Synaptic, transcriptional and chromatin genes disrupted in autism

A list of authors and their affiliations appears at the end of the paper.

The genetic architecture of autism spectrum disorder involves the interplay of common and rare variants and their impact on hundreds of genes. Using exome sequencing, here we show that analysis of rare coding variation in 3,871 autism cases and 9,937 ancestry-matched or parental controls implicates 22 autosomal genes at a false discovery rate (FDR) < 0.05, plus a set of 107 autosomal genes strongly enriched for those likely to affect risk (FDR < 0.30). These 107 genes, which show unusual evolutionary constraint against mutations, incur de novo loss-of-function mutations in over 5% of autistic subjects. Many of the genes implicated encode proteins for synaptic formation, transcriptional regulation and chromatin-remodelling pathways. These include voltage-gated ion channels regulating the propagation of action potentials, pacemaking and excitability—transcription coupling, as well as histone-modifying enzymes and chromatin remodellers—most prominently those that mediate post-translational lysine methylation/demethylation modifications of histones.

Features of subjects with autism spectrum disorder (ASD) include compromised social communication and interaction. Because the bulk of risk arises from de novo and inherited genetic variation, characterizing which genes are involved informs ASD neurobiology and reveals part of what makes us social beings.

Whole-exome sequencing (WES) studies have proved fruitful in uncovering risk-conferring variation, especially by enumerating de novo variation, which is sufficiently rare that recurrent mutations in a gene provide strong evidence for a causal link to ASD. De novo loss-of-function (LoF) single-nucleotide variants (SNVs) or insertion/deletion (indel) variants are found in 6.7% more ASD subjects than in matched controls and implicate nine genes from the first 1,000 ASD subjects analysed. Moreover, because there are hundreds of genes involved in ASD risk, ongoing WES studies should identify additional ASD genes as an almost linear function of increasing sample size.

Here we conduct the largest ASD WES study so far, analysing 16 sample sets comprising 15,480 DNA samples (Supplementary Table 1 and Extended Data Fig. 1). Unlike earlier WES studies, we do not rely solely on counting de novo LoF variants, rather we use novel statistical methods to assess association for autosomal genes by integrating de novo, inherited and case-control LoF counts, as well as de novo missense variants predicted to be damaging. For many samples original data from sequencing performed on Illumina HiSeq 2000 systems were used to call SNVs and indels in a single large batch using GATK (v2.6). De novo mutations were called using enhancements of earlier methods (Supplementary Information), with calls validating at extremely high rates.

After evaluation of data quality, high-quality alternative alleles with a frequency of <0.1% were identified, restricted to LoF (frameshifts, stop gains, donor/acceptor splice site mutations) or probably damaging missense (Mis3) variants (defined by PolyPhen-2 (ref. 18)). Variants were classified by type (de novo, case, control, transmitted, non-transmitted) and severity (LoF, Mis3), and counts tallied for each gene.

Some 13.8% of the 2,270 ASD trios (two parents and one affected child) carried a de novo LoF mutation—significantly in excess of both the expected value (8.6%, P = 10^{-14}) and what was observed in 510 control trios (7.1%, P = 1.6 × 10^{-5}) collected here and previously published. Eighteen genes (Table 1) exhibited two or more de novo LoF mutations. These genes are all known or strong candidate ASD genes, but given the number of trios sequenced and gene mutability, we would expect to observe this in approximately two such genes by chance. While we expect only two de novo Mis3 events in these 18 genes, we observe 16 (P = 9.2 × 10^{-11}, Poisson test). Because most of our data exist in cases and controls and because we observed an additional excess of transmitted LoF events in the 18 genes, it is evident that the optimal analytical framework must involve an integration of de novo mutation with variants observed in cases and controls and transmitted or untransmitted from carrier parents. Investigating beyond de novo LoFs is also critical given that many ASD risk genes and loci have mutations that are not completely penetrant.

Transmission and de novo association

We adopted TADA (transmission and de novo association), a weighted, statistical model integrating de novo, transmitted and case-control variation. TADA uses a Bayesian gene-based likelihood model including per-gene mutation rates, allele frequencies, and relative risks of particular classes of sequence changes. We modelled both LoF and Mis3 sequence variants. Because no aggregate association signal was detected for inherited Mis3 variants, they were not included in the analysis. For each gene, variants of each class were assigned the same effect on relative risk. Using a prior probability distribution of relative risk across genes for each class of variants, the model effectively weighted different classes of variants in this order: de novo LoF > de novo Mis3 > transmitted LoF, and allowed for a distribution of relative risks across genes for each class. The strength of association was assimilated across classes to produce a gene-level Bayes factor with a corresponding FDR q value. This framework increases the power compared to the use of de novo LoF variants alone (Extended Data Fig. 2).

TADA identified 33 autosomal genes with an FDR < 0.1 (Table 1) and 107 with an FDR < 0.3 (Supplementary Tables 2 and 3 and Extended Data Fig. 3). Of the 33 genes, 15 (45.5%) are known ASD risk genes; 11 have been reported previously with mutations in ASD patients but were not classed as true risk genes owing to insufficient evidence (SUV420H1 (refs 11, 15), ADNP [12], BCL11A [13], CACNA2D3 (refs 15, 21), CTTNB282 (ref. 15), GABRB3 (ref. 21), CDC42BBP [14], APHIA4 [15], NR3C2 (ref. 15), SETD5 (refs 14, 22) and TRIO [16]) and 7 are completely novel (ASH1L, MLL3 (also known as KMT2C), ETFB, NAA15, MYO9B, MIB1 and VIL1). ADNP mutations have recently been identified in 10 patients with ASD and other shared clinical features [12]. Two of the newly discovered genes,
ASHIL and MLL3, converge on chromatin remodelling. MYO9B plays a key role in dendritic arborisation29. MIB1 encodes an E3 ubiquitin ligase critical for neurogenesis26 and is regulated by miR-137 (ref. 26), a microRNA that regulates neuronal maturation and is implicated in schizophrenia risk27.

When the WES data from genes with an FDR < 0.3 were evaluated for the presence of deletion copy number variants (CNVs) (such CNVs are functionally equivalent to LoF mutations), 34 CNVs meeting quality and frequency constraints (Supplementary Information) were detected in 5,781 samples (Extended Data Fig. 1). Of the 33 genes with an FDR < 0.1, 3 contained deletion CNVs mapping to 3 ASD subjects and one parent. Of the 74 genes meeting the criterion 0.1 ≤ FDR < 0.3, about one-third could be false positives. Deletion CNVs were found in 14 of these genes and the data supported risk status for 10 of them (Extended Data Table 1 and Extended Data Fig. 4). Two of these ten, NRXN1 and SHANK3, were previously implicated in ASD30,31. The risk from deletion CNVs, as measured by the odds ratio, is comparable to that from LoF SNVs in cases versus controls or transmission of LoF variants with a modest impact for LoF variants (odds ratio range 2–4, whether de novo or transmitted or present in ASD subjects). dnLoF, de novo mutations, can have a large impact, others smaller, and still others have no effect at all. In addition, mis-annotation of variants, among other confounds, can yield false variant calls in subjects (Supplementary Information). These confounds can often be overcome by examining the data in a manner orthogonal to gene discovery. For example, females have greatly reduced rates of ASD relative to males (a ‘female protective effect’). Consequently, and regardless of whether this is diagnostic bias or biological protection, females have a higher liability threshold, requiring a larger genetic burden before being diagnosed28,29. A corollary is that if a variant has the same effect on autism liability in males as it does in females, that variant will be present at a higher frequency in female ASD cases compared to males. Importantly, the magnitude of the difference is proportional to risk as measured by the odds ratio; hence, the effect on risk for a class of variants can be estimated from the difference in frequency between males and females.

Genes with an FDR < 0.1 show profound female enrichment for de novo events (P = 0.005 for LoF, P = 0.004 for Mis3), consistent with de novo events having large impacts on liability (odds ratio ≥ 20; Extended Data Fig. 5). However, genes with an FDR between 0.1 and 0.3 show substantially less enrichment for female events, consistent with a modest impact for LoF variants (odds ratio range 2–4, whether transmitted or de novo) and little to no effect from Mis3 variants. The class of variants (for example, LoF) some genes have a large impact, others smaller, and still others have no effect at all. In addition, mis-annotation of variants, among other confounds, can yield false variant calls in subjects (Supplementary Information). These confounds can often be overcome by examining the data in a manner orthogonal to gene discovery. For example, females have greatly reduced rates of ASD relative to males (a ‘female protective effect’). Consequently, and regardless of whether this is diagnostic bias or biological protection, females have a higher liability threshold, requiring a larger genetic burden before being diagnosed28,29. A corollary is that if a variant has the same effect on autism liability in males as it does in females, that variant will be present at a higher frequency in female ASD cases compared to males. Importantly, the magnitude of the difference is proportional to risk as measured by the odds ratio; hence, the effect on risk for a class of variants can be estimated from the difference in frequency between males and females.

Genes with an FDR < 0.1 show profound female enrichment for de novo events (P = 0.005 for LoF, P = 0.004 for Mis3), consistent with de novo events having large impacts on liability (odds ratio ≥ 20; Extended Data Fig. 5). However, genes with an FDR between 0.1 and 0.3 show substantially less enrichment for female events, consistent with a modest impact for LoF variants (odds ratio range 2–4, whether transmitted or de novo) and little to no effect from Mis3 variants. The class of variants (for example, LoF) some genes have a large impact, others smaller, and still others have no effect at all. In addition, mis-annotation of variants, among other confounds, can yield false variant calls in subjects (Supplementary Information). These confounds can often be overcome by examining the data in a manner orthogonal to gene discovery. For example, females have greatly reduced rates of ASD relative to males (a ‘female protective effect’). Consequently, and regardless of whether this is diagnostic bias or biological protection, females have a higher liability threshold, requiring a larger genetic burden before being diagnosed28,29. A corollary is that if a variant has the same effect on autism liability in males as it does in females, that variant will be present at a higher frequency in female ASD cases compared to males. Importantly, the magnitude of the difference is proportional to risk as measured by the odds ratio; hence, the effect on risk for a class of variants can be estimated from the difference in frequency between males and females.

Genes with an FDR < 0.1 show profound female enrichment for de novo events (P = 0.005 for LoF, P = 0.004 for Mis3), consistent with de novo events having large impacts on liability (odds ratio ≥ 20; Extended Data Fig. 5). However, genes with an FDR between 0.1 and 0.3 show substantially less enrichment for female events, consistent with a modest impact for LoF variants (odds ratio range 2–4, whether transmitted or de novo) and little to no effect from Mis3 variants. The class of variants (for example, LoF) some genes have a large impact, others smaller, and still others have no effect at all. In addition, mis-annotation of variants, among other confounds, can yield false variant calls in subjects (Supplementary Information). These confounds can often be overcome by examining the data in a manner orthogonal to gene discovery. For example, females have greatly reduced rates of ASD relative to males (a ‘female protective effect’). Consequently, and regardless of whether this is diagnostic bias or biological protection, females have a higher liability threshold, requiring a larger genetic burden before being diagnosed28,29. A corollary is that if a variant has the same effect on autism liability in males as it does in females, that variant will be present at a higher frequency in female ASD cases compared to males. Importantly, the magnitude of the difference is proportional to risk as measured by the odds ratio; hence, the effect on risk for a class of variants can be estimated from the difference in frequency between males and females.

Genes with an FDR < 0.1 show profound female enrichment for de novo events (P = 0.005 for LoF, P = 0.004 for Mis3), consistent with de novo events having large impacts on liability (odds ratio ≥ 20; Extended Data Fig. 5). However, genes with an FDR between 0.1 and 0.3 show substantially less enrichment for female events, consistent with a modest impact for LoF variants (odds ratio range 2–4, whether transmitted or de novo) and little to no effect from Mis3 variants. The class of variants (for example, LoF) some genes have a large impact, others smaller, and still others have no effect at all. In addition, mis-annotation of variants, among other confounds, can yield false variant calls in subjects (Supplementary Information). These confounds can often be overcome by examining the data in a manner orthogonal to gene discovery. For example, females have greatly reduced rates of ASD relative to males (a ‘female protective effect’). Consequently, and regardless of whether this is diagnostic bias or biological protection, females have a higher liability threshold, requiring a larger genetic burden before being diagnosed28,29. A corollary is that if a variant has the same effect on autism liability in males as it does in females, that variant will be present at a higher frequency in female ASD cases compared to males. Importantly, the magnitude of the difference is proportional to risk as measured by the odds ratio; hence, the effect on risk for a class of variants can be estimated from the difference in frequency between males and females.

Genes with an FDR < 0.1 show profound female enrichment for de novo events (P = 0.005 for LoF, P = 0.004 for Mis3), consistent with de novo events having large impacts on liability (odds ratio ≥ 20; Extended Data Fig. 5). However, genes with an FDR between 0.1 and 0.3 show substantially less enrichment for female events, consistent with a modest impact for LoF variants (odds ratio range 2–4, whether transmitted or de novo) and little to no effect from Mis3 variants. The
results are consistent with inheritance patterns: LoF mutations in FDR < 0.1 genes are rarely inherited from unaffected parents whereas those in the 0.1 ≤ FDR < 0.3 group are far more often inherited than they are de novo mutations.

By analysing the distribution of relative risk over inferred ASD genes\(^{20}\), the number of ASD risk genes can be estimated. The estimate relies on the balance of genes with multiple de novo LoF mutations versus those with only one: the larger the number of ASD genes, the greater proportion that will show only one de novo LoF. This approach yields an estimate of 1,150 ASD genes (Supplementary Information). While there are many more genes to be discovered, many will have a modest impact on risk compared to the genes in Table 1.

**Enrichment analyses**

Gene sets with an FDR < 0.3 are strongly enriched for genes under evolutionary constraint\(^{19} (P = 3.0 \times 10^{-7})\), consistent with the hypothesis that heterozygous LoF mutations in these genes are ASD risk factors. Over 5% of ASD subjects carry de novo LoF mutations in our FDR < 0.3 list. We also observed that genes in the FDR < 0.3 list had a significant excess of de novo non-synonymous events detected by the largest schizophrenia WES study so far\(^{20} (P = 0.0085\); Fig. 1a), providing further evidence for overlapping risk loci between these disorders and independent confirmation of the signal in the gene sets presented here.

We found significant enrichment for genes encoding messenger RNAs targeted by two neuronal RNA-binding proteins: FMRP\(^{31}\) (also known as FMR1), mutated or absent in fragile X syndrome\((P = 1.20 \times 10^{-17})\), 34 targets\(^{31}\), of which 11 are corroborated by an independent data set\(^{32}\), and RBFOX (RBFOX1/2/3)\((P = 0.0024\); 20 targets, of which 12 overlap with FMRP), with RBFOX1 shown to be a splicing factor dysregulated in ASD\(^{33,34}\) (Fig. 1a). These two pathways expand the complexity of ASD neurobiology to post-transcriptional events, including splicing and translation, both of which sculpt the neural proteome.

We found nominal enrichment for human orthologues of mouse genes encoding synaptic\((P = 0.031)\) and post-synaptic density (PSD) proteins\(^{35} (P = 0.046\); Fig. 1a, b and Supplementary Tables 4–6). Enrichment analyses for InterPro, SMART or Pfam domains (FDR < 0.05 and a minimum of five genes per category) reveal an overrepresentation of DNA- or histone-related domains: eight genes encoding proteins with InterPro zinc-finger FYVE PHD domains (142 such annotated genes in the genome; FDR = 7.6 \times 10^{-4}), and five with Pfam Su(var)3-9, enhancer-of-zeste, trithorax (SET) domains (39 annotated in the genome; FDR = 8.2 \times 10^{-4}).

**Integrating complementary data**

To implicate additional genes in risk for ASD, we used a model called DAWN (detecting association with networks)\(^{36}\). DAWN evokes a hidden Markov random field framework to identify clusters of genes that show strong association signals and highly correlated co-expression in a key tissue and developmental context. Previous research suggests human mid-fetal prefrontal and motor-somatosensory neocortex is a critical nexus for risk\(^{36}\), thus we evaluated gene co-expression data from that tissue together with TADA scores for genes with an FDR < 0.3. Because this list is enriched for genes under evolutionary constraint, we generalized DAWN to incorporate constraint scores (Supplementary Information). When TADA results, gene co-expression in mid-fetal neocortex and constraint scores are jointly modelled, DAWN identifies 160 genes that plausibly affect risk\((P = 0.018)\).

A subnetwork obtained by seeding the 160 DAWN genes within a high-confidence protein–protein interactome\(^{37}\) confirmed that the putative genes are enriched for neuronal functions. We kept the largest connected component, containing 95 seed DAWN genes, 50 of which were in the FDR < 0.3 gene set. The DAWN gene products form four natural

[Figure 2 | ASD genes in neuronal networks. Protein–protein interaction network created by seeding TADA and DAWN-predicted genes. Only intermediate genes that are known to interact with at least two TADA and/or DAWN genes are included. Four natural clusters (C1–C4) are demarcated with black ellipses. All nodes are sized on the basis of degree of connectivity.]
clusters on the basis of network connectivity (Fig. 2). We visualized the enriched pathways and biological functions for each of these clusters on ‘canvases’ (Extended Data Fig. 6). Many of the previously known ASD risk genes fall in cluster C3, including genes involved in synaptic transmission and cell–cell communication. Cluster C4 is enriched for genes related to transcriptional and chromatin regulation. Many TADA and DAWN genes in this cluster interact tightly with other transcription factors, histone-modifying enzymes and DNA-binding proteins. Five TADA genes in the cluster C2 are bridged to the rest of the network through MAPT, as inferred by DAWN. The enrichment results for cluster C2 indicate that genes implicated in neurodegenerative disorders could also have a role in neurodevelopmental disorders.

**Emergent results**

Amongst the critical synaptic components found to be mutated in our study are voltage-gated ion channels involved in fundamental processes including the propagation of action potentials (for example, the Na\(_A\), 1.2 channel), neuronal pacemaking and excitability–transcription coupling (for example, the Ca\(_{4,1}\)3 channel) (Fig. 1b). We identified four LoF and five Mis3 variants in SCN2A (Na\(_A\), 1.2), three Mis3 variants in CACNA1D (Ca\(_{1,2}\)a, 1.3) and two LoF variants in CACNA2D3 (\(\gamma\,\delta\)-3 subunit). Remarkably, three de novo Mis3 variants in SCN2A affected residues mutated in homologous genes in patients with other syndromes, including Brugada syndrome (SCNSA) or epilepsy disorders (SCN1A) (Arg379His and Arg 937His). These arginines, as well as the threonine mutated in Thr1420Met, cluster to the P-loops forming the ion selectivity filter, located in proximity to the inner ring (DEKA motif) (Fig. 1c). Because homologous channels mutated in these arginines do not conduct inward Na\(^{+}\) currents, Arg379His and Arg937His mutations might have similar effect.

Two de novo CACNA1D variants (Gly407Arg and Ala749Gly) emerged at positions proximal to residues mutated in patients with primary aldosteronism and neurological deficits (Fig. 1d). The reported mutations interfere with channel activation and inactivation. Amongst variants found in cases, Ala59Val maps to the NSCaTE domain, also important for Ca\(^{2+}\)-dependent inactivation, and Ser1977Leu and Arg2021His co-cluster in the carboxy-terminal proline-rich domain, the site of interaction with SHANK3, a key PSD scaffolding protein. Mutations in RIM12 and RIMBP2, which can associate with Ca\(_{4,1}\), 1.3, were found in our cohort (but with an FDR > 0.3).

Chromatin remodelling involves histone-modifying enzymes (encoded by histone-modifier genes, HMGs) and chromatin remodelers (readers) that recognize specific histone post-translational modifications and orchestrate their effects on chromatin. Our gene set is enriched in HMGs (9 HMGs out of 152 annotated in HIsome, Fisher’s exact test, \(P = 2.2 \times 10^{-7}\)). Enrichment in the gene ontology term ‘histone-lysine N\(^{m}\)-methyltransferase activity’ (5 genes out of 41 so annotated; FDR = 2.2 \times 10^{-7}) highlights this as a prominent pathway.

Lysines on histones 3 and 4 can be mono-, di- or tri-methylated, providing a versatile mechanism for either activation or repression of transcription. Of 107 TADA genes, five are SET lysine methyltransferases, four are jumonji lysine demethylases, and two are readers (Fig. 3a). RBFOX1 co-isolates with histone H3 trimethyl Lys 4 (H3K4me3)42, and our data set is enriched in targets shared by RBFOX1 and H3K4me3 (\(P = 0.0166\); Fig. 1a and Supplementary Table 4). Some de novo missense variants targeting these genes map to functional domains (Extended Data Fig. 7).

For the H3K4me2 reader CHD8, we extended our analyses in search of additional de novo variation in the cases of the case-control sample. By sequencing complete parent–child trios for many CHD8 variants, five variants were found to be de novo, two of which affect essential splice sites and cause LoF by exon skipping or activation of cryptic splice sites in lymphoblastoid cells (Fig. 3b).

Given the role of HMGs in transcription, we reasoned that TADA genes might be interconnected through transcription ‘routes’. We searched for a connected network (seeded by 9 TADA HMGs) in a transcription factor interaction network (ChEA)43. We found that 46 TADA genes are directly interconnected in a 55-gene cluster (Extended Data Fig. 8) (\(P = 0.002; 1,000\) random draws), for a total of 69 when including all known HMGs (Fig. 4) (\(P = 0.001; 1,000\) random draws).

Examining the Human Gene Mutation Database we found that the 107 TADA genes included 21 candidate genes for intellectual disability, 3 for epilepsy, 17 for schizophrenia, 9 for congenital heart disease and 6 for metabolic disorders (Fig. 5).

**Conclusions**

Complementing earlier reports, ASD subjects show a clear excess of de novo LoF mutations above expectation, with a concentration of such events in a handful of genes. While this handful has a large effect on risk, most ASD genes have a much smaller impact. This gradient emerges most notably from the contrast of risk variation in male and female ASD subjects. Unlike some earlier studies, but consistent with expectation, the data also show clear evidence for effect of de novo missense SNVs.
on risk; for risk generated by LoF variants transmitted from unaffected parents; and for the value of case-control design in gene discovery. By integrating data on \textit{de novo}, inherited and case-control variation, the yield of ASD gene discoveries was doubled over what would be obtained from a count of \textit{de novo} LoF variants alone. ASD genes almost uniformly show strong constraints against variation, a feature we exploit to implicate other genes in risk.

Three critical pathways for typical development are damaged by risk variation: chromatin remodelling, transcription and splicing, and synaptic function. Chromatin remodelling controls events underlying the formation of neural connections, including neurogenesis and neural differentiation\textsuperscript{44}, and relies on epigenetic marks as post-translational modifications of histones. Here we provide extensive evidence for HMGs and readers in sporadic ASD, implicating specifically lysine methylation and extending the mutational landscape of the emergent ASD gene CHD8 to missense variants. Splicing is implicated by the enrichment of RBFOX targets in the top ASD candidates. Risk variation also affects multiple classes and components of synaptic networks, from receptors and ion channels to scaffolding proteins. Because a wide set of synaptic genes is disrupted in idiopathic ASD, it seems reasonable to suggest that altered chromatin dynamics and transcription, induced by disruption of relevant genes, leads to impaired synaptic function as well.

De novo mutations in ASD\textsuperscript{11–15}, intellectual disability\textsuperscript{45} and schizophrenia\textsuperscript{30} cluster to synaptic genes, and synaptic defects have been reported in models of these disorders\textsuperscript{46}. Integrity of synaptic function is essential for neural physiology, and its perturbation could represent the intersection between diverse neuropsychiatric disorders\textsuperscript{47}.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 18 May; accepted 18 August 2014.
Published online 29 October; corrected online 12 November 2014 (see full-text HTML version for details).

1. Ronald, A. & Hoekstra, R. A. Autism spectrum disorders and autistic traits: a decade of new twin studies. \textit{Am. J. Med. Genet. B Neuropsychiatr. Genet.} \textbf{156}, 255–274 (2011).
2. Sebat, J. et al. Strong association of \textit{de novo} copy number mutations with autism. \textit{Science} \textbf{316}, 445–449 (2007).
3. Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. \textit{Nature} \textbf{466}, 368–372 (2010).
4. Klei, L. et al. Common genetic variants, acting additively, are a major source of risk for autism. \textit{Mol. Autism} \textbf{3}, 9 (2012).
39. Volkers, L.
40. Voineagu, I.
28. Robinson, E. B., Lichtenstein, P., Anckarsater, H., Happe, F. & Ronald, A. Examining de novo mutations in autism spectrum disorders. Nature 485, 237–241 (2012).
12. O’Roak, B. J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485, 246–250 (2012).
4. Neale, B. M. et al. Patterns and rates of de novo mutations in autism spectrum disorders. Nature 485, 242–245 (2012).
1. Iossifov, I. et al. De novo gene disruptions in children on the autistic spectrum. Nature 470, 285–299 (2011).
2. DePristo, M. A. et al. A framework for accurate and interpretable discovery of medical genomics. Nature 467, 1067–1072 (2010).
11. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485, 239–242 (2012).
7. De novo mutations are enriched in genes related to neuronal function and autism. Nature 485, 240–242 (2012).
17. DePristo, M. A. et al. De novo mutations are enriched in genes related to neuronal function and autism. Nature 485, 240–242 (2012).
10. Glessner, J. T. et al. Rare complete knockouts in humans: population distribution and significance role in autism spectrum disorders. Nature Genet. 41, 569–573 (2009).
9. Penzes, P., Cahill, M. E., Jones, K. A., VanLeeuwen, J. E. & Woolfrey, K. M. Dendritic spine pathology in neuropsychiatric disorders. Nature Rev. Neurosci. 14, 347–359 (2013).
8. Rauch, A. et al. Range of genetic mutations associated with severe non-syndromic congenital intellectual disability: an exome sequencing study. Lancet 380, 1674–1682 (2012).
7. Zoghbi, H. Y. Post-stroke developmental disorders: meeting at the synapse? Science 302, 826–830 (2003).

**Supplementary Information** is available in the online version of the paper.
Extended Data Figure 1 | Workflow of the study. The workflow began with 16 sample sets, as listed in Supplementary Table 1. DNA was obtained, and exomes were captured and sequenced. After variant calling, quality control was performed: duplicate subjects and incomplete families were removed and subjects with extreme genotyping, de novo, or variant rates were removed. Following cleaning, 3,871 subjects with ASD remained. Analysis proceeded separately for SNVs and indels, and CNVs. De novo and transmission/non-transmission variants were obtained for trio data (published de novo variants from 825 trios\textsuperscript{11,13–15} were incorporated). This led to the TADA analysis, which found 33 ASD risk genes with an FDR < 0.1; and 107 with an FDR < 0.3. CNVs were called in 2,305 ASD subjects. BAM, binary alignment/map; MAF, minor allele frequency.
Extended Data Figure 2 | Expected number of ASD genes discovered as a function of sample size. The multiple LoF test (red) is a restricted version of TADA that uses only the de novo LoF data. TADA (blue) models de novo LoF, de novo Mis3, LoF variants transmitted/not transmitted and LoF variants observed in case-control samples. The sample size (n) indicates either n trios for which we record de novo and transmitted variation (TADA), or n trios for which we record only de novo events (multiple LoF), plus n cases and n controls.
Extended Data Figure 3 | Heat map of the numbers of variants used in TADA analysis from each data set in genes with an FDR < 0.3. Left, variants in affected subjects; right, unaffected subjects. For the counts, we only included de novo LoF and Mis3 variants, transmitted/untransmitted and case-control LoF variants. These variant counts are normalized by the length of coding regions of each gene and sample size of each data set (|trio| + |case| for the left, |trio| + |control| for the right). Description of the samples can be found in Supplementary Table 1.
Extended Data Figure 4 | Genome browser view of the CNV deletions identified in ASD-affected subjects. The deletions are displayed in red if with unknown inheritance, in grey if inherited, and in black in unaffected subjects. Deletions in parents are not shown. For deletions within a single gene, all splicing isoforms are shown.
Extended Data Figure 5 | Frequency of variants by gender. Frequency of de novo (dn) and transmitted (Tr) variants per sample in males (black) and females (white) for genes with an FDR < 0.1 (top row), FDR < 0.3 (middle row), or all TADA genes (bottom row). The P values were determined by one-tailed permutation tests (*P < 0.05; **P < 0.01; ***P < 0.01).
Extended Data Figure 6 | Enrichment terms for the four clusters identified by protein–protein interaction networks. P-values calculated using mouse-genome-informatics–mammalian-phenotype (MGI_Mammalian phenotype, blue), Kyoto encyclopaedia of genes and genomes (KEGG) pathways (red), and gene ontology biological processes (yellow) are indicated.
Extended Data Figure 7 | De novo variants in SET lysine methyltransferases and jumonji lysine demethylases. Mis3 variants are in black, LoF in red, and variants identified in other disorders in grey (Fig. 5). ARID, AT-rich interacting domain; AWS, associated with SET domain; BAH, bromo adjacent homology; bromo, bromodomain; FYR C, FY-rich C-terminal domain; FYR N, FY-rich N-terminal domain; HiMG, high mobility group box; JmjC, jumonji C domain; JmjN, jumonji N domain; PHD, plant homeodomain; PWWP, Pro-Trp-Trp-Pro domain; SET, Su(var)3-9, enhancer-of-zeste, trithorax domain.
Extended Data Figure 8 | Transcription regulation network of TADA genes only. Edges indicate transcription regulators (source nodes) and their gene targets (target nodes) based on the ChEA network.
### Extended Data Table 1 | CNVs hitting TADA genes

| Gene    | ASD subject | Unaffected parent* | Unaffected | Odds Ratio† |
|---------|-------------|---------------------|------------|-------------|
|         |             | Unknown Inheritance | Inherited  | Tr-ASD | NT | Tr-not-ASD |          |
|        |             |                     |            |        |    |           |          |
| **q-value < 0.1** |          |                     |            |        |    |           |          |
| ANK2    | 1           |                     |            |        |    |           |          |
| ASXL3   | 1           |                     |            |        |    |           |          |
| **VIL1** | 1           | 1                   | 1          | 1.49   |    |           |          |
| **0.1 ≤ q-value < 0.3: Evidence for role in ASD** |          |                     |            |        |    |           |          |
| UTP6    | 1           |                     |            |        |    |           |          |
| DNAH10  | 1           | 1                   | 1          | 1.49   |    |           |          |
| ATP1B1  | 1           |                     |            |        |    |           |          |
| GGNBP2  | 1           |                     |            |        |    |           |          |
| NRXN1   | 2           | 1                   | 1          | 2.99   |    |           |          |
| WHSC1   | 1           |                     |            |        |    |           |          |
| **HDLBP‡** | 1           | 2                   | 1          | 1      | 1  |           | 2.24     |
| CERS4   | 1           | 1                   | 1          |        |    |           | 1.49     |
| SHANK3  | 4           |                     |            |        |    |           |          |
| IQGAP2  | 1           |                     |            |        |    |           |          |
| **0.1 ≤ q-value < 0.3: Evidence against role in ASD** |          |                     |            |        |    |           |          |
| EP400   |             |                     |            |        |    |           |          |
| **SLCO1B1‡ §** | 1           | 1                   | 1          | 1      |    |           | 0.996    |
| **SLCO1B3§** | 1           | 1                   | 2          |        |    |           | 0.37     |
| **KDM6B** |             |                     |            |        |    |           |          |

Count of deletion CNVs inferred from sequence for ASD subjects and those unaffected by ASD. Number of subjects and family status: 849 ASD subjects without family information; 1,467 ASD subjects in families; 2,766 unaffected parents; 319 unaffected siblings of ASD subjects; 373 unaffected subjects without family information. NT, parent a carrier but CNV not transmitted to affected child; Tr-ASD, transmitted to ASD subject from carrier parent; Tr-not-ASD, parent transmits a CNV to an unaffected child.

* No parents in this count were affected; seven parents in the study were affected, none carried a CNV reported in the table and these subjects did not enter the calculation.

† To compute the odds ratio we count the number of affected carriers (a), unaffected carriers (including parents) (b), affected subjects who do not have the CNV (c), and unaffected non-carriers (d). The odds ratio = (ad)/(bc).

‡ One parent transmits the CNV to an affected and unaffected offspring; to obtain the total count of controls with a CNV, subtract one.

§ Genes are adjacent in the genome (see Extended Data Fig. 4). For three subjects both genes are affected by the same CNV (1 ASD and 2 unaffected subjects).