Exome analyses reveal new autism genes in synaptic, transcriptional, and chromatin networks

Supplementary Methods

Samples

The goal of the ASC[1] is to leverage all existing and ongoing whole exome studies, as well as whole genome sequencing studies as they become available. The ASC membership continues to expand (see www.autismsequencingconsortium.org for an up-to-date list). Study design differs across the many datasets available to the ASC, however, the ASC is committed to incorporating data irrespective of design. To achieve our goal will require assessment of study design, such as appropriate controls (e.g., parents or ancestry-matched control samples), and appropriate modeling of these data. By accumulating massive data and by appropriate modeling, the expectation is enhanced power with no increase in false positive rate (see below). Moreover the ASC can ultimately compare the impact of ascertainment on yield of genetic findings. For example, we will eventually compare epidemiological samples with convenience samples; samples ascertained to be simplex with samples ascertained as multiplex; and samples receiving ADI and ADOS evaluation with 'real-world' clinical case series. While it remains an open question, ascertainment might not have as much impact as some in the field anticipate. Indeed the evidence to date indicates that such differences in study design have only modest impact on genetic yield. For example, while heritability due to common variation is perforce higher in multiplex families[2], the yield of *de novo* variation is not significantly different[3], although the sample size was small and modest differences cannot be ruled out; even the earliest studies showed etiological heterogeneity for rare high-risk alleles in multiplex families[4]. On important dimensions, such as IQ, our design and series of studies will eventually help sort out why low versus high IQ seem to have at most moderate effect on heritability[5] yet could have a larger impact on genetic yield. A complete summary of samples, ascertainment and diagnostic strategy, sequencing approach, and associated references is found in Supplementary Table 1. As indicated in the references, 175 trios published in Neale et al[3], 238 in Sanders et al[6], 189 in O'Roak et al[7], and 343 in Iossifov et al[8] were also included. All subjects provided informed consent and the research was approved by institutional human subjects boards.

Calling Variants

DNA samples were captured and sequenced as summarized in Supplementary Table 1, and the publications cited therein. The majority of unpublished (and many of the published) samples were sequenced at Broad Institute as previously described[3]. Sites made every effort to ensure that sequencing read depth exceeded 10 for 90% of the exome and 20 for 80%. SNV and small indels were called with the GATK's Unified Genotyper[9]. CNV were called with XHMM, as discussed below. Variants were annotated with SnpEff version 2.0.5 or 3 with GATK output format compatibility[10]. Samples were

segregated into family and case-control subsets. *De novo* and transmitted variants were called in family subsets. *De novo* calling is discussed below.

PLINK/SEQ (http://atgu.mgh.harvard.edu/plinkseq/) and custom code were used to filter outlier samples and select transmitted and case-control variants. We combined multiple sample sets into PLINK projects so that rare variants could be selected. Outlier filtering used individual statistics obtained with the 'pseq i-stats' command. For each project we filtered all individuals more than 4 SD from the mean of all individuals in the PLINK project in any of the following statistics: count of alternate, minor, or heterozygous genotypes; number of called variants; or genotyping rate. A trio was dropped if any member was an outlier.

Only alternate variants with a 'PASS' FILTER and 20≤MQ were considered. We selected loss of function (LoF) variants with SnpEff effect in {FRAME_SHIFT, STOP_GAINED, SPLICE_SITE_ACCEPTOR, SPLICE_SITE_DONOR}, and missense variants classified as probably damaging in at least one transcript by the PolyPhen-2 HumDiv model[11], which we refer to as missense3 (Mis3) variants. Multi-allelic variants were filtered out.

We obtained transmitted variants from family data, selecting those with alternate allele frequency ≤ 0.1%, as assessed among parents. Called genotypes in parents and children were required to have 10 ≤ DP, and 20 ≤ PL (genotype likelihood) for all non-called genotypes. Heterozygous calls were required to have had AB reference and AB alternate ≥ 0.3, and homozygous calls to have AB reference or AB alternate ≥ 0.95. A transmitted variant had to be present in the child and heterozygous in a parent. We obtained case and control variants from case-control subsets, selecting those with alternate allele frequency ≤ 0.1% among all passing subjects. These variants were required to have 10 ≤ DP and 30 ≤ GQ.

We tallied these variants into a gene by variant-type matrix for analysis by TADA. Subjects in case–control and family sample sets are distinct, so they were handled separately. Multiple variants that hit the same gene in the same subject were counted as a single variant, as follows: if a case or control subject had multiple variants in the same gene, then only a single count of the most damaging variant was used in the TADA analysis. For example, if a case had two loss of function and one missense variant in a given gene, then they were counted as having one loss of function variant.

In the affected children in trios we avoided multiple counting of *de novo* and transmitted variants separately from non-transmitted variants because the former are evidence in support of the affected gene being involved in autism and the latter support the hypothesis that the gene is not involved in autism. If an affected child had multiple *de novo* or transmitted variants in one gene, then a single count of the most damaging variant was used in the TADA analysis. In decreasing order of deleterious effect, these variants are a *de novo* loss of function, a *de novo* missense, a transmitted loss of function, and a transmitted missense. Thus, for example, if an affected child had one *de novo* missense and two transmitted loss of function variants in the same gene, then only

one *de novo* missense was counted. Similarly, the tallies of the non-transmitted variants counted only one of the most deleterious variant in a single gene in a single child.

Molecular validation

I. Validation of de novo variation

Predicted *de novo* events were validated by Sanger sequencing of the genomic DNA extracted from peripheral blood samples of the carrier and both parents. Results from targeted validation of 56 events identified as of potentially medium quality were used to enhance the *de novo* caller. High-quality variants [P(*de novo*) > 0.95] validated at an extremely high rate and were highly sensitive when validation of *de novo* mutations blind to probability estimates were performed. Overall, we tested 199 variants, and obtained a validation rate of 98.5% (196/199 variants) and a *de novo* validation rate of 97.5% (194/199 variants).

II. Validation of splice site mutations

Lymphoblastoid cells from carriers of CHD8 splice sites variants and their parents were grown in RPMI-1640 medium (Invitrogen) supplemented with 15% BenchMark™ Fetal Bovine Serum (Gemini Bio-Products) and 1% Penicillin-Streptomycin (Invitrogen). Between the 3rd and 5th passages, cells were harvested, pelleted, washed with 1X dPBS (Invitrogen) and RNA extracted using Trizol® (Invitrogen). 1 μg total RNA was used for cDNA synthesis using SuperScript® III reverse transcriptase (Invitrogen). cDNA was amplified using exon-specific primers (exons 25 and 30, #ENST00000557364) as follows: SDR151 5'-CGGAGGCCAGAATAAATGGC-3' and SDR152 5'-GTCTCCCCTTTCCCAGGTCT-3'. PCR products were purified using QIAquick PCR Purification Kit (Qiagen), subcloned into TOPO® TA Cloning® Kit for Sequencing (Invitrogen), and subjected to Sanger sequencing.

***De novo* caller**

We developed a Python application (DeNovoFinder) that expands on our previous method[3,12] for identifying *de novo* mutations from exome sequencing data. We first identified a relatively small set of candidate *de novo* mutations at highest quality (PASSing) sites defined by a standard GATK – VQSR pipeline where the child is called heterozygous and the parents both reference homozygotes. The input to this procedure is a standard VCF containing calls for all family members and a simple file describing familial relationships between individuals in the VCF.

To first insure that the child call is reliable, following our previous study we require the child PL for homozygous reference genotype is greater than 30. PL, a standard component of the VCF, represents a "Phred-scale" likelihood which is equal to -10log10(p) – such that 30 corresponds to p=0.001 for the genotype in question. In addition, we made use of the strict filters used within GATK, and further required that the proportion of non-reference reads was greater than 20% and the depth of sequence in the child was at least 10% of the mean depth of the two parents. In our previous study[3], all variants found in children matching these criteria were confirmed in validation, so we have extremely high confidence that these constitute genuine variable sites.

The major competing hypotheses amounts to comparing the likelihood that one parent is truly heterozygous and has been falsely called reference with the likelihood that this is a true *de novo* site. We therefore implemented a novel algorithm that uses population and sample allele frequency information to provide a Bayesian probability estimate that an apparent *de novo* mutation constitutes a true *de novo* as opposed to a missed heterozygous call in the parent. While the PL information from the parents provides an accurate picture of the p(data | genotype), the prior probability of heterozygous genotype must be derived from population data. To calculate this, we conservatively take the maximum allele frequency from two sources: the extensively curated NHLBI-ESP reference database and the sequenced ASC population sample from which the trio is drawn. Including both datasets permits use of both the accuracy that comes from the size of well-curated reference but insures against false low frequency estimates should there be an occasional variant missed in the reference resource but present in many copies in the current data. The probability of a site being present in a parent but absent from the reference data and all other samples in our data is simply the average number of 'singleton' sites unique to an individual (roughly 100) divided by the exome target size in bp whereas the prior probability of a *de novo* mutation is the mean number of *de novo* mutations (roughly 1) divided by the same exome size in bp.

The probabilities of the two hypotheses are then calculated using Bayes' theorem and the relative probability of p(*de novo*) / { p(*de novo*) + p(one or both parents het) } is reported as the probability of *de novo* mutation. Sites for which p(*de novo*) was estimated to be greater than .99 were considered high quality sites and constitute the systematic set of variants included in all analyses. Extensive validation of sites was performed and confirmed that 0 out of 147 high quality sites (SNVs and indels) was found to be inherited upon Sanger sequencing, confirming the validity of the $p > 0.99$ estimate. We further pursued the small fraction of overall sites that had $0.99 > p(dn) > 0.50$ and found (within a set of high quality heterozygote calls in the child) 10 of 14 were true *de novo*, the remainder inherited. This further confirmed the validity of the probability estimate and as this constituted a small but significantly real category (estimated to add ~2% true events) we added these to the analysis.

Sample specific quality control

To identify duplicate samples, discrepancy between nominal versus genetically determined sex, sample contamination, and familial relationships, we first identified a set of SNPs that would be captured by WES for most if not all samples. These variants were identified as having a non-call rate < 0.005; a minor allele frequency MAF > 0.05; marker name starting with "rs"; and reported alleles in the set A, C, G, and T. For chromosome Y, markers the non-call rate was raised to < 0.05.

To identify discrepancy between nominal versus genetically determined sex, sex was inferred from genotypes on chromosome X and Y. For chromosome X the option "--sexcheck" in PLINK was used to determine the estimated homozygosity based on X markers. To determine sex based on chromosome Y genotypes, we determined the call rate of Y markers in the high quality marker set. The appropriate cut-offs to assign male

or female sex based on these two measures varied by dataset. Therefore plots of the measures were generated and cut-offs were derived empirically from these plots. Individuals for which the reported sex did not match sex inferred from genotypes were removed from the data because they are possible sample swaps. In addition, individuals identified as possible Klinefelter syndrome (XXY males) were removed.

To identify possible sample contamination the inbreeding coefficient for each sample was calculated based on the autosomal loci using the "--het" option in PLINK. Extremely negative values for this variable indicate an overabundance of heterozygous genotype calls and this, in turn, is an indication of a mixture of DNA from different sources. Likely contaminated samples were removed from the analysis.

Data were then checked for duplicate and MZ samples using the "--genome" option in PLINK for the autosomal markers. Based on comparison of all possible pairs within the data set, pairs of individuals with estimated relationship values > 0.90 were assumed to be duplicates and one individual of each pair was removed. In a "duplicate pair", if one sample came from family data and the other from a case-control set, the family sample was retained; in all other scenarios the sample to delete was chosen as random.

For family data, pedigree information was checked for consistency with estimated relationships using the "--genome --rel-check" option in PLINK. Data were checked for consistency with the expected p(IBD=0), p(IBD=1), and p(IBD=2). When a pair of samples showed an estimate deviating by more than 0.20 from expectation, the pair was flagged. In addition, information on Mendelian error counts ("--mendel" in PLINK) was used to determine which sample was causing the problem in each pair. Problem samples were removed from the analysis.

	p(IBD=0)	p(IBD=1)	p(IBD=2)
Unrelated	1.00	0.00	0.00
Parent-offspring	0.00	1.00	0.00
Full-sibs	0.25	0.50	0.25
Half-sibs	0.50	0.50	0.00

Calling copy number variation for top genes

We screened all samples for which we had BAM files (Supplementary Data Table 1) for potential CNV, as detected by XHMM[13]. We called CNV largely as outlined previously[13]. In brief, we (a) calculated read depth with GATK, (b) determined thresholds for per-sample read depth, per-sample standard deviation, and per-target read depth, (c) ran XHMM, and (d) filtered results based on per-individual CNV count and total size, CNV XHMM score, CNV size, exon count, and minor allele frequency.

CNV were called in 8 separate batches, corresponding to groups of samples that were sequenced together as follows: ARRA Autism Sequencing Consortium case-control, Boston Autism Consortium/Finland (combined), Germany, PAGES cases, Middle Eastern, Simons Simplex Collection, Central Valley of Costa Rica, and TASC.

Seaver Autism Center Assessment Core samples were sequenced in several batches along with Central Valley of Costa Rica and TASC samples, and were included with the appropriate batches for CNV calling. XHMM thresholds for minimum and maximum mean sample read depth, maximum sample standard deviation, and minimum/maximum mean target read depth were determined on a per-batch basis. The resulting XHMM calls were then merged, creating an initial master set of 55,138 CNV. All further processing and analysis was carried out on this merged set.

The merged set was filtered as follows: 1) CNV with XHMM quality score SQ < 65 were removed, as were CNV with a predicted size CNV under 1 kb, as in prior work[13] (this somewhat conservative threshold was chosen to maximize confidence in called CNV, at the expense of some probable false negatives); 2) Individuals with an unusually high number of CNV or total length of CNV were removed (we used a cutoff of 3 SD above the mean for both measures, removing individuals with >33 CNV or >8.46 Mb total CNV); and, 3) A minor allele frequency (MAF) filter of 0.1% was applied. The frequencies of CNV were calculated across all parents, cases, and controls, excluding children to avoid over-counting transmitted CNV. We identified all regions of the exome with CNV in >0.1% of these samples, and then removed from the full set (including children) all CNV that overlapped these regions by greater than 50%.

After following the steps above, the initial set was reduced to a set of 5010 rare, high-confidence CNV, amongst which 34 CNV hitting 17 likely ASD genes ($q < 0.3$) as identified through the TADA analyses were chosen for further examination.

Transmission And *De novo* Association test (TADA)

He et al. (2013) recently published a statistical method, named TADA, for the analysis of exome sequencing data from families and case-control studies[14]. TADA performs gene-level analysis by integrating information from *de novo* mutations, inherited variants from parents and standing variants in the population in a unified statistical framework. We will briefly review the published TADA model and describe several refinements implemented in the current work.

For a given gene, we have exome sequencing data from N parent-child trios (we will describe the case-control data later). All the rare variants (defined as MAF < 0.1%) are called for each subject. A variant in a trio is then classified as *de novo*, transmitted or non-transmitted (see Figure 2 of the TADA paper). We define variants by category such as loss-of-function (LoF) or missense. All rare variants from a category are collapsed and treated as a single variant. This collapsing step allows us to combine information from multiple *de novo* mutations and from *de novo* mutation and in inherited variants in the same gene to gain power. We can then count, for each gene, the number of *de novo* mutations (within each category of variants) in N families, and similarly the number of families for which the mutant allele is transmitted or non-transmitted. He et al. (2013) assume that these counts follow Poisson distributions[14] whose rates are simple functions of the underlying parameters including mutation rates (μ), population frequency of the mutant genotype (q, note that this is about twice the mutant allele frequency) and the relative risk (RR) of the mutations (γ). Furthermore, the case-control data can be

incorporated easily: the numbers of subjects with the mutant genotype follow Poisson distributions, in a way that is similar to transmitted-nontransmitted data. The TADA model of these counts, x_d for the number of *de novo* mutations, x_1 for the number of transmitted mutant alleles plus the number of mutant alleles in cases and x_0 for the number of non-transmitted mutant alleles plus the number of mutant alleles in controls, can be summarized as:

$$x_d \sim \text{Pois}(2N\mu\gamma),\ x_1 \sim \text{Pois}(qN_1\gamma),\ x_0 \sim \text{Pois}(qN_0) \qquad (0.1)$$

where N is the number of trios, N_1 is sum of N and the number of cases, and N_0 is the sum of N and the number of controls. These parametric distributions allow us to perform likelihood-based inference. To increase the power further, we use a Bayesian strategy: incorporating the prior that some classes of variants may be more damaging and thus have higher RRs than other classes. The Bayes factor (BF) of one class of variants of a gene is defined as:

$$B = \frac{P(x \mid H_1)}{P(x \mid H_0)} = \frac{\int P(x \mid \gamma, q) P(\gamma \mid H_1) P(q \mid H_1) d\gamma dq}{\int P(x \mid q) P(q \mid H_0) dq} \qquad (0.2)$$

where H_1 and H_0 stand for the alternative model (the gene is a risk gene) and the null model ($\gamma=1$), respectively. We will defer the discussion of the prior distributions of γ and q for now. The BF at the gene level is the product of the BFs from all classes of variants. We consider two classes in our experiments: LoF and "probably damaging" missense variants predicted by PolyPhen-2[11] (denoted as Mis3), and the LoF variants receive higher prior RR (thus are weighted more heavily).

For this study, we perform three refinements of the published TADA method. We improve the model for RR, update our strategy for model parameterization, and implement a different, and much faster, way for controlling false discovery rates (FDR). These changes are described below.

I. The updated TADA model

The original TADA model assumes that for a given class of variants, e.g. LoF, a *de novo* mutation and an inherited variant in the same gene has the same RR. This is based on the assumption that all LoF mutations disrupt the gene function, thus should have a similar effect on the phenotype, regardless of the origin of the mutations. Nevertheless, more recent studies demonstrate that not all LoF mutations are alike: even a truncated protein may be partially functional, and depending on where the truncation events occur, some LoF mutations may be more damaging than the others[15]. Furthermore, because of alternative splicing, not all isoforms of a gene are expressed in a disease-related tissue (brain in our case), thus a LoF mutation in an exon not expressed in the tissue may not affect the protein function at all. By using a list of published ASD genes, we find that the estimated average RR of *de novo* LoF mutations is higher than that of inherited LoFs (see below). There is clearly an evolutionary argument supporting this observation: the most detrimental mutations in the population are likely under strong natural selection and thus tend to be eliminated, while *de novo*

mutations have not been subject to selection yet. So in the new model, we allow the RR of *de novo* mutations and that of inherited variants to be different in Equation (0.1), and the two are denoted as γ_d and γ respectively. We assume two different prior distributions for these two parameters, using Gamma distributions (the conjugate prior of the Poisson distribution):

$$\gamma_d \sim \text{Gamma}(\bar{\gamma}_d\beta_d, \beta_d),\ \gamma \sim \text{Gamma}(\bar{\gamma}\beta, \beta) \qquad (0.3)$$

where $\bar{\gamma}_d$ and $\bar{\gamma}$ are prior average RRs for *de novo* and inherited (including case-control) variants, respectively. The parameters β_d and β control the dispersion or variance of the prior distributions. The values of these parameters, particularly the prior mean RRs, roughly correspond to the weights of a type of data in the BF. So the prior mean RR of the *de novo* mutations is higher than that of the inherited ones, and the prior mean RR of LoF mutations is higher than that of missense mutations.

This change of the RR model also means we will compute the BF differently than in the original TADA model. The BF of one class of variants per gene can be factorized as the product of BFs from *de novo* and from inherited data:

$$B = \frac{P(x \mid H_1)}{P(x \mid H_0)} = \frac{P(x_d \mid H_1) \times P(x_1, x_0 \mid H_1)}{P(x_d \mid H_0) \times P(x_1, x_0 \mid H_0)} = B_d \times B_i. \qquad (0.4)$$

To compute B_d, we compute the model evidence from *de novo* data alone:

$$P(x_d \mid H_0) = \text{Pois}(x_d \mid 2N\mu) \qquad (0.5)$$

$$P(x_d \mid H_1) = \int P(x_d \mid \gamma_d) P(\gamma_d \mid H_1) = \text{NegBin}(x_d \mid \bar{\gamma}_d\beta_d, \frac{2N\mu}{\beta_d + 2N\mu}) \qquad (0.6)$$

where NegBin represents the Negative Binomial distribution. The model evidence under H_1 follows from the standard Bayesian calculation for Poisson distribution with Gamma prior. To compute B_i, we first write it as:

$$B_i = \frac{P(x_1, x_0 \mid H_1)}{P(x_1, x_0 \mid H_0)} = \frac{P(x_0 \mid H_1)}{P(x_0 \mid H_0)} \times \frac{P(x_1 \mid H_1, x_0)}{P(x_1 \mid H_0, x_0)}. \qquad (0.7)$$

As we will see, this simplifies some calculations. We next define our prior distribution for q, the population frequency of the mutant genotype:

$$q \mid H_j \sim \text{Gamma}(\rho_j, \nu_j),\ j = 0,1. \qquad (0.8)$$

Note that in theory, the prior distributions could be different for the null and alternative models. Here we take equal priors and then compute the terms in Equation (0.2):

$$P(x_0 \mid H_j) = \int P(x_0 \mid q) P(q \mid \rho_j, \nu_j) dq = \text{NegBin}(x_0 \mid \rho_j, \frac{N_0}{\nu_j + N_0}),\ j = 0,1, \qquad (0.9)$$

where NegBin(.) stands for the density function of Negative Binomial distribution. This again follows from the standard Gamma-Poisson distribution. And we have:

$$P(x_1 \mid H_j, x_0) = \int P(x_1 \mid \gamma, q) P(\gamma \mid H_j) P(q \mid H_j, x_0) d\gamma dq, \;\; j = 0,1 . \quad (0.10)$$

In this equation, the first two terms in the integrand have be defined (γ=1 under H_0), and the last term is the posterior probability of q after "seeing" the data x_0, which follows Gamma distribution:

$$q \mid H_j, x_0 \sim \text{Gamma}(\rho_j + x_0, \nu_j + N_0) \quad (0.11)$$

Under the null model, this can be computed analytically:

$$P(x_1 \mid H_0, x_0) = \text{NegBin}(x_1 | \rho_0 + x_0, \frac{N_1}{\nu_0 + N_0 + N_1}) . \quad (0.12)$$

Under the alternative model, the integration can only be solved numerically.

II. The updated parameterization scheme

For each class of variants (e.g. LoF), we have eight different parameters for the prior distributions:

$$\phi = (\bar{\gamma}_d, \beta_d, \bar{\gamma}, \beta, \rho_1, \nu_1, \rho_0, \nu_0) . \quad (0.13)$$

The meanings of each of these parameters have been defined above. We will first describe how we estimate/set the parameters related to RRs. For *de novo* LoF and Mis3 mutations, we have estimated their average RRs in He et al. (2013), and the value for the LoF mutations is about 20 and that of the Mis3 mutations is about 4.7[14]. We set β_d*=1*, following the results of that paper (the prior mean RR is a much more important parameter than β). In order to estimate the prior RR of the inherited variants, we curate a list of 26 published ASD genes, including 20 genes in Text S1 (see Section 8) of the published TADA analysis[14] and six additional ones. These six genes are: NLGN3[16], SHANK2[17,18], SHANK3[19,20], SYNGAP1, DLGAP2[21] and EPHB2[22]. We estimate the frequency difference of the LoF/Mis3 variants in these genes in cases vs. in controls. Over all 26 genes, the LoF variants are 2.3 fold enriched in cases compared with controls, so we choose $\bar{\gamma} = 2.3$. We set *β=4.0* so that (1) most of the probability mass falls in the range of values greater than 1 (if *β* is too small, there is a significant fraction of protective variants, which is unrealistic for LoF variants); (2) allow some variability of the RR: if *β* is too large, say greater than 10, then the range of RR is too narrow. Regardless of the choice, the results are not highly sensitive to this parameter. We did not find significant enrichment of Mis3 variants in cases vs. controls, so we ignore the Mis3 inherited variants in computing the gene-level BF.

In He et al., we estimate the parameters related to q by an Empirical Bayes procedure and we allow them to be different for non-risk and risk genes[14]. In general, estimating the prior parameters under H_1 is difficult because only a small fraction of genes are risk genes, and the risk genes are difficult to identify. In this work, we simplify parameter estimation by using the same prior distributions for q under H_1 and H_0, i.e.

$\rho_1 = \rho_0 = \rho, \nu_1 = \nu_0 = \nu$. To estimate ρ, ν , we note that ρ / ν is the mean prior frequency of *q*. We estimate the mean frequency across all genes (for LoF variants only, since we will not use inherited Mis3 variants) in the samples combining cases and controls, and this is about 5×10^{-4}. Next we note that ν reflects the strength of the prior distribution, or the equivalent sample size encoded in the prior. We choose ν =200 so that it is relatively small comparing with the actual samples for both LoF and Mis3 variants.

III. The updated FDR control procedure

In the published work, we determine the p-values of the gene-level BFs by a sampling procedure, and determine the FDR at a given p-value threshold by the Benjamini-Hochberg method. This procedure of FDR control is computationally expensive, and in the current work, we replace it with a Bayesian FDR control that is sometimes called "direct posterior probability" approach[23]. Specifically, for each gene, we convert its BF (B_i) to the posterior probability that the gene is a risk gene:

$$p_i = \frac{\pi B_i}{(1-\pi) + \pi B_i} \qquad (0.14)$$

where π is the prior probability of being a risk gene, or the fraction of risk genes in the genome. We choose π=0.06 in this study, corresponding to about 1000 ASD genes in 17,000 human genes. This estimate has been independently made by several groups, including ours[6,14]. Once we have the posterior probabilities, we apply the Bayesian FDR procedure as described in[23] to determine the FDR at any specified value of BF.

IV. Simulation to assess the power of TADA

We generate simulation data for all genes in the human genome (18,700 genes) using the TADA model. Specifically, at the first step, we sample an indicator variable Z_i for each gene, which follows the Bernoulli distribution with probability of success 0.06 (we estimate there are about 1,150 ASD genes, so the ratio of ASD genes is 1,150/18,700=0.06). Next, we generate the counts of *de novo* and transmitted/ not transmitted events for each gene from N trios, where N varies from 1000 to 5000. (Alternatively we assume the sample size of cases and controls is equal to the number of trios and we record only *de novo* events from trio families, i.e., N = 1000 indicates 1000 trios for *de novo* events only and 1000 cases and 1000 controls for counts of standing variation.) Note that we only specify below how to sample the parameters q and γ. Once we have these parameters, the counts from a gene are sampled from the Poisson distributions in Equation 0.1.

(1) If a gene does not affect risk (Z_i=0), its *de novo* relative risk (RR) is 1 for both LoF and Mis3 mutations. For the case-control data, we only generate simulated data for LoF variants, since Mis3 variants will not be used (see above). We sample q_i, the frequency of LoF variants, for each gene from the prior distribution Gamma(0.66, 1947) (see the last paragraph of this section about how these parameters were obtained).

(2) If a gene is a risk gene (Z_i=1), we sample its *de novo* RR for LoF and Mis3 mutations from the prior distributions Gamma(18.0, 1) and Gamma(5.4, 1) respectively (18.0 and 5.4 are prior mean RRs, see the section "Transmission And *De novo* Association test (TADA)"). For the case-control data, we first sample the RR of the LoF variants from the distribution Gamma(2.3, 4.0), as explained in the section above about model parameterization. We next determine LoF q_i according to the value of the sampled RR instead of sampling it from the prior distribution. The motivation for this step is that modeling the dependency between q and RR will likely make the simulation more realistic. In general, we expect highly penetrant mutations to be under strong natural selection, thus q of such variants will be low. Specifically, if the value of the sampled RR is γ_i, we set $q_i=\mu_i/(C\cdot \gamma_i)$, where μ_i is the mutation rate and C is a constant. The assumption here is that q follows from mutation-selection balance, and the selection coefficient is proportional to the RR. The value of C is chosen so that the equation is satisfied for an "average" gene whose γ_i=1, and q_i, μ_i are equal to the genomewide average.

Once we have the full simulated data from all genes, we apply TADA and a restricted version of TADA to the dataset. The restricted version uses only the *de novo* LoF data for the genes, and serves as a baseline to assess the performance of TADA. It is similar to the simple multiplicity test, which considers a gene significant if it has X or more *de novo* LoF mutations. We use this restricted form of TADA instead of the naïve multiplicity test because (1) unlike the multiplicity test, it does take into account the gene's mutation rates, thus discounting large and highly mutable genes; and (2) we will have a uniform strategy for FDR control. Both TADA and the restricted TADA compute a BF for each gene, and we use the FDR control procedure described above to determine the number of genes above a certain FDR threshold (we use 0.1). This number, averaged over 10 simulations, will be used as a measure of the power of the methods (Extended Data Fig. 2).

Estimating the prior distribution of q: when a gene is non-risk, we sample q of this gene (LoF) from a prior distribution and we describe how we estimate the parameters of this distribution. Let x_i be the LoF count of the i-th gene in the controls, it follows the distribution Poisson(q_iN_0), where q_i is the LoF frequency and N_0 is the sample size of controls. As an approximation, we assume all the genes are non-risk genes, then q_i follows the prior Gamma(ρ_0,v_0). It is easy to show that the marginal distribution of x_i, integrating out q_i, is negative binomial:

$$x_i \sim \mathrm{NegBin}(\rho_0, \frac{N_0}{\nu_0 + N_0}) \qquad (0.15)$$

We thus fit the entire distribution of x_i with Negative Binomial to determine the values of ρ_0 and v_0, and this gives the values of 0.66 and 1947.0, respectively.

Analyses of differences in mutation rates across genders

Inherent in both the mathematics of TADA and in our conception of the biology of ASD is the notion that there is variation between genes in their potential to contribute to ASD. We propose that for a given class of variants (e.g., LoF) some genes have very

large OR, others smaller, and still others have no effect at all. If a variant has a very large OR, that variant may effectively be a Mendelian cause of ASD and necessarily absent from control samples (parents or controls). We see this pattern for LoF variants in several of our top genes (Table 1). On the other hand, mis-annotation of variants (for example, incorrectly calling a variant LoF because the appropriate reading frame is not properly identified) can cause both a loss of power to identify ASD genes and false positives, depending on how these mis-annotated variants are distributed among samples. Finally, because we do not yet know the prior distribution of potential effects accurately, any estimation we make from a "small" dataset has the potential to over-estimate the effect size, *i.e.* show a "winner's curse." Here we use differences in mutation rates between males as females as an orthogonal approach, free from the above confounds, to estimating effect sizes.

To assess differences in the mutation rates between males and females, we analyzed all the *de novo* and transmitted variants input into the TADA model. We calculated variant rates by gender and segregated them by likelihood of autism risk: genes with q<0.1, genes with q<0.3 and all genes. P values were generated by a single-tailed test that permuted the gender labels 1,000,000 times and calculated the frequency of permutations that had a more extreme difference between male and female mutation rate than was observed.

I. A Liability Model of ASD

To model a qualitative trait, presence or absence of ASD, using standard quantitative genetics concepts, we imagine that there is an unobserved, normally distributed variable called "liability" that determines whether or not an individual is diagnosed with ASD[24]. We assume that liability, L, has mean 0 and variance 1 in the general population. Individuals with L greater than some threshold t are diagnosed with ASD and individuals with L < t are considered "typical". Under this model, the prevalence difference between males and females is viewed as a difference in thresholds for males and females. For a male to be diagnosed with ASD, his liability must be larger than t_m. For a female to be diagnosed with ASD her liability must be larger than t_f. Since ASD is more common in males than females, we conclude that $t_m < t_f$. For all that follows we will assume that the prevalence of ASD is 1 in 42 in males (implying t_m ~1.98), and the prevalence of ASD is 1 in 189 females (implying t_f ~2.56)[25].

When considering the effects of individual alleles on liability, we employ an elaboration to the standard quantitative genetics model which is sometimes called the "mixed model of inheritance"[26]. We assume that individual alleles make additive contributions to liability, so that for some allele, A_1, individuals with 0 copies of the allele have mean $-\mu$, variance 1 liability, but individuals with 1 copy have mean $\alpha-\mu$, variance 1, and individuals with 2 copies have mean $2\alpha-\mu$, variance 1 liability. Assuming Hardy-Weinberg equilibrium for genotypes, and the frequency of A_1 equaling p, $\mu = 2\alpha p^2+\alpha 2pq = 2\alpha p$. Here μ is a normalizing factor to ensure the overall population has mean liability 0.

The fundamental assumption we make for all that follows is the effect of an allele, α, is identical in males and females. The simple assumption that alleles have the same

effect on liability in males and female allows us to make substantial and sometimes surprising predictions concerning the relative distribution of risk alleles in males versus females.

Under the liability model, the penetrance of a genotype, Pr{Disease|Genotype} can be easily calculated given a and the appropriate thresholds. In particular, in males

$Pr\{Disease \mid A_1A_1\} = P_m(A_1A_1) = 1 - \Phi(t_m-(2a-m))$.

$Pr\{Disease \mid A_1A_2\} = P_m(A_1A_2) = 1 - \Phi(t_m-(a-m))$.

$Pr\{Disease \mid A_2A_2\} = P_m(A_2A_2) = 1 - \Phi(t_m+m))$.

Similarly, in females

$Pr\{Disease \mid A_1A_1\} = P_f(A_1A_1) = 1 - \Phi(t_f-(2a-m))$.

$Pr\{Disease \mid A_1A_2\} = P_f(A_1A_2) = 1 - \Phi(t_f-(a-m))$.

$Pr\{Disease \mid A_2A_2\} = P_f(A_2A_2) = 1 - \Phi(t_f+m))$,

where Φ^{-1} is an cumulative normal function. With the penetrances and allele frequency in hand, the frequency of A_1 in cases versus controls, relative risks, etc. can be calculated immediately in the usual fashion[26].

This model makes several, perhaps surprising, predictions for any allele that has the same effect on liability in males and females, because of the higher disease prevalence in males than females. First, the allele will have a higher penetrance in males than it does in females. Second, the allele will have a smaller relative risk in males than in females. Finally, the allele will be at a higher frequency in female cases than it is in male cases. The magnitude of the difference is proportional to the overall effect size α, where larger effects cause greater differences between males and females. The table below gives several illustrative examples for a rare allele (P<0.01).

Overall Odds Ratio	1.1	2	3	4	10	20	100
Penetrance Males	0.0260	0.0446	0.0640	0.0823	0.1759	0.2932	0.6719
Penetrance Females	0.0059	0.0115	0.0180	0.0247	0.0659	0.1314	0.4478
Relative Risk Males	1.09	1.92	2.80	3.68	8.75	17.00	83.92
Relative Risk Females	1.11	2.18	3.44	4.76	13.27	28.45	152.67
Frequency Female Case / Male Case	1.02	1.16	1.26	1.35	1.69	2.02	3.00

As is clear from the table, for relative risks typical of GWAS discoveries (~1.1) male/female differences are largely insignificant. However, for extremely large effect alleles, this model predicts very substantial differences between male and female penetrances and disease allele frequencies. Given the observed frequencies of various classes of variation in males versus females, we can use this model to estimate average relative risk for each class of variation.

Sub-exome enrichment

Enrichment analyses count the number of genes in common between two sets of genes, such as the set of genes with FDR < 0.3 versus a set of genes defined by a common functional role, to determine if there is greater overlap than expected by chance. Enrichment analyses can inform on this FDR list at two levels. Rigorously defined statistical excesses with independently pre-defined gene sets provide both additional confirmation of the veracity of the ASD signal, as well as, depending on the nature of the gene set, provide insight into the biological origin of that signal. Our enrichment analyses control for size and mutability of the genes in these sets, which is important given the established non-randomness of such mutation rates with respect to functional categories of genes[12].

I. Constructing sub-exomes from the literature

FMRP target lists containing 842 and 939 genes were extracted from Darnell et al[27] (Table S2A), Ascano et al[22] and Suhl et al[28]. RBFOX1/2/3 *in vivo* RNA interaction sites were extracted from Weyn-Vanhentenryck et al. 2014[29]. More specifically, we created two sub-exomes from Supplementary Table 1 and Table 6. In Table 1 we took 1048 genes mapped with significant HITS-CLIP peaks in the exon or CDS regions. In Table 6 we took 587 RBFOX gene targets with alternative splicing events predicted by their integrative model. The list of genes that are targets of both RBFOX1 and H3K4me3 was extracted from Feng et al, Table S9[30] (a total of 478 genes). Human orthologues of mouse synaptosome (152) and PSD (1080) genes were extracted from the Genes2Cognition database (http://www.genes2cognition.org). Genes with *de novo* non-synonymous mutations in schizophrenia were extracted from Fromer et al., 2014[31], also including previous studies[32,33]. Constrained genes were extracted from Samocha et al[12]. A summary is shown in Supplementary Table 4.

II. Computing empirical P value

For each comparison between 107 ASD genes with q < 0.3 and the sub-exome, we first constructed the empirical distribution by sampling the same number of genes as in the sub-exome from the gene pool without replacement for 10,000 times. We used all genes with an inferred mutation rate (as did TADA) as the gene pool, and weighted the sampling probability by normalized mutation rate (i.e., genes with larger mutation rates were more likely to be sampled). The P value was computed by counting the number of sampled gene lists that had at least as many overlapping genes as the original 107 ASD gene list, divided by the number of iterations.

HMG and ChEA networks

Histone modifier genes (HMG) were annotated using *HI*stome database[34], and chromatin remodeling factors according to the list compiled by Huang et al[35] and manually curated information from published literature. We used a manually curated transcription network to create subnetworks for TADA genes and HMGs. The background network consists of 92 transcription factors and 31,932 gene targets, with 89,933 interactions extracted from 87 publications (ChEA[27]). Only direct interactions among 107 TADA genes and HMGs were kept. For the permutation test, we selected 999 random gene sets of size 107, from the set of 18,736 background genes in the

TADA gene list (using the sampling procedure described above). For each random set, we performed the same analyses done with the original 107 TADA genes. We counted the number of genes in the resulting network for each analyses and calculated the P values based on the ranking of the original TADA list among all random draws.

DAWN (Detecting Association With Networks)

Based on the TADA scores[14] alone, only a modest number of genes are significantly associated with ASD. To identify more genes associated with ASD additional biological information can be modeled. Using a new approach called DAWN[36], Liu et al. (2014) model two kinds of data: rare variations from exome sequencing and gene co-expression in the mid-fetal prefrontal and motor-somatosensory neocortex, a critical nexus for risk[37]. Using these data, DAWN identified 160 genes that plausibly affect risk.

The DAWN algorithm casts the ensemble data as a hidden Markov random field in which the graph structure is determined by gene co-expression. It combines these interrelationships with node-specific observations, namely gene identity, expression, genetic data and the estimated effect on risk. Here we extend the DAWN approach by incorporating information about gene constraint.

I. Algorithm for network estimation

The first step of DAWN requires an estimate of the gene network, i.e., the adjacency matrix. In Liu et al. (2014) the network is estimated using a thresholded version of the correlation matrix[36]. Because the resulting network is quite dense, clusters of highly correlated genes are combined to create multigene nodes. When incorporating information about constrained genes into the model, however, it is better if each node represents a single gene. For this reason we modified the original DAWN algorithm to produce a sparse network with single-gene nodes.

We estimate the network using a sparse regression technique to select the non-zero partial correlations. Following Meinshausen and Buhlmann (2006)[38], we apply the lasso to each neighborhood regression and then construct the adjacency matrix by aggregating the non-zero partial correlation obtained for each regression. Some adjustments were made to this approach to focus on key nodes in the network based on genetic information and pairwise correlations.

To determine the right choice for the smoothing parameter we rely on the fact that many biological networks follow a power law[39].

II. The DAWN Algorithm

Let $\boldsymbol{I} = (I_1, \ldots, I_n)$ be a binary vector indicating which genes are associated with ASD. This is the "hidden state". The original DAWN model, M_0, assumes that the distribution of $\boldsymbol{I}$ follows an Ising model with density

$$P(\boldsymbol{I} = \boldsymbol{\eta}) \propto \exp\left(b\mathbf{1}^{'}\boldsymbol{\eta} + c\boldsymbol{\eta}^{'}\boldsymbol{\Omega}\boldsymbol{\eta}\right). \qquad (1)$$

To incorporate constraint information, we propose the generalized Ising model, M_1, that incorporates the directed network indicating which genes are constrained. The density function of the generalized Ising model is as follows:

$$P(\boldsymbol{I} = \boldsymbol{\eta}) \propto \exp\left(b\mathbf{1}'\boldsymbol{\eta} + c\boldsymbol{\eta}'\boldsymbol{\Omega}\boldsymbol{\eta} + d\boldsymbol{H}'\ \boldsymbol{\eta}\right) \qquad (2)$$

where $\boldsymbol{H} = (h_1, \ldots, h_n)$ is the indicator of constrained genes, and $d > 0$ reflects the enhanced probability of risk for constrained genes.

The corresponding p-values derived from TADA is converted to Z-scores ($\boldsymbol{Z}$) to obtain a measure of the evidence of disease association for each gene. It follows immediately that each of the Z-scores under the null hypothesis $I = 0$ has a standard normal distribution. We assume that under the alternative $I = 1$ the Z-scores approximately follow a shifted normal distribution. To fit M_0 we apply the iterative algorithm described in Liu et al. (2014) to estimate the parameters of the model[36]. Minor adjustments of the DAWN algorithm permit the estimation of the additional parameter d in M_1.

III. Testing the Constraint Effect

If d>0 this suggests that the constraint covariate is a predictor of risk for ASD. To test whether or not d is significantly larger than zero, we compare the observed statistic dˆ with d obtained under the null hypothesis of no association. We do so using a smoothed bootstrap simulation that involves simulating data with the same clustering of genetic signals, but without an association with the constraint sites.

To simulate Z from M_0, we first simulate the hidden states I from the distribution (1). Initial values of I are given to each node in the simulated graph, with a proportion of r being 0.5. Then, we apply a Metropolis-Hasting algorithm to update I until convergence:

- Apply the algorithm to model M_0 to obtain estimates of the model parameters.
- Using the estimated null model, simulate Iˆ by the Metropolis Hastings algorithm, then simulate Zˆ.
- Using model M_1 estimate the parameters for the simulated data.
- Iteratively conduct step (2-3) N times, then compute the empirical p-value for d by comparing the realized and simulated values. 400 simulations were performed.

IV. DAWN PPI network

The 160 genes from DAWN were seeded in the curated high confidence protein protein interaction network[3]. 95 genes were found in the largest connected component and were connected with Dijkstra's shortest path algorithm. Direct interaction between seed nodes as well as indirect interaction through immediate intermediates were recorded. The organic clustering algorithm in graph editor yEd (http://www.yworks.com/en/products_yed_about.html) was used to identify node clusters. The resulting nodes in each cluster were extracted and fed into Enrichr[40] for enrichment analysis. Enriched terms for each cluster using Mouse Genome Informatics-Mammalian Phenotype (MGI-MP), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Ontology (GO) terms were also visualized with Enrichr[40] grid feature.

Nav1.2 and Cav1.3 structural domains

The domains shown in Fig. 1c and 1d were extracted from the Protein Data Bank (PDB) and refer to Uniprot #Q99250 ($Na_v1.2$) and #Q01668 ($Ca_v1.3$). The $Na_v1.2$ EF-hand and IQ domain were extracted from #2KAV and #2KXW, respectively. For the NSCaTE motif of $Ca_v1.3$ #2LQC was used. The vestigial EF-hand domain, and the pre-IQ and IQ domains are based on $Na_v1.5$ (PDB #2KBI) and $Ca_v1.2$ (PDB #3OXQ), and adapted as described previously[41].

Supplementary Notes

Consortia

The Autism Sequencing Consortium

The authors gratefully acknowledge the support of their colleagues, including additional members of the ASC. The ASC membership continues to expand: an up-to-date list of members of the ASC can be found at www.autismsequencingconsortium.org; the site includes a list of existing, new, and provisional members, as well as more information about joining the ASC.

The DDD Study

Nadia Akawi, Saeed Al-Turki, Kirsty Ambridge, Jeffrey Barrett, Daniel Barrett, Tanya Bayzetinova, Nigel Carter, Stephen Clayton, Eve Coomber, Helen Firth, Tomas Fitzgerald, David FitzPatrick, Sebastian Gerety, Susan Gribble, Matthew Hurles, Philip Jones, Wendy Jones, Daniel King, Netravathi Krishnappa, Laura Mason, Jeremy McRae, Parker Michael, Anna Middleton, Ray Miller, Katherine Morley, Vijaya Parthiban, Elena Prigmore, Diana Rajan, Alejandro Sifrim, Adrian Tivery, Margriet van Kogelenberg, Caroline Wright

Homozygosity Mapping Collaborative for Autism

Mazhar Adli, Al Noor Centre for Children with Special Needs, Sadika Al-Awadi, Lihadh Al-Gazali, Zeinab I. Alloub, Samira Al-Saad, Muna Al-Saffar, Bulent Ataman, Soher Balkhy, A. James Barkovich, Brenda J. Barry, Laila Bastaki, Margaret Bauman, Tawfeg Ben-Omran, Nancy E. Braverman, Maria H. Chahrour, Bernard S. Chang, Haroon R. Chaudhry, Michael Coulter, Alissa M. D'Gama, Azhar Daoud, Dubai Autism Center, Valsamma Eapen, Jillian M. Felie, Stacey B. Gabriel, Generoso G. Gascon, Micheal E. Greenberg, Ellen Hanson, David A. Harmin, Asif Hashmi, Sabri Herguner, R. Sean Hill, Fuki M. Hisama, Sarn Jiralerspong, Robert M Joseph, Samir Khalil, Najwa Khuri-Bulos, Omar Kwaja, Benjamin Y. Kwan, Elaine LeClair, Elaine T. Lim, Manzil Centre for Challenged Individuals, Kyriakos Markianos, Madelena Martin, Amira Masri, Brian Meyer, Ganeshwaran H. Mochida, Eric M. Morrow, Nahit M. Mukaddes, Ramzi H. Nasir, Saima Niaz, Kazuko Okamura-Ikeda, Ozgur Oner, Jennifer N. Partlow, Annapurna Poduri, Anna Rajab, Leonard Rappaport, Jacqueline Rodriguez, Klaus Schmitz-Abe, Sharjah Autism Centre, Yiping Shen, Christine R Stevens, Joan M Stoler, Christine M. Sunu, Wen-Hann Tan, Hisaaki Taniguchi, Ahmad Teebi, Christopher A. Walsh, Janice Ware, Bai-Lin Wu, Seung-Yun Yoo, Timothy Yu

UK10K Consortium

Richard Anney, Mohammad Ayub, Anthony Bailey, Gillian Baird, Jeff Barrett, Douglas Blackwood, Patrick Bolton, Gerome Breen, David Collier, Paul Cormican, Nick Craddock, Lucy Crooks, Sarah Curran, Petr Danecek, Richard Durbin, Louise Gallagher, Jonathan Green, Hugh Gurling, Richard Holt, Chris Joyce, Ann LeCouteur, Irene Lee, Jouko Lönnqvist, Shane McCarthy, Peter McGuffin, Andrew McIntosh, Andrew McQuillin, Alison Merikangas, Anthony Monaco, Dawn Muddyman, Michael O'Donovan, Michael Owen, Aarno Palotie, Jeremy Parr, Tiina Paunio, Olli Pietilainen, Karola Rehnström, David Skuse, Jim Stalker, David St. Clair, Jaana Suvisaari, Hywel Williams

Supplementary References

1 Buxbaum, J. D. *et al.* The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron* **76**, 1052-1056 (2012).

2 Klei, L. *et al.* Common genetic variants, acting additively, are a major source of risk for autism. *Molecular autism* **3**, 9 (2012).

3 Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012).

4 Szatmari, P. *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature genetics* **39**, 319-328 (2007).

5 Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nature genetics* **46**, 881-885 (2014).

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

7 O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).

8 Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-299 (2012).

9 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498 (2011).

10 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92 (2012).

11 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249 (2010).

12 Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nature genetics* (2014).

13 Poultney, C. S. *et al.* Identification of Small Exonic CNV from Whole-Exome Sequence Data and Application to Autism Spectrum Disorder. *American journal of human genetics* **93**, 607-619 (2013).

14 He, X. *et al.* Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS genetics* **9**, e1003671 (2013).

15 Guo, Y. *et al.* Dissecting disease inheritance modes in a three-dimensional protein network challenges the "guilt-by-association" principle. *American journal of human genetics* **93**, 78-89 (2013).

16 Jamain, S. *et al.* Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nature genetics* **34**, 27-29 (2003).

17 Berkel, S. *et al.* Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. *Nature genetics* **42**, 489-491 (2010).

18 Leblond, C. S. *et al.* Genetic and functional analyses of SHANK2 mutations suggest a multiple hit model of autism spectrum disorders. *PLoS genetics* **8**, e1002521 (2012).

19 Durand, C. M. *et al.* Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nature genetics* **39**, 25-27 (2007).

20 Moessner, R. *et al.* Contribution of SHANK3 mutations to autism spectrum disorder. *American journal of human genetics* **81**, 1289-1297 (2007).

21 Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-372 (2010).

22 Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475 (2012).

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

24 Falconer, D. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann Hum Genet* **29**, 51-76 (1965).

25 http://www.cdc.gov/ncbddd/autism/data.html.

26 Morton, N. E. & MacLean, C. J. Analysis of family resemblance. 3. Complex segregation of quantitative traits. *American journal of human genetics* **26**, 489-503 (1974).
27 Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-261 (2011).
28 Suhl, J. A., Chopra, P., Anderson, B. R., Bassell, G. J. & Warren, S. T. Analysis of FMRP mRNA target datasets reveals highly associated mRNAs mediated by G-quadruplex structures formed via clustered WGGA sequences. *Human molecular genetics* (2014).
29 Weyn-Vanhentenryck, S. M. *et al.* HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell reports* **6**, 1139-1152 (2014).
30 Feng, J. *et al.* Chronic cocaine-regulated epigenomic changes in mouse nucleus accumbens. *Genome biology* **15**, R65 (2014).
31 Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014).
32 Xu, B. *et al.* De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature genetics* **44**, 1365-1369 (2012).
33 Girard, S. L. *et al.* Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature genetics* **43**, 860-863 (2011).
34 Khare, S. P. *et al.* HIstome--a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic acids research* **40**, D337-342 (2012).
35 Huang, H. T. *et al.* A network of epigenetic regulators guides developmental haematopoiesis in vivo. *Nat Cell Biol* **15**, 1516-1525 (2013).
36 Liu, L. *et al.* DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. *Molecular autism* **5**, 22 (2014).
37 Willsey, A. J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007 (2013).
38 Meinshausen, N. & Buhlmann, P. High-dimensional graphs and variable selection with the lasso. *Ann Stat* **34**, 1463-1462 (2006).
39 Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **4**, Article17 (2005).
40 Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics* **14**, 128 (2013).
41 Ben Johny, M., Yang, P. S., Bazzazi, H. & Yue, D. T. Dynamic switching of calmodulin interactions underlies Ca2+ regulation of CaV1.3 channels. *Nature communications* **4**, 1717 (2013).