose loss of function causes Fragile X syndrome and
271 autistic features. All of the three genes have been identified as haploinsufficient genes[47,48] (Table
272 2, hypergeometric test, P = 0.00079) and are highly expressed in the brain (top 25% of all genes,
273 P = 0.015). The genes also tend to be evolutionarily constrained using either RIVS or another
274 metric based on frequency of LoF variants in ExAC (Table 2). Evolutionary constraint in the
275 human population has been shown to be a strong predictor of autism genes[45].
276

277 We performed multiple network analyses to further establish the link of new candidate genes to
278 autism. DAWN[49,50] is a recently developed method that predicts autism risk genes by virtue of the
279 genes' association with known ASD genes in co-expression networks of early developing brain.
280 We found two out of three new genes (*JUP* and *ARHGAP5*) have DAWN q-value < 0.05 in at
281 least one of the two critical developmental spatial-temporal windows for ASD[51] (Table 2,
282 enrichment P = 0.00041). Using GeneMania[52], we found additional evidence supporting the
283 connections between high-confidence ASD genes and our candidate genes. Between the two gene
284 sets, there are 75 co-expression links in GeneMania (Fig. 3a), significantly higher than chance
285 expectation (P = 0.014). To better understand the functions and biological processes that the
286 candidate genes are involved in, we obtained the top 30 neighboring genes of each candidate using
287 GeneMania (Supplementary Table 6). Gene Ontology (GO) enrichment analysis of the top 30
288 neighbors of the whole set of new ASD genes identified multiple biological processes related to
289 neuro-development, such as "negative regulation of neurogenesis" and "forebrain development"
290 (Table 3).
291

292 We next expanded the analysis to the 11 new genes at FDR < 0.3. We observed significant
293 enrichment of multiple gene annotations (Supplementary Table 7), including haploinsufficient
294 genes (P = 0.014), FMRP targets (P = 0.010), Gene Ontology "Brain" (P = 0.024), constrained
295 genes (P = 0.0038 using RVIS and P = 0.025 using variant frequency in ExAC), high expression
296 level in the early developing brain from BrainSpan (P =0.011) and genes significantly co-
297 expressed with known ASD genes from DAWN analysis (P = 0.00013). The pathogenic roles of
298 these genes in ASD are further supported by previous studies (Supplementary Table 8).
299
300 We next explore the mechanisms by which non-coding mutations may contribute to ASD. We
301 found an interesting example in *PLXNB1* (a new risk gene with q = 0.22). This gene has important
302 functions in neurodevelopment (see Discussions). In the promoter region of *PLXNB1*, an SNV
303 disrupts the binding motif of the transcription factor RFX2 (Fig. 3b). Notably, in the motif-
304 matching region, multiple bases including the one hit by the DNM, are highly conserved based on
305 GERP scores (Fig. 3b). RFX2 is a transcription factor highly expressed in the tegmental region of
306 the the fetal mid-brain, a region linked to cognitive abilities[53] and social reward processing[54,55].
307 An intronic variant in *RFX2* has been associated with schizophrenia and bipolar disorder[56].
308 Additionally, the RFX2 binding motif is enriched in regulatory regions of several co-expression
309 gene modules (Modules 1, 5, 7, and 15)[37] that are active in the early developing brain. Taken
310 together, our analyses offer a testable hypothesis: the mutation in *PLXNB1* promoter changes the
311 affinity of RFX2, and as a result, perturbs the expression of *PLXNB1* and hence increases the risk
312 of ASD.
313
314 <u>Enhancers with recurrent *de novo* mutations</u>
315
316 Our previous analyses were performed at the gene level, and only considered enhancers within 10
317 kb of TSSs. We expanded these analyses to identify specific H3K27 enhancers with roles in autism
318 across the genome, based on the notion that multiple DNMs in a single enhancer are unlikely to
319 occur by chance. We found 25 enhancers with two or more SNVs in ASD cases, significantly
320 higher than random expectation (Fig. 3c, P = 0.0025). Since the expected number of recurrent
321 enhancers by chance is about 12, we estimated that half of these enhancers are likely to be causal
322 enhancers (Supplementary Table 9). Many recurrent enhancers are distant from genes, so the

323 nearest genes may not represent their true targets. We used predicted enhancer-gene pairs based
324 on cross-tissue correlation between enhancer activity and gene expression from Roadmap
325 Epigenomics (http://khuranalab.med.cornell.edu/roadmap_stringent_enhancers.txt). We found a
326 recurrent enhancer targeting *ZMIZ1*, a gene missed in previous analysis as it is more than 250 kb
327 away from the enhancer (Fig. 3d). The region contains two other DNMs in two enhancers, one of
328 which also has correlated activities with the ZIMZ1 promoter. A target of FMRP, *ZMIZ1* is highly
329 expressed in the brain and its protein product interacts with neuron-specific chromatin remodeling
330 complex (nBAF) which is important in regulating synapatic functions[57,58]. Several nBAF members
331 have been linked to autism, such as ARID1B and BCL11A[59]. The pathogenic potential of *ZMIZ1*
332 is further implicated by the observation of a *de novo* gene-disrupting translocation in an individual
333 with intellectual disability[60]. These evidence together strongly support the role of ZMIZ1 in autism,
334 and also highlight the mechanism that DNMs may increase ASD risk by disrupting distal
335 regulatory elements.
336
337 We next performed recurrent DNM analysis at the level of Topologically associated domains
338 (TADs)[61]. These are megabase–sized chromatin interaction domains that are stable across cell
339 types and have been proposed to demarcate transcriptional regulatory units[61]. We reason that if we
340 identify TADs with excess regulatory SNVs, it may be straightforward to assign target genes of
341 these regulatory SNVs. Based on estimated mutation rates, we found two TADs with a significant
342 (at FDR < 0.125) number of regulatory SNVs (Supplementary Fig. 5 and Supplementary Table
343 10). In both TADs, there are only two or three genes, and we conjecture that SRBD1 and MRSA
344 are likely the underlying ASD genes in the two TADs, respectively (see Discussion).
345
346 <u>Power of mapping ASD risk genes with WGS and WES</u>
347
348 We used simulations to address how the power of a DNM-focused WES or WGS study depends
349 on its sample size and sequencing budget, using our insights of the *de novo* genetic architecture of
350 ASD (Table 1). We randomly chose 1000 genes as risk genes based on previous estimates of the
351 total number of ASD risk genes[8,62]; randomly sampled mutations according to mutation rates and
352 their likely relative risks (causal genes tend to have more deleterious mutations compared to
353 expectations) and then applied TADA to identify risk genes at q < 0.1 (see Section 6 of

354 Supplementary Methods for simulation details). We found that the power of our simulated WES
355 studies is roughly in line with empirical discoveries (Fig. 4a). For example, at N = 2,500 trios, we
356 would expect to identify about 30-40 ASD genes, similar to the actual number (33) from a previous
357 study[8].The power of the simulated WGS design is about 20% higher than that of the WES (Fig.
358 4a). We next investigated whether the additional power gained from WGS is justifiable on the
359 basis of cost. At the current per-sample cost level (WES: $500 and WGS: $1000), we found that
360 WES is still more cost-effective (Fig. 4b).
361
362
363 **Discussion**
364
365 Our analysis provides clear evidence that non-coding mutations play an important role in the
366 etiology of autism. We found elevated rates of *de novo* SNVs in ASD subjects within brain
367 enhancers that are likely to affect transcription factor binding. Near potential risk genes of ASD
368 and other neuropsychiatric disorders, there is an excess of ASD-derived DNMs in enhancer regions.
369 We also found enrichment of splicing-affecting SNVs in ASD and neuropsychiatric genes as well
370 as in constrained genes. Compared with coding mutations, we found that these non-coding
371 mutations account for a smaller but substantial proportion of *de novo* ASD risk. Using a statistical
372 framework, TADA, we were able to leverage both coding and regulatory mutations for gene
373 mapping and identified several new ASD genes. Our study on non-coding mutations also led to
374 the characterization of ASD risk units with excess DNMs at the level of enhancers and TAD
375 regions. Altogether, our findings highlight the importance of regulatory variants in ASD and their
376 utility in mapping risk genes and functional elements. The analytic framework for *de novo*
377 mutations we developed in this work is applicable to any trio-based sequencing studies.
378
379 Our work is one of the first studies that directly assess the role of non-coding regulatory variants
380 in complex diseases[23]. Previously, our knowledge of non-coding variants came mostly from
381 GWAS. The challenge with GWAS is that regulatory elements are generally much shorter (~1kb)
382 than regions of linkage disequilibrium (LD, hundreds of kb on average). It is thus far from
383 straightforward to assess the quantitative contribution of non-coding variants, or identify specific
384 regulatory elements underlying diseases from GWAS[63]. In contrast, using DNMs from WGS, we

385 eliminate the impact of LD and work directly on putative causal mutations. The results of our
386 findings thus can be interpreted more straightforwardly. For instance, the enhancers with recurrent
387 DNMs we identified are good candidates for further experimental analysis.
388
389 Distinguishing likely pathogenic variants from benign ones is essential when analyzing non-coding
390 variants. We only observed a modest enrichment of DNMs in H3K27ac enhancers near putative
391 risk genes, but not across all genes. This suggests that most SNVs within H3K27ac sites are
392 probably not causative mutations. We tested various methods for annotating variants, including
393 GERP++[25], PhyloP[41], CADD[24], and Eigen[42], but none was found to enrich pathogenic mutations.
394 Two possible explanations are: (1) Cross-species evolutionary constraint may be only modestly
395 correlated with pathogenicity in humans[64]. (2) Two of these methods, CADD and Eigen, were
396 trained on epigenomic data from mostly non-brain tissues. To better analyze the non-coding
397 genome, we need a more comprehensive repertoire of regulatory information in brain, within
398 specific spatial-temporal contexts[65].
399
400 A recent study estimated the contributions of common variants, rare inherited variants and DNMs
401 to the risk of autism, and concluded that most of the autism risk is explained by common variants[44].
402 Our goal here is not to explain ASD risk *per se*, but use DNMs to explore the relative contributions
403 of coding and non-coding variants. While risk partition among different types of variants has been
404 pursued in GWAS[63], linkage disequilibrium limits our ability to do this, as discussed previously.
405 Our work thus provides independent evidence that non-coding variants make a substantial
406 contribution to the risk of complex diseases[63]. Comparing with the previous study of autism, our
407 estimate of ASD risk attributable to coding mutations is somewhat higher (1.9% vs. 1.1%) and in
408 particular we found a significant contribution from missense DNMs (0.83% vs. 0.04%). We
409 believe that this higher estimate is largely due to our different modeling assumptions: we treated
410 all mutations in a category (e.g., missense) as a mixture of causal and non-causal mutations,
411 whereas the previous study treated all mutations in a category equally (see Methods)[44]. We
412 estimated that *de* novo coding (1.9%), non-coding (0.9%) and copy number variants (1.46%,
413 estimated by a previous study[44]) together contribute at least 4.2% of ASD risk. We think that we
414 significantly under-estimated the contribution of non-coding mutations to ASD risk as we did not

take into account mutations in more distal regulatory regions or in ASD-relevant enhancers not covered by our datasets, and indels or CNVs disrupting regulatory sequences.

We have identified three new ASD risk genes whose expression patterns and biological functions are consistent with their putative roles in ASD risk. JUP is a member of the catenin/cadherin superfamily which has important roles in neuron connections and interactions[66]. It is strongly expressed in the primate prefrontal cortex and hippocampus[67]. A*PBB1* is an adaptor protein localized in the nucleus and has a role in Alzheimer's disease. It is down-regulated in ASD cerebellum than control[68], and its microexons are mis-regulated in the brains of ASD individuals[20]. ARHGAP5 negatively regulates RHO GTPase activity. It is required for normal neural development, and is implicated in axonal branching and synapse formation[69,70].

The pathogenicity of most of the new ASD genes at FDR < 0.3 is supported by functional and association studies (summarized in Supplementary Table 8). For example, PLXNB1 is a receptor in the GTPase signaling pathway that regulates the development of hippocampal neurons[71,72], remodeling of dendrites[73], and the plasticity of axons[74,75]. *PLXNB1* expression has been found to be altered in the prefrontal cortex of schizophrenia patients[76]. Another example is DPYSL2, a multifunctional adaptor protein within the central nervous system[77]. DPYSL2 interacts with various binding partners to regulate multiple biological processes critical in neuron development, including axon initiation and elongation[77,78], neurite outgrowth[79], neuronal differentiation [80], and neurotransmitter release[81]. Functional variants in *DPYSL2* have been found to increase the risk of schizophrenia[82,83]. Only two of our new ASD genes (*MSL2* and *PPM1D*) have not been functionally implicated in neuro-development. However, *PPM1D* has recently been identified as a novel risk gene for intellectual disability[84].

We also found that the two TAD regions with excess regulatory SNVs in ASD are supported by copy number variant (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[85]. In a later independent study, CNVs in this TAD region were found to be enriched in ASD cases versus controls[85]. These results suggest that *SRBD1* is likely the risk gene in this TAD. In the other TAD region, ASD-associated duplication of 8p23.1-

446 8p23.2 introduces a breakpoint between MSRA and RP1L1[86]. MSRA is a member of the
447 methionine-sulfoxide reductase system whose function is to alleviate oxidative stress. Increased
448 exposure to oxidative stress plays an important role in the pathogenesis of ASD[87]. In addition,
449 GWAS studies have established associations of MSRA with schizophrenia[88] and bipolar disorder[89].
450 All these evidence together suggests that MSRA is a strong candidate gene for autism.
451
452 We investigated the power of WES and WGS studies using simulation. We note that the power of
453 WGS is significantly underestimated for the same reasons, described previously, that the
454 contribution of non-coding mutations to ASD risk is under-estimated. Additionally, WGS would
455 enable much better identification of structural variants, which collectively make a significant
456 contribution to the autism risk. Finally, we did not account for the cost of recruiting patients, which
457 would further favor choosing WGS over WES, as WGS requires a smaller cohort to achieve the
458 same level of power.
459
460 In summary, our work established genome-wide DNM analysis as a powerful strategy for mapping
461 genes and regulatory elements underlying autism. By leveraging multiple types of mutations
462 perturbing the activity and expression of genes, we gain considerable power over using protein-
463 coding variants alone. Furthermore, comparing with protein-coding genes, regulatory elements are
464 typically active in specific cell types and brain regions, thus the knowledge of autism-related
465 regulatory elements may help us identify important neuronal subtypes and circuits relevant to
466 autism, a crucial step for understanding its mechanisms[3]. With lowering cost of genome
467 sequencing and better annotations of regulatory genomes in the brain[65], we anticipate that whole-
468 genome DNM studies will transform our understanding of neuropsychiatric disorders.
469
470
471 **Methods**
472
473 <u>DNMs from whole genome sequencing data</u>
474
475 The detailed information for each WGS dataset is summarized in Table S1. To remove erroneously
476 called *de novo* SNVs, we excluded 8 individuals with more than 140 (2 times more than the median

477 of ASD DNMs per individual) DNMs and removed all recurrent DNMs (i.e. exactly the same
478 mutation in multiple individuals). Our unpublished (Wu et al.) DNM data are from WGS of 32
479 ASD trios of Han Chinese ancestry (see Section 1 of Supplementary Methods for details, data are
480 available at http://wwwdev.ebi.ac.uk/eva/?eva-study=PRJEB14713).
481
482 Collection of brain cis-regulatory regions
483
484 We used H3K27ac sites in fetal and adult brains to define cis-regulatory regions in our main
485 analyses. Fetal brain sites from human cortex at embryonic stages 5 p.c.w., 7 p.c.w., and 12 p.c.w.
486 were obtained from a recent study[37]. For each stage, only peak regions consistent between two
487 biological replicates were selected. Adult brain sites were obtained from Roadmap Epigenomics
488 Project[38]. They include H3K27ac sites from human angular gyrus, anterior caudate, cingulate
489 gyrus, middle hippocampus, inferior temporal lobe, mid-frontal lobe and substantia nigra. We used
490 MACS2 to call peaks from raw data, and only kept peak regions consistent between two biological
491 replicates for each brain region. We used BEDtools[90] to merge H3K27ac sites from fetal and adult
492 brain.
493
494 To define brain-specific enhancers in Fig. 2d, we compared H3K27ac data from human embryonic
495 cortex with publicly available H3K27ac datasets for the colon, the esophagus and limbs[37,91]. For
496 each H3K27ac dataset, only replicated peaks were included and any peak with 1-bp overlap to a
497 Gencode(v19) annotation feature was excluded. A single composite multi-sample enhancer
498 annotation was generated by merging replicated peaks across all tissue samples. The level of
499 H3K27ac signal in each region for each sample was quantified by averaging read counts per
500 kilobase per million mapped reads (RPKM) in each region from each replicate. Each region was
501 represented by a vector of a length equal to the total number of tissues considered, with each
502 element representing the RPKM value of marking in that region for a single tissue. Each vector
503 was standardized to have zero-mean and unit-variance. The matrix of these normalized tissue
504 quantification values was then subjected to k-means clustering to identify sets of sites exhibiting
505 the strongest marking in each tissue relative to all other samples in the comparison.
506
507 Annotation of regulatory mutations

508

509 We annotated DNMs within regulatory regions using various bioinformatics tools. To limit our
510 analysis to motifs that are most likely to be ASD-relevant, we used JASPAR core vertebrate TF
511 motifs[92] that meet two criteria: (1) they are enriched in regulatory sequences of gene modules that
512 are adjacent to regulatory sequences active in developing cortex of human, but not other primates[37];
513 and (2) their cognate transcription factors are in the top 50% brain-expressed genes (Expression
514 levels based on mean expression levels across Brainspan samples). To identify motif-changing
515 SNVs, for each SNV, we generated a 41-bp wild-type sequence centering on the reference allele,
516 along with a 41-bp mutant sequence centering on the mutated allele. We then used FIMO[93] to
517 obtain its maximal motif score difference (log(ratio of p-values)) in absolute value between its
518 wild-type and mutant sequences. A SNV whose maximal absolute motif score difference is in the
519 top 10% among the difference of all SNVs was defined as a motif SNV. We used ANNOVAR to
520 annotate coding mutation types and obtain GERP++ scores for all mutations[94]. CADD scores were
521 downloaded from http://cadd.gs.washington.edu/score, (default parameters, v1.3). Other
522 annotations that we used include PhyloP scores from the 100-way PhyloP wiggle file from UCSC
523 (/goldenpath/hg19/phyloP100way/) and Eigen-PC scores[42].
524

525 <u>Lists of genes used in burden analyses</u>

526

527 Known ASD genes include genes with q-value < 0.3 from a combined analysis of CNVs, indels,
528 and WES data using TADA[35], SFARI category I (high confidence) and II (strong candidate) genes
529 (https://sfari.org), AutismKB genes[95] and ASD genes summarized in a previous study[96].
530 Neuropsychiatric genes are a larger set of genes likely involved in neuropsychiatric disorders,
531 including genes with TADA q-value < 0.5[35] , SFARI genes (https://sfari.org), AutismKB genes[95],
532 ASD risk genes summarized in a previous study[96], intellectual disability genes[97], the union of gene
533 sets enriched with SCZ *de novo* coding mutations[98] and high confidence postsynaptic density
534 genes[99]. The set of nonASD genes are the 1000 genes with the highest TADA q-values[35]. The size
535 of this gene set is comparable to that of neuropsychiatric genes. Intolerant genes include genes
536 with top 5% RVIS[64] and haploinsuffcient genes obtained from two sources, one using copy number
537 variations (genes with predicted haploinsufficient probability greater than 0.95)[47], and the other
538 using estimated mutation rate[48]. To define tolerant gene, we started with genes with RVIS scores

539 in the bottom 10%[64], genes with haploinsufficient probability smaller than 0.1[47], and genes that
540 were used as control genes for LoF deficient genes[48]. We then removed any genes that were in the
541 intolerant gene set. To define gene groups based on their expression levels, we used the average
542 expression level for each gene across all developing brain tissues from BrainSpan data.
543
544 Burden analysis of different types of *de novo* mutations
545
546 In burden analysis, we accounted for the difference in mutation rate between ASD subjects
547 (~60/individual) and controls (~39/individual) as follows. Instead of comparing absolute mutation
548 rates, we normalized the mutation counts in the regions of interest by the mutation counts of some
549 background sequences that should not be enriched with risk mutations. This relative mutational
550 frequency is thus comparable between ASD and controls.
551
552 Specifically, for the burden analysis of coding sequences, we first calculated the ratio of
553 nonsynonymous mutations and synonymous mutations for both ASD and control samples. This
554 ratio can be thought of as the relative frequency of nonsynonymous mutations in cases or controls.
555 The burden of nonsynonymous mutations is then the ratio between the relative frequency of
556 nonsynonymous mutation of ASD subjects and that of control subjects:
557
558 Coding burden = $\frac{ASD\ \ nonsynonymous\ SNVs/ASD\ synonymous\ SNVs}{control\ nonsynonymous\ SNVs/control\ synonymous\ SNVs}$
559
560 The p-value of the burden was calculated from Fisher's exact test.
561
562 We defined splicing SNVs as SNVs whose delta-splicing scores predicted by SPIDEX[26] are in the
563 lower 10% of all the SNVs in our studies. At this cutoff, enriched splicing SNVs near known ASD
564 genes have been observed previously[26]. We defined non-spicing mutations as SNVs whose delta-
565 splicing scores are in the middle 10%. The burden of splicing SNVs was calculated in a manner
566 similar to coding mutation, using synonymous SNVs as background mutations.
567

568 For the burden of regulatory SNVs, we followed the same strategy but used all regulatory SNVs
569 as background instead of synonymous mutations. The burden of regulatory SNVs with a certain
570 annotation (or in a certain group of genes) is expressed as:
571
572 Regulatory SNV burden = $\frac{ASD\ regulatory\ SNVs\ with\ a\ certain\ annotation/all\ ASD\ regulatory\ SNVs}{control\ regulaotrySNVs\ with\ a\ certain\ annotation/all\ control\ regulatory\ SNVs}$
573
574 Burden analysis by simulation
575
576 To estimate the burden of ASD regulatory SNVs in a gene set, we sampled the number of
577 regulatory ASD SNVs of a gene using a multinomial distribution, with the total number given by
578 the observed number of ASD regulatory SNVs. The multinomial probability of a gene is
579 proportional to the total mutation rate of H3k27ac regulatory regions within 10 kb of the TSS of
580 that gene. We performed simulations 10,000 times and generated the null distribution of the total
581 number of mutations that land in the gene set of interest. We then used this distribution to derive
582 the empirical enrichment p-value and fold of enrichment (observed mutation number in the gene
583 set divided by the mean of simulation distribution). To estimate the burden of motif-changing
584 SNVs, we first shuffled the positions of ASD regulatory SNVs within all H3K27ac regulatory
585 regions (under the same mutation-rate-based multinomial distribution used above), and then, for
586 each new position, generated a mutant allele based on the mutational spectrum of all ASD SNVs.
587 We counted the number of motif-changing SNVs from each of 10000 simulations, and generated
588 a null distribution of this number, which is used to obtain enrichment p-values of the observed
589 number of motif-changing SNVs.
590
591 Mutation rates of regulatory sequences
592
593 We applied the tri-nucleotide mutation model developed by a previous study[45] to the whole-
594 genome to derive the base-level mutation rate. The mutation rate for each functional unit (e.g., one
595 enhancer) is defined as the sum of the per-base mutation rates[27] in the unit. To approximate the
596 mutation rate of splicing SNVs for each gene, we counted the number of splicing SNVs across
597 genes in controls, and used the ratio between this number and the total number of synonymous

598 mutations in controls as a scaling factor. We then multiplied this scaling factor to the synonymous
599 mutation rate of each gene to estimate its splicing SNV mutation rate.
600
601 Estimating the relative risks of de novo mutations
602
603 We used a strategy developed previously to estimate the average relative risk of a given type of
604 mutation (section 6 in the Supporting Text of TADA). It is also described in the user guide of the
605 software implementation of TADA[27]. Note that the relative risk of a type of mutation (e.g. missense)
606 is interpreted as the average relative risk of a missense mutation in a causal gene. Let λ be the
607 burden estimated for the mutational types of interest across all genes. Because causal genes
608 represent only a small proportion of all genes, the relative risks in causal genes should be higher
609 than λ. Let π be the fraction of genes that are risk genes, the relative risk of causal genes can be
610 calculated as $\frac{\lambda-1}{\pi}+1$. We used $\pi = 0.06$ in our analysis of coding mutations, which corresponds
611 to there being 1000 ASD risk genes[9,27]. For regulatory SNVs and splicing SNVs, the genome-
612 wide burdens of these two types of mutations are not significant, so we instead limited our analysis
613 to previously indicated ASD genes with TADA q-value <0.5[35]. We then converted the burden in
614 these genes to relative risks with $\pi = 0.5$ using the same equation. When estimating the relative
615 risk of splicing SNVs, we removed SNVs that were predicted to be nonsynonymous mutations or
616 that fell into the brain H3K27ac regions within 10kb of TSSs.
617
618 Partitions of *de novo* ASD risk into coding and non-coding mutations
619
620 We treated ASD liability (risk) as a continuous trait, and estimated the percentage of variance in
621 ASD liability explained by four types of mutations (Fig. 2f). The variation of ASD liability
622 explained by the *j-th* type of mutations could be expressed as the equation below (see
623 Supplementary Methods):
624
625 $$V_j = \beta_j^2 * p_j * (1 - p_j)$$
626

627 where β_j is the effect size of the *j-th* type of mutations at the liability scale and p_j is the probability
628 that an individual carries a mutation of type *j*. Note that only causal mutations contribute to ASD
629 liability, so both β_j and p_j are defined for mutations affecting causal genes. We calculated β_j from
630 the relative risk of j-th type of mutation using standard quantitative genetic calculations. To obtain
631 p_j, we calculate the total mutation rate of type j mutations (note that an individual has two copies
632 of genomes), and then multiply this by 0.06 (fraction of ASD risk genes) to obtain the rate of causal
633 mutations of type j (see Section 4 of Supplementary Methods for details).
634
635 TADA analysis to map ASD risk genes
636
637 A detailed description of TADA can be found in Section 5 of Supplementary Methods and in
638 previous publications[8,27]. To combine evidence of the coding and non-coding mutations for a
639 particular gene, we multiplied the pre-computed coding-mutation Bayes factors from published
640 work[35] by the non-coding Bayes factors derived from our study. If a gene does not have brain
641 H3K27ac regions within 10 kb of its TSS, we assigned its non-coding Bayes factor to be 1. We
642 translated the combined Bayes factor of each gene into a q-value using Bayesian FDR control.
643
644 To define new ASD genes at a particular q-value cutoff, we chose the genes with a combined q-
645 value below the cutoff and a coding q-value above the cutoff. We then filtered out new ASD genes
646 that do not have H3K27ac regions within 10kb of TSSs. Note that even if a gene has no evidence
647 from non-coding mutations, its q-value could change, because the q-value depends on the
648 distribution of Bayes factors of all genes. We also lost a few genes at each q-value cutoff: These
649 genes have no regulatory SNVs and thus receive a small penalty to their Bayes factors.
650
651 DAWN analysis
652
653 DAWN (Detecting Association With Networks) algorithm[49,50] is a guilt-by-association-based gene
654 prediction algorithm. The fundamental assumption underlying the algorithm is that risk genes are
655 part of a functional cluster and thus, tightly connected. A gene has a high posterior risk probability
656 if it is interacts in a network with other risk genes or has a high prior risk probability or both. We
657 used TADA to assign each gene a p-value for the prior evidence of being an ASD risk gene. For

658 the underlying network, we constructed partial co-expression networks for two spatio-temporal
659 windows, namely the mid-fetal prefrontal cortex (PFC) and the infancy mediodorsal cerebellar
660 cortex (MD-CBC), which are indicated as high risk windows for ASD[51]. BrainSpan microarray
661 dataset is used as the source for spatio-temporal gene expression data
662 (http://developinghumanbrain.org).
663
664 DAWN was run separately for each above-mentioned network. We used regularization parameter
665 (lambda) = 0.12, p-value cutoff = 0.1 and correlation thresholds 0.7 for PFC and 0.85 for MD-
666 CBC, respectively. In Table 2, posterior risk scores (q-values) are shown for the candidate genes.
667 A Dash means that the corresponding gene is not co-expressed with other risk genes in any of the
668 spatio-temporal windows.
669
670 GeneMania analysis
671
672 GeneMania[52] is a tool for studying interactions among genes in a network using various types of
673 information, such as gene co-expression and protein-protein interactions (PPIs). We first studied
674 the connection between our candidate genes with high-confidence ASD genes (genes with coding
675 TADA FDR < 0.1 and genes in SFARI categories I and II, https://sfari.org) using co-expression
676 data. We then chose 30 top neighboring genes for each of our new candidate genes using multiple
677 networks, including co-localization, genetic interactions, shared pathway, physical interactions,
678 predicted interactions and shared protein domains.
679
680 Enhancers with recurrent *de novo* mutations
681
682 For this analysis, we used brain H3K27ac regions not overlapping with exons. We observed 25
683 enhancers with at least two *de novo* SNVs in ASD samples. We performed simulations to assess
684 its signfiicance. In each simulation, we randomly re-distributed *de novo* SNVs of all brain
685 enhancers, following a multinomial distribution. The multinomial probability of an enhancer is the
686 ratio between the mutation rate of that enhancer and the sum of mutation rates across all enhancers.
687 We performed simulations 10,000 times and obtained the distribution of the number of enhancers
688 with recurrent SNVs.

TADs with recurrent *de novo* SNVs

For each TAD region, we calculated its 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[27,45]. We then calculated the Poisson *p*-value of each TAD region and used the Benjamini-Hochberg procedure to control FDR.

Acknowledgements

We thank all members of Dr. Xin He's lab for their help. We thank members of Autism Sequencing Consortium for providing helpful comments. This work was partially supported by a SFARI (Simons Foundation Autism Research Initiative) award from Simons Foundation (award ID: 385027).

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; 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, tissue-specific enhancers and contributed ideas to the burden analysis and TAD analysis; All authors contributed to the writing of the manuscript.

Competing financial interests

The authors declare no competing financial interests.

720 **References**

721

722 1. Folstein, S. E. & Rosen-Sheidley, B. Genetics of austim: complex aetiology for a
723 heterogeneous disorder. *Nat. Rev. Genet.* **2,** 943–955 (2001).
724 2. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental*
725 *Disorders*. *Arlington* (2013). doi:10.1176/appi.books.9780890425596.744053
726 3. Sahin, M. & Sur, M. Genes, circuits, and precision therapies for autism and related
727 neurodevelopmental disorders. *Science* **350,** aab3897–aab3897 (2015).
728 4. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum
729 disorders. *Nature* **459,** 528–533 (2009).
730 5. Anney, R. *et al.* Individual common variants exert weak effects on the risk for autism
731 spectrum disorders. *Hum. Mol. Genet.* **21,** 4781–4792 (2012).
732 6. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron*
733 **74,** 285–299 (2012).
734 7. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum
735 disorders. *Nature* **485,** 242–245 (2012).
736 8. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism.
737 *Nature* **515,** 209–215 (2014).
738 9. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum
739 disorder. *Nature* **13,** 216–221 (2014).
740 10. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait
741 associations. *Nucleic Acids Res.* **42,** 1001–1006 (2014).
742 11. Ma, Q. *et al.* A role for non-coding mutations in schizophrenia. **27,** 14299–14307 (2013).
743 12. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional
744 connections with IRX3. *Nature* **507,** 371–375 (2014).
745 13. Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma.
746 *Science* **339,** 957–9 (2013).
747 14. Flanagan, S. E. *et al.* Next-generation sequencing reveals deep intronic cryptic ABCC8
748 and HADH splicing founder mutations causing hyperinsulinism by pseudoexon activation.
749 *Am. J. Hum. Genet.* **92,** 131–136 (2013).
750 15. Chen, W. *et al.* Intronic deletions of tva receptor gene decrease the susceptibility to

751 infection by avian sarcoma and leukosis virus subgroup A. *Sci. Rep.* **5,** 9900 (2015).
752 16. Claussnitzer, M. *et al.* FTO obesity variant circuitry and adipocyte browning in humans.
753 *N. Engl. J. Med.* **373,** 895–907 (2015).
754 17. Krumm, N., O'Roak, B. J., Shendure, J. & Eichler, E. E. A de novo convergence of autism
755 genetics and molecular neuroscience. *Trends Neurosci.* **37,** 95–105 (2014).
756 18. Zafeiriou, D. I., Ververi, A. & Vargiami, E. Childhood autism and associated
757 comorbidities. *Brain Dev.* **29,** 257–272 (2007).
758 19. Didiot, M. C. *et al.* The G-quartet containing FMRP binding site in FMR1 mRNA is a
759 potent exonic splicing enhancer. *Nucleic Acids Res.* **36,** 4902–4912 (2008).
760 20. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in
761 autistic brains. *Cell* **159,** 1511–23 (2014).
762 21. Sadakata, T. *et al.* Autistic-like phenotypes in Cadps2-knockout mice and aberrant
763 CADPS2 splicing in autistic patients. *J. Clin. Invest.* **117,** 931–943 (2007).
764 22. Talebizadeh, Z. *et al.* Novel splice isoforms for NLGN3 and NLGN4 with possible
765 implications in autism. *J. Med. Genet.* **43,** e21 (2006).
766 23. Turner, T. N. *et al.* Genome Sequencing of Autism-Affected Families Reveals Disruption
767 of Putative Noncoding Regulatory DNA. *Am. J. Hum. Genet.* **98,** 58–74 (2016).
768 24. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human
769 genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
770 25. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under
771 selective constraint using GERP++. *PLoS Comput. Biol.* **6,** e1001025 (2010).
772 26. Xiong, H. Y. *et al.* RNA splicing. The human splicing code reveals new insights into the
773 genetic determinants of disease. *Science* **347,** 1254806 (2015).
774 27. He, X. *et al.* Integrated model of de novo and inherited genetic variants yields greater
775 power to identify risk genes. *PLoS Genet.* **9,** e1003671 (2013).
776 28. Jiang, Y.-H. *et al.* Detection of clinically relevant genetic variants in autism spectrum
777 disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93,** 249–63 (2013).
778 29. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease
779 risk. *Nature* **488,** 471–475 (2012).
780 30. Michaelson, J. J. *et al.* Whole-genome sequencing in autism identifies hot spots for de
781 novo germline mutation. *Cell* **151,** 1431–1442 (2012).

782 31. Yuen, R. K. C. *et al.* Whole-genome sequencing of quartet families with autism spectrum
783 disorder. *Nat. Med.* **21,** 185–191 (2015).
784 32. Wong, W. S. W. *et al.* New observations on maternal age effect on germline de novo
785 mutations. *Nat. Commun.* **7,** (2016).
786 33. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome
787 sequencing. *Science* **328,** 636–9 (2010).
788 34. Campbell, C. D. *et al.* Estimating human mutation rate using autozygosity in a founder
789 population. *Nat. Genet.* **44,** 1277–1281 (2013).
790 35. Sanders, S. J. *et al.* Insights into autism spectrum disorder genomic architecture and
791 biology from 71 risk loci. *Neuron* **87,** 1215–1233 (2015).
792 36. Li, J. *et al.* Genes with de novo mutations are shared by four neuropsychiatric disorders
793 discovered from NPdenovo database. *Mol. Psychiatry* **21,** 290–297 (2015).
794 37. Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and
795 enhancer activity during human corticogenesis. *Science* **347,** 1155–9 (2015).
796 38. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,**
797 317–330 (2015).
798 39. Maurano, M. T. *et al.* Large-scale identification of sequence variants influencing human
799 transcription factor occupancy in vivo. *Nat. Genet.* **47,** 1393–1401 (2015).
800 40. Pai, A. A., Pritchard, J. K. & Gilad, Y. The genetic and mechanistic basis for variation in
801 gene regulation. *PLoS Genet.* **11,** e1004857 (2015).
802 41. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral
803 substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).
804 42. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating
805 functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48,** 214–
806 220 (2016).
807 43. Akhtar, M. M., Scala, G., Cocozza, S., Miele, G. & Monticelli, A. CpG islands under
808 selective pressure are enriched with H3K4me3, H3K27ac and H3K36me3 histone
809 modifications. *BMC Evol. Biol.* **13,** 145 (2013).
810 44. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat. Genet.*
811 **46,** 881–885 (2014).
812 45. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human

813 disease. *Nat. Genet.* **46,** 944–950 (2014).
814 46. Newton, M. A., Noueiry, A., Sarkar, D. & Ahlquist, P. Detecting differential gene
815 expression with a semiparametric hierarchical mixture method. *Biostatistics* **5,** 155–176
816 (2004).
817 47. Huang, N., Lee, I., Marcotte, E. M. & Hurles, M. E. Characterising and predicting
818 haploinsufficiency in the human genome. *PLoS Genet.* **6,** e1001154 (2010).
819 48. Petrovski, S. *et al.* The intolerance of regulatory sequence to genetic variation predicts
820 gene dosage sensitivity. *PLoS Genet.* **11,** e1005492 (2015).
821 49. Liu, L. *et al.* DAWN: a framework to identify autism genes and subnetworks using gene
822 expression and genetics. *Mol. Autism* **5,** 22 (2014).
823 50. Liu, L., Lei, J. & Roeder, K. Network assisted analysis to reveal the genetic basis of
824 autism. *Ann. Appl. Stat.* **9,** 1571–1600 (2015).
825 51. Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical
826 projection neurons in the pathogenesis of autism. *Cell* **155,** 997–1007 (2013).
827 52. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-
828 time multiple association network integration algorithm for predicting gene function.
829 *Genome Biol.* **9,** S4 (2008).
830 53. Robbins, T. W. & Roberts, A. C. Differential regulation of fronto-executive function by
831 the monoamines and acetylcholine. *Cereb. Cortex* **17 Suppl 1,** i151–i160 (2007).
832 54. Lin, A., Rangel, A. & Adolphs, R. Impaired learning of social compared to monetary
833 rewards in autism. *Front. Neurosci.* **6,** 1–7 (2012).
834 55. Sepeta, L. *et al.* Abnormal social reward processing in autism as indexed by pupillary
835 responses to happy faces. *J. Neurodev. Disord.* **4,** 17 (2012).
836 56. Wang, K. S., Liu, X. F. & Aragam, N. A genome-wide meta-analysis identifies novel loci
837 associated with schizophrenia and bipolar disorder. *Schizophr. Res.* **124,** 192–199 (2010).
838 57. Li, X. *et al.* ZMIZ1 preferably enhances the transcriptional activity of androgen receptor
839 with short polyglutamine tract. *PLoS One* **6,** 1–12 (2011).
840 58. Wu, J. I. *et al.* Regulation of dendritic development by neuron-specific chromatin
841 remodeling complexes. *Neuron* **56,** 94–108 (2007).
842 59. Vogel-Ciernia, A. & Wood, M. A. Neuron-specific chromatin remodeling: A missing link
843 in epigenetic mechanisms underlying synaptic plasticity, memory, and intellectual

844 disability disorders. *Neuropharmacology* **80,** 18–27 (2014).
845 60. Córdova-Fletes, C. *et al.* A de novo t(10;19)(q22.3;q13.33) leads to ZMIZ1/PRR12
846 reciprocal fusion transcripts in a girl with intellectual disability and neuropsychiatric
847 alterations. *Neurogenetics* **16,** 287–298 (2015).
848 61. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of
849 chromatin interactions. *Nature* **485,** 376–380 (2012).
850 62. Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the
851 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70,** 863–
852 885 (2011).
853 63. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-
854 wide association summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).
855 64. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance
856 to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9,**
857 e1003709 (2013).
858 65. Akbarian, S. *et al.* The PsychENCODE project. *Nat. Neurosci.* **18,** 1707–1712 (2015).
859 66. Takeichi, M. The cadherin superfamily in neuronal connections and interactions. *Nat. Rev.*
860 *Neurosci.* **8,** 11–20 (2007).
861 67. Smith, A., Bourdeau, I., Wang, J. & Bondy, C. A. Expression of Catenin family members
862 CTNNA1, CTNNA2, CTNNB1 and JUP in the primate prefrontal cortex and
863 hippocampus. *Mol. Brain Res.* **135,** 225–231 (2005).
864 68. Zeidán-Chuliá, F. *et al.* Altered expression of Alzheimer's disease-related genes in the
865 cerebellum of autistic patients: a model for disrupted brain connectome and therapy. *Cell*
866 *Death Dis.* **5,** e1250 (2014).
867 69. Matheson, S. F. *et al.* Distinct but overlapping functions for the closely related p190
868 RhoGAPs in neural development. *Dev. Neurosci.* **28,** 538–550 (2006).
869 70. Rico, B. *et al.* Control of axonal branching and synapse formation by focal adhesion
870 kinase. *Nat. Neurosci.* **7,** 1059–69 (2004).
871 71. Laht, P., Otsus, M., Remm, J. & Veske, A. B-plexins control microtubule dynamics and
872 dendrite morphology of hippocampal neurons. *Exp. Cell Res.* **326,** 174–184 (2014).
873 72. Oinuma, I., Ito, Y., Katoh, H. & Negishi, M. Semaphorin 4D/Plexin-B1 stimulates PTEN
874 activity through R-Ras GTPase-activating protein activity, inducing growth cone collapse

875 in hippocampal neurons. *J. Biol. Chem.* **285,** 28200–28209 (2010).
876 73. Tasaka, G.-I., Negishi, M. & Oinuma, I. Semaphorin 4D/Plexin-B1-mediated M-Ras GAP
877 activity regulates actin-based dendrite remodeling through Lamellipodin. *J. Neurosci.* **32,**
878 8293–8305 (2012).
879 74. Wolman, M. A., Regnery, A. M., Becker, T., Becker, C. G. & Halloran, M. C.
880 Semaphorin3D regulates axon axon interactions by modulating levels of L1 cell adhesion
881 molecule. *J. Neurosci.* **27,** 9653–9663 (2007).
882 75. Vodrazka, P. *et al.* The semaphorin 4D-plexin-B signalLing complex regulates dendritic
883 and axonal complexity in developing neurons via diverse pathways. *Eur. J. Neurosci.* **30,**
884 1193–1208 (2009).
885 76. Gilabert-Juan, J. *et al.* Semaphorin and plexin gene expression is altered in the prefrontal
886 cortex of schizophrenia patients with and without auditory hallucinations. *Psychiatry Res.*
887 **229,** 850–857 (2015).
888 77. Hensley, K., Venkova, K., Christov, A., Gunning, W. & Park, J. Collapsin response
889 mediator protein-2: An emerging pathologic feature and therapeutic target for
890 neurodisease indications. *Mol. Neurobiol.* **43,** 180–191 (2011).
891 78. Cole, A. R. *et al.* GSK-3 phosphorylation of the Alzheimer epitope within collapsin
892 response mediator proteins regulates axon elongation in primary neurons. *J. Biol. Chem.*
893 **279,** 50176–50180 (2004).
894 79. Quach, T. T. *et al.* Involvement of collapsin response mediator proteins in the neurite
895 extension induced by neurotrophins in dorsal root ganglion neurons. *Mol. Cell. Neurosci.*
896 **25,** 433–443 (2004).
897 80. Patrakitkomjorn, S. *et al.* Neurofibromatosis type 1 (NF1) tumor suppressor,
898 neurofibromin, regulates the neuronal differentiation of PC12 cells via its associating
899 protein, CRMP-2. *J. Biol. Chem.* **283,** 9399–9413 (2008).
900 81. Brittain, J. M. *et al.* An atypical role for collapsin response mediator protein 2 (CRMP-2)
901 in neurotransmitter release via interaction with presynaptic voltage-gated calcium
902 channels. *J. Biol. Chem.* **284,** 31375–31390 (2009).
903 82. Liu, Y. *et al.* Functional variants in DPYSL2 sequence increase risk of schizophrenia and
904 suggest a link to mTOR signaling. *G3 (Bethesda).* **5,** 61–72 (2014).
905 83. Lee, H. *et al.* Changes in Dpysl2 expression are associated with prenatally stressed rat

906 offspring and susceptibility to schizophrenia in humans. *Int. J. Mol. Med.* **35,** 1574–1586
907 (2015).
908 84. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 novel candidate
909 genes for intellectual disability. *Nat. Publ. Gr.* 5–10 (2016). doi:10.1101/052670
910 85. Matsunami, N. *et al.* Identification of rare DNA sequence variants in high-risk autism
911 families and their prevalence in a large case/control population. *Mol. Autism* **5,** 5 (2014).
912 86. Glancy, M. *et al.* Transmitted duplication of 8p23.1–8p23.2 associated with speech delay,
913 autism and learning difficulties. *Eur. J. Hum. Genet.* **17,** 37–43 (2009).
914 87. Rossignol, D. A. & Frye, R. E. Evidence linking oxidative stress, mitochondrial
915 dysfunction, and inflammation in the brain of individuals with autism. *Front. Physiol.* **5**
916 **APR,** 1–15 (2014).
917 88. Ma, X. *et al.* A genome-wide association study for quantitative traits in schizophrenia in
918 China. *Genes, Brain Behav.* **10,** 734–739 (2011).
919 89. Ni, P. *et al.* Methionine sulfoxide reductase A (MsrA) associated with bipolar I disorder
920 and executive functions in A Han Chinese population. *J. Affect. Disord.* **184,** 235–238
921 (2015).
922 90. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing
923 genomic features. *Bioinformatics* **26,** 841–842 (2010).
924 91. Cotney, J. *et al.* The evolution of lineage-specific regulatory activities in the human
925 embryonic limb. *Cell* **154,** 185–196 (2013).
926 92. Mathelier, A. *et al.* JASPAR 2014: An extensively expanded and updated open-access
927 database of transcription factor binding profiles. *Nucleic Acids Res.* **42,** 1–6 (2014).
928 93. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given
929 motif. *Bioinformatics* **27,** 1017–1018 (2011).
930 94. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and
931 wANNOVAR. *Nat Protoc* **10,** 1556–1566 (2015).
932 95. Xu, L. M. *et al.* AutismKB: An evidence-based knowledgebase of autism genetics.
933 *Nucleic Acids Res.* **40,** 1–7 (2012).
934 96. Betancur, C. Etiological heterogeneity in autism spectrum disorders: More than 100
935 genetic and genomic disorders and still counting. *Brain Res.* **1380,** 42–77 (2011).
936 97. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism

937 spectrum disorders. *Am. J. Hum. Genet.* **94,** 677–694 (2014).
938 98. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia.
939 *Nature* **506,** 185–190 (2014).
940 99. Bayés, A. *et al.* Comparative study of human and mouse postsynaptic proteomes finds
941 high compositional conservation and abundance differences for key synaptic proteins.
942 *PLoS One* **7,** (2012).
943

Figure 1

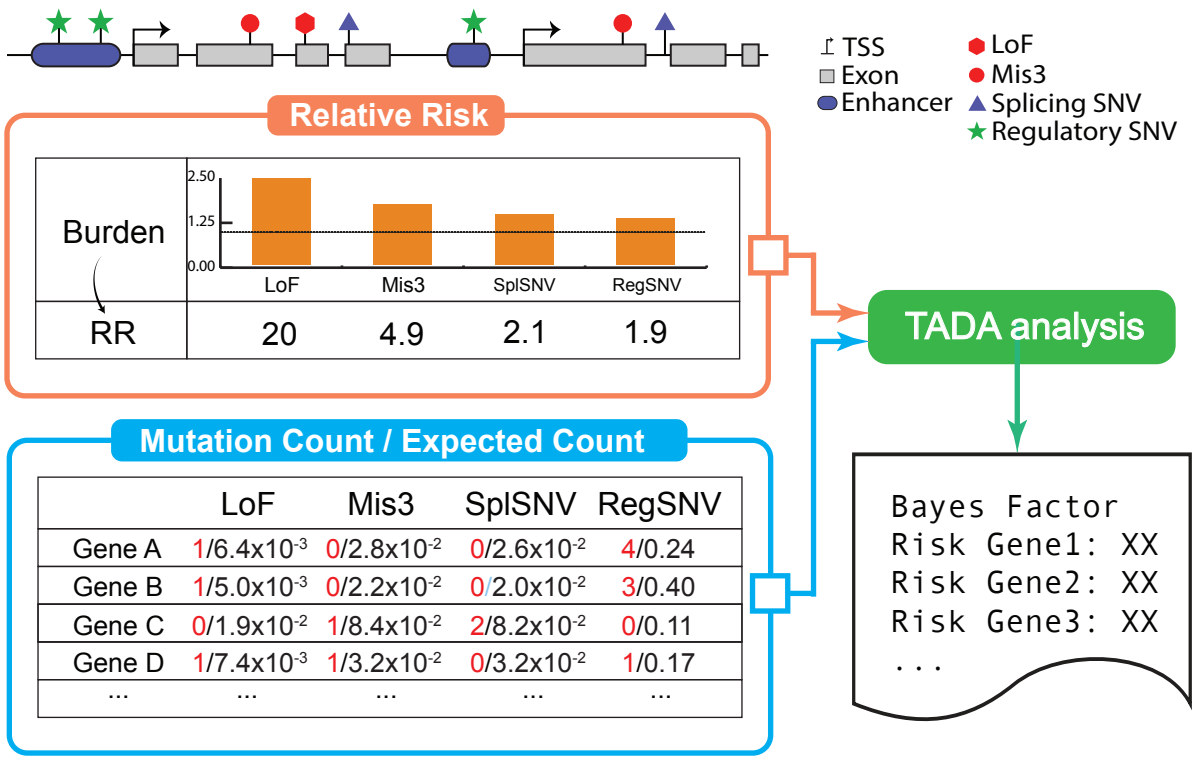

Figure 1. Overview of study design. We consider four mutational types, including LoF: Loss-of-Function coding mutations; Mis3: probably damaging nonsynonymous coding mutations predicted by Polyphen-2; SplSNV: SNVs predicted to affect pre-mRNA splicing by SPIDEX; RegSNV: non-coding SNVs that fall into brain H3K27ac regions within 10 kb of TSSs. RR denotes average relative risks of different types of mutations in risk genes. In the Mutation Count/ Expected Count Box, each cell contains the observed mutation count (colored in red) and the expected count (colored in black) (the numbers are for illustration only). TADA combines information from all mutation types related to a gene to test how likely it is to be a risk gene. The average relative risk of a mutation type is used in TADA to weigh the contribution of that type.

Figure 2

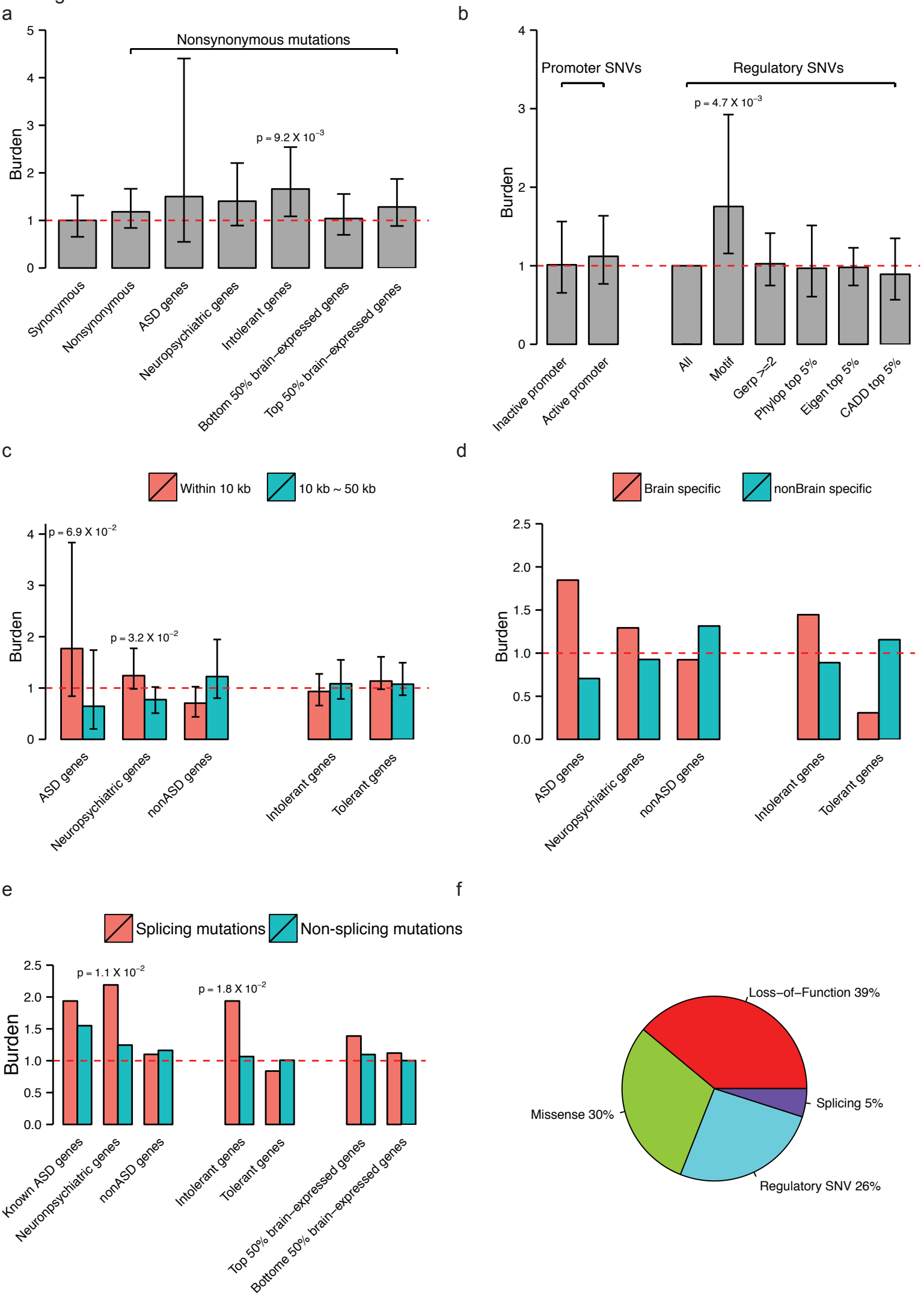
a
Nonsynonymous mutations
Burden
p = 9.2 X 10^-3
Synonymous
Nonsynonymous
ASD genes
Neuropsychiatric genes
Intolerant genes
Bottom 50% brain−expressed genes
Top 50% brain−expressed genes
b
Promoter SNVs
Regulatory SNVs
Burden
p = 4.7 X 10^-3
Inactive promoter
Active promoter
All
Motif
Gerp >=2
Phylop top 5%
Eigen top 5%
CADD top 5%
c
Within 10 kb
10 kb ~ 50 kb
Burden
p = 6.9 X 10^-2
p = 3.2 X 10^-2
ASD genes
Neuropsychiatric genes
nonASD genes
Intolerant genes
Tolerant genes
d
Brain specific
nonBrain specific
Burden
ASD genes
Neuropsychiatric genes
nonASD genes
Intolerant genes
Tolerant genes
e
Splicing mutations
Non-splicing mutations
Burden
p = 1.1 X 10^-2
p = 1.8 X 10^-2
Known ASD genes
Neuronpsychiatric genes
nonASD genes
Intolerant genes
Tolerant genes
Top 50% brain−expressed genes
Bottome 50% brain−expressed genes
f
Loss−of−Function 39%
Splicing 5%
Missense 30%
Regulatory SNV 26%

Figure 2. Burdens analyses of different types of de novo mutations. The error bars in (a), (b), and (c) represents the 95% confidence intervals of burdens, based on the lower bounds and upper bounds of the odds ratios from Fisher's exact tests. We did not show error bars in (d) and (e) as they are too large to show. (a) De novo nonsynonymous mutations from WGS studies in multiple gene sets. See Methods for the definitions of these gene sets. (b) De novo SNVs in H3K27ac regions within 10 kb of TSSs. In the barplot, we show burdens of promoter/5'UTR SNVs and regulatory SNVs (including all SNVs in brain H3K27ac regions within 10 kb of TSSs). (c) Regulatory SNVs and distal regulatory SNVs (10 kb ~ 50 kb to TSSs) in different gene sets. (d) Comparison between burdens of SNVs in brain-specific H3K27ac regions and of those in other-tissue-specific H3K27ac regions. (e) Burdens of SNVs that are predicted to affect splicing vs. SNVs not predicted to have such effects (non-splicing SNVs). (f) Partition of de novo ASD risk into coding and non-coding mutations.

Figure 3

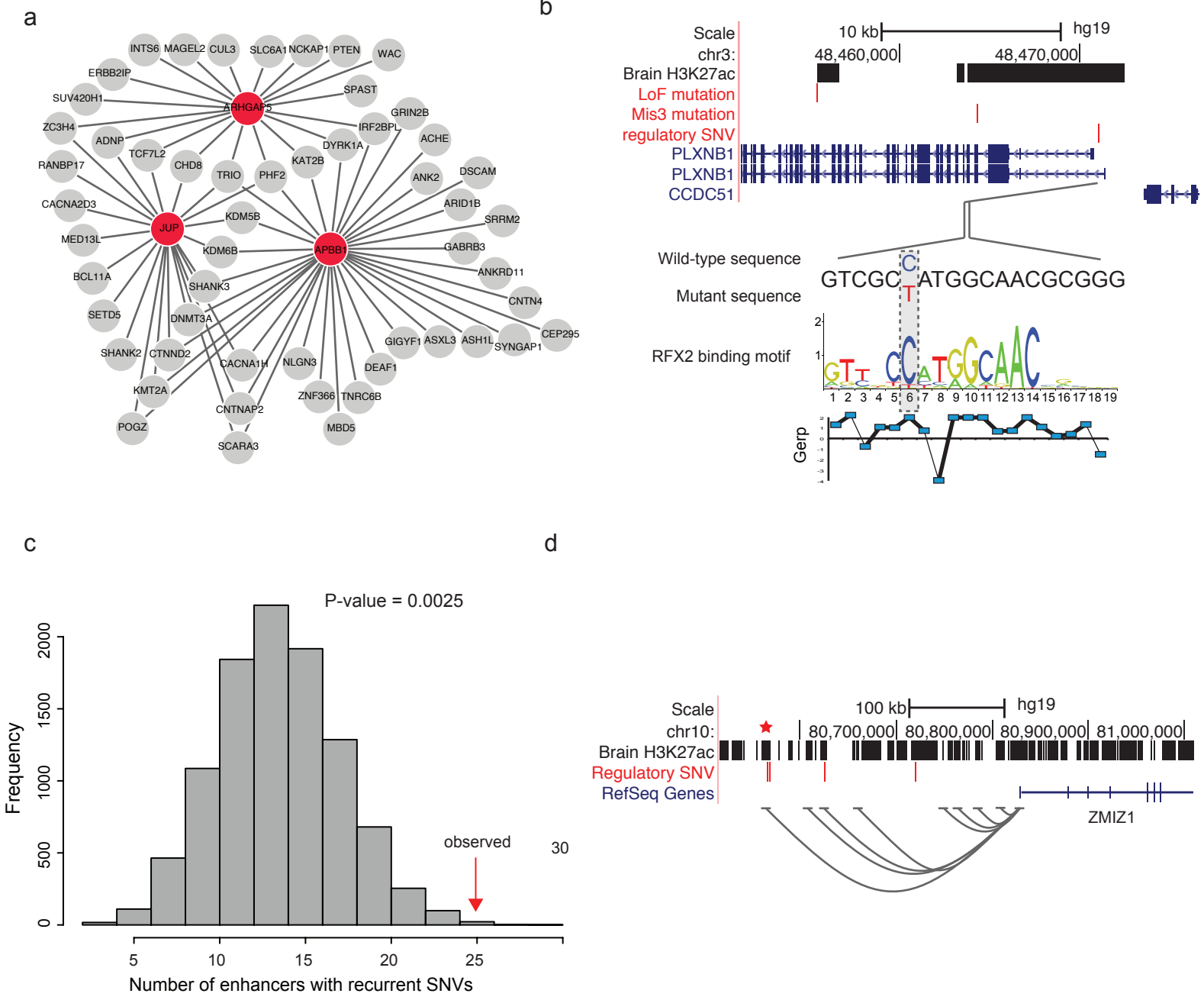

Figure 3. Predicted risk genes and enhancers of ASD. (a) GeneMania network analysis of the three new ASD genes. Red circles represent new ASD genes while grey circles represent known ones. The figure shows only the co-expression networks: two genes are connected if their co-expression across multiple 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. (c) Distribution of the number of enhancers with recurrent (more than two) de novo SNVs from 10,000 simulations. The vertical red arrow marks the observed number of enhancers with recurrent de novo SNVs from ASD data. (d) A plausible distal enhancer (marked by a star) with recurrent ASD SNVs has correlated activities with ZMIZ1 promoter. Grey curves represent correlated activities between enhancers and ZMIZ1 promoter.

Figure 4

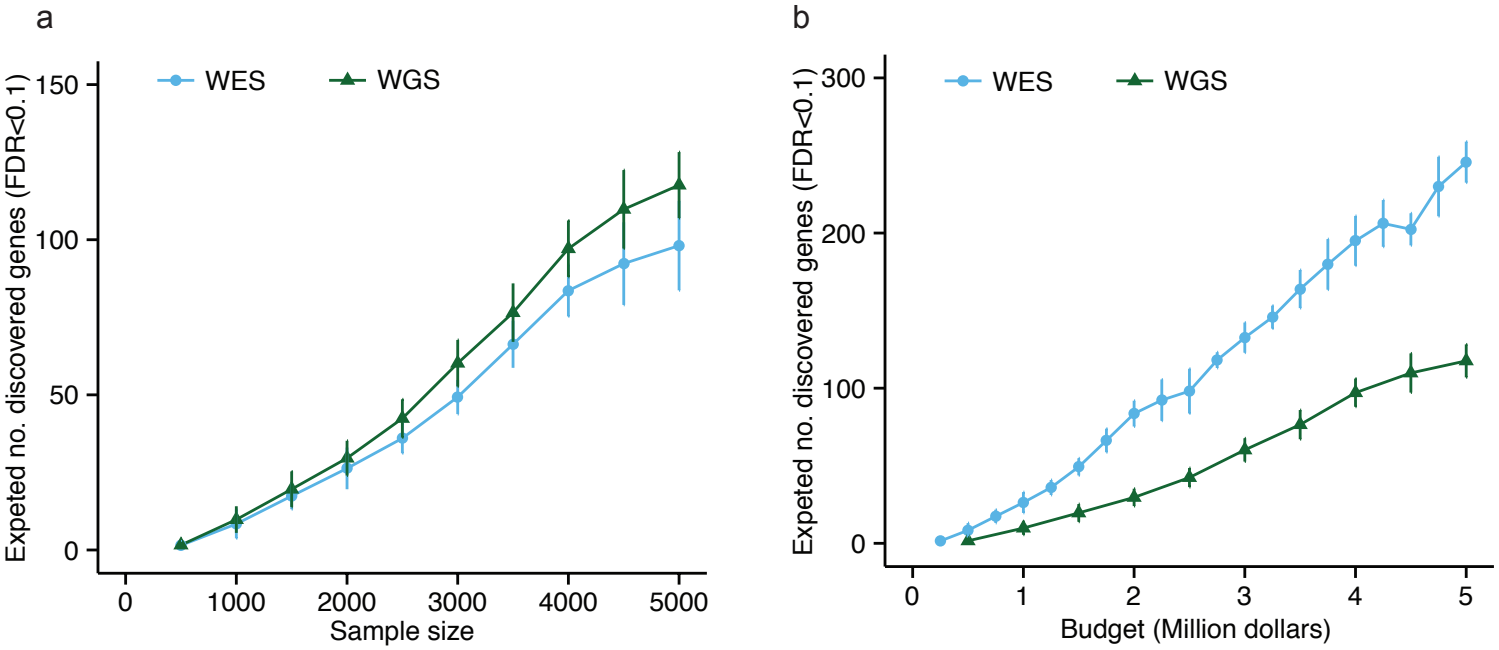

Figure 4. Comparison of power between WES and WGS from simulations. Power is measured as the expected number of discovered ASD risk genes at $q < 0.1$, and is obtained at each level of sample size (a) or sequencing cost (b).

Table 1. Mutational exposure (number of mutations from ASD risk genes per subject) and relative risk of different types of *de novo* mutations.

Mutation Class	**Mutational Exposure**	**Relative Mutational Exposure(%)**	**Relative risk**	**Variance of ASD liability explained (100%)**
Mis3	0.0175	11.6	4.7	0.83
Loss-of-function	0.00405	2.68	20	1.08
Regulatory SNV	0.115	76.1	1.9	0.72
Splicing SNV	0.0145	9.60	2.1	0.14

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

Gene name	JUP	APBB1	ARHGAP5	Enrichment p-value
LoF	1	1	1	
Mis3	0	0	1	
Regulatory SNV	4	3	1	
HI	Y	Y	Y	**7.90e-4**
RVIS (%)	2.74	19.93	20.02	**1.11e-2**
ExAC zscore (%)	30.62	14.25	9.69	1.38e-1
FMRP targets	N	Y	N	1.25e-1
BrainSpan expression (%)	19.29	3.41	22.03	**1.49e-2**
DAWN	5.00e-4	-	2.00e-4	**4.06e-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).

Table 3. Biological processes enriched among neighborhood genes of the three new ASD genes.

Term	P-value	Adjusted P-value	Overlap	Neighbor Genes
Adherens junction organization	1.17e-05	1.06e-02	4/69	DSP, CTNND1, CTNNA1(APBB1); THY1(ARHGAP5)
Negative regulation of neurogenesis	3.59e-05	1.27e-02	5/192	APP, NOTCH1, LRP4(JUP); CTNNA1(APBB1); THY1(ARHGAP5)
Neuromuscular junction development	5.00e-05	1.27e-02	3/34	APP, ERBB2, LRP4(JUP)
egative regulation of nervous system development	5.57e-05	1.27e-02	5/211	APP, NOTCH1, CTNNA1(APBB1); LRP4(JUP); THY1(ARHGAP5)
Negative regulation of cell development	9.92e-05	1.51e-02	5/239	APP, NOTCH1, LRP4 (JUP); CTNNA1(APBB1); THY1(ARHGAP5)
Kidney development	1.00e-04	1.52e-02	4/122	MME, TSHZ3, LRP4 (JUP); ZBTB16(APBB1)
Neuron remodeling	1.64e-04	1.98e-02	2/8	APP(JUP); RND1(ARHGAP5)
Forebrain development	1.74e-04	1.98e-02	3/53	APP, NOTCH1, APLP2(JUP)
Negative regulation of neuron differentiation	2.01e-04	2.04e-02	4/147	APP, NOTCH, LRP4(JUP); THY1(ARHGAP5)
Mating behavior	2.41e-04	2.20e-02	2/10	APP, APLP2(JUP)

Neighborhood genes are the top 30 connected genes from GeneMania analysis for the three new ASD genes. In the last column, after each set of neighbor genes (separated by semicolon) in parenthesis is their common connected new ASD gene. In the “Overlap” column, the number of neighborhood genes in each GO category is followed by the total number of genes in this GO category.