A framework for the interpretation of *de novo* mutation in human disease

Kaitlin E. Samocha[1,2,3,4], Elise B. Robinson[1,2,3], Stephan J. Sanders[4,5], Christine Stevens[2,3], Aniko Sabo[7], Lauren M. McGrath[8], Jack A. Kosmicki[1,9,10], Karola Rehnström[11,12], Swapan Mallick[13], Andrew Kirby[1,2], Dennis P. Wall[9,10], Daniel G. MacArthur[1,2], Stacey B. Gabriel[2], Mark dePristo[14], Shaun M. Purcell[1,2,8,15,16,17], Aarno Palotie[8,11,12], Eric Boerwinkle[7,18], Joseph D. Buxbaum[15,16,17,19,20,21], Edwin H. Cook, Jr.[22], Richard A. Gibbs[7], Gerard D. Schellenberg[23], James S. Sutcliffe[24], Bernie Devlin[25], Kathryn Roeder[26,27], Benjamin M. Neale[1,2,3], and Mark J. Daly[1,2,3*]

Affiliations:

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, 02114

[2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA, 02142

[3]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA, 02142

[4]Program in Genetics and Genomics, Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, 02114

[5]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, 06520

[6]Department of Genetics, Yale University School of Medicine, New Haven, CT, 06520

[7]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, 77030

[8]Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA, 02114

[9]Center for Biomedical Informatics, Harvard Medical School, Boston, MA, 02115

[10]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA, 02115

[11]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

[12]Wellcome Trust Sanger Institute, Cambridge, UK

[13]Department of Genetics, Harvard Medical School, Boston, MA, 02115

[14]Synapdx, Lexington, MA, 02421

[15]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, 10029

[16]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, 10029

[17]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, 10029

[18]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, 77030

[19]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY, 10029

[20]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, 10029

[21]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, 10029

[22]Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, 60608

[23]Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104

[24]Center for Molecular Neuroscience, Vanderbilt University, Nashville, TN, 37232

[25]Department of Psychiatry, University of Pittsburgh Medical School, Pittsburgh, PA, 15213

[26]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, 15232

[27]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA, 15232

*Correspondence to: mjdaly@atgu.mgh.harvard.edu

Supplementary Note **4**
Basis of the mutational model 4

Adjustments to the mutational model 4
Depth adjustment 5
Divergence adjustment 5
Replication timing 5
Using rare variants instead of singletons 6

Definition of constrained genes 6

Evaluation of *de novo* findings 7
Global excesses of mutations 7
Number of genes with multiple mutations 7
Single genes with multiple mutations 7
Global mutation rates separated by IQ group 8

Overlap between gene lists of interest and *de novo* containing genes 9

Phenotype of individuals with *de novo* LoF mutations in FMRP targets 10

Comparison of constraint metrics 10
Comparing the power of our method to that of NS:S ratio 10
Comparison of constrained genes to the RVIS metric 11
Confirming the association between constraint and *de novo* mutations 11
Comparison of three different metrics of constraint 12

Autism spectrum disorder samples 13
Supplementary References **14**

Supplementary Tables **15**
Supplementary Table 1: Gene specific probabilities of mutation 15
Supplementary Table 2: Top constrained genes 15
Supplementary Table 3: Genes with multiple mutations 16
Supplementary Table 4: Per gene p-values for ASD cases 17
Supplementary Table 5: Per gene p-values for unaffected siblings 19
Supplementary Table 6: *De novo* mutations for high and low IQ groups 20
Supplementary Table 7: IQ groups of FMRP-I and ASD cases 21
Supplementary Table 8: Sex of FMRP-I and ASD cases 21
Supplementary Table 9: Comparison of different sets of constrained genes 22

Supplementary Figures
Supplementary Figure 1: An outline of the steps used in the model of *de novo* mutation
Supplementary Figure 2: The expected and observed fraction of genes with a *de novo* mutation in cases and controls for four gene sets of interest
Supplementary Figure 3: Correlation between constraint score and RVIS

Basis of the mutational model

The steps involved in the creation of the model are outlined in **Supplementary Figure 1**. Briefly, we determined the probability of a given base mutating into one of the three other possible bases as well as the coding impact of each possible mutation. We added probabilities across a gene to create per-gene probabilities of all mutation types under study: synonymous, missense, nonsense, and splice site.

The first, and most important, step of making a model based on sequence context is to establish the mutability of a given base. Krawczak and colleagues determined that the best context for determining the mutability of a single base is to include both the 5' and 3' bases[1]. Following the lead of other groups, we took this trinucleotide context as sufficient for determining mutability[2]. We used 1000 Genomes intergenic regions that are orthologous between humans and chimps as the basis for our mutation rate table. Across the sequence, we tallied the number of observations for each of the 64 possible trinucleotides and, for each SNP, considered the chimp allele to be ancestral and determined the trinucleotide (XY_1Z) to trinucleotide (XY_2Z) change that occurred. To determine the probability of a given trinucleotide mutating, we divided the number of mutations in that trinucleotide context by the number of occurrences of the trinucleotide. This probability is adjusted by a proportionality constant, λ, that gives the mutation rate of that trinucleotide for a single generation. The mutation rate for the given nucleotide is then proportionally divided between the three possible trinucleotides to which it could mutate. In the end, we have a mutation rate table that contains the probability of any of the 192 possible mutations.

We then use the mutation rate table and the sequence context to determine the per-gene probability of mutation based on the sequence of the gene. For a given base in the gene, the trinucleotide sequence context is determined. The probability of the middle base mutating to one of the three other bases is queried in the mutation rate table and the type of change it would create is determined. The probability of mutation is added to a running total for the type of mutation it would cause. This is repeated for the two other possible mutations for every coding base in the gene as well as the bases in the conserved splice sites for all genes in the genome. In the end, there is a per-gene probability of each type of mutation under study: synonymous, missense, nonsense, and splice site. We determine the probability of a frameshift mutation by multiplying the probability of a nonsense mutation by 1.25, the relative rate of singleton frameshift to singleton nonsense mutations found in exome sequencing data from roughly 2,000 ASD cases and controls. All probabilities of mutation are listed in **Supplementary Table 1**.

Adjustments to the mutational model

In order to evaluate the predictive value of the model of *de novo* mutation probability, we extracted the number of synonymous singletons – seen only once in the data set – found in each gene from the National Heart, Lung and Blood Institute's Exome Sequencing Project (ESP). The number of these singletons in each gene was correlated to both gene length and the probability of synonymous mutation determined by our model. While gene length alone showed a high

correlation with the number of synonymous singletons (0.835), the probability of a synonymous mutation was significantly higher (0.854, $p < 10^{-16}$).

Depth adjustment

We first investigated the role that depth of coverage could have on the predictions of mutation rates. The ability to call a *de novo* event is dependent on how well sequenced the location of the event is. Therefore, bases that are not covered at all should not contribute to the overall probability of mutation for the gene. In order to account for differences in sequencing coverage, we created a way to determine what fraction of a base's mutation probabilities should be added to the total for the gene based on the coverage. For each base, we looked up the number of trios in which all members had 10x coverage or greater and used that number to determine the appropriate discount. For bases with almost all trios having 10x coverage, the probability of mutation was not adjusted. However, as the number of trios with 10x coverage dropped, the probability of mutation was multiplied by an adjustment factor in between 0.9 and 1. To determine the endpoints of the adjustment, we compared the ratio of the observed number of synonymous singletons to the overall probability of a synonymous mutation for a high confidence set of bases to sets of bases with fewer trios passing at 10x. The depth adjusted probabilities of synonymous mutation showed a significantly greater correlation to the number of synonymous singletons in the ESP data set when compared to gene length alone (0.891, $p < 10^{-16}$).

Divergence adjustment

Divergence between humans and other primates is known to correlate with the relative number of SNPs in large regions[3]. We postulated that local divergence rates could be added to the model as a regional term that captured the local deviation from the base mutation rate. We used human-macaque divergence information to determine the divergence score – defined as the number of divergent sites over screened sites for the region containing the gene as well as 1 MB upstream and downstream – for each gene. We used linear models to determine the best equation to adjust the per-gene probabilities of mutation to incorporate the divergence score. In the end, the probability of mutation is adjusted slightly for the divergence score. For genes with no divergence information, the average divergence score is used. This, however, lead to a global increase in the predicted rate of mutation, so all probabilities of mutation were modified so that the sum of all probabilities after divergence adjustment was equal to the sum of probabilities from before the adjustment. This adjustment of predictions significantly increased the correlation with the synonymous singletons in the ESP data (0.910, $p < 10^{-16}$).

Replication timing adjustment

Replication timing has also been associated with overall mutation rate, with later replicating DNA having a higher rate of mutation[4]. We used replication timing Z scores from Koren *et al* to create a replication timing score for each gene[5]. The replication timing score is defined as the average replication timing score across the length of the gene. The replication timing score was used in linear models. It did significantly add to the mutational model (p=0.005), but had a very slight overall effect. Further investigation revealed that the model was predicting more synonymous changes as the average replication Z score increased, and thereby was already accounting for the adjustments that the replication score was adding. We did not include the replication timing adjustment in any further analyses.

Using rare variants instead of singletons

To increase power for our definition of constrained genes, we extracted the number of rare (<0.01%) synonymous variants found in each gene in the ESP data set. For this set of large counts per gene, the correlation between the number of rare synonymous variants and the gene length was 0.880. The probability of synonymous mutation as defined by our full model and the number of rare synonymous variants was 0.940. Due to the stochastic nature of small counts in the ESP data set, the maximum correlation we could achieve is 0.975, indicating that our model captured ~66% of the remaining correlation that we could achieve above gene length.

Definition of constrained genes

A traditional approach to identifying genes that appear to be under constraint is to compare the ratio of nonsynonymous to synonymous substitutions (known as the K_a/K_s or d_N/d_s). Given that the correlation between the probability of a synonymous mutation and the number of rare synonymous variants in a gene was high, we wanted to use our model to predict the number of rare missense variants as a way to evaluate genes under constraint in an approach similar to the K_a/K_s. We determined the expected number of variants by fitting a linear model based on the probability of mutation and the observed number of synonymous variants. The autosomes were fit separately from the X chromosome. The equations were applied using the probability of a missense mutation to create an expected number of rare missense variants in the ESP dataset. For both synonymous and missense variants, we created a signed Z score of the chi-squared deviation of observation from expectation. Negative values indicate more variants than expected, while positive values are tied to fewer variants observed than expected.

In order to define the set of genes that appeared to be under excessive constraint, we used three filters: (1) the predicted number of rare synonymous variants should be 5 or greater, (2) the observed number of rare synonymous variants should not be significantly lower than expectation ($p > 0.001$), and (3) the observed number of missense singletons should be significantly lower than expectation ($p < 0.001$). The reason for restricting to genes with 5 or more expected synonymous singletons is so that true deviations from expectation can be separated from deviations caused by sampling problems. Using these filters, we identified 1,003 genes that appeared to be under excessive constraint. They represent roughly 5% of the genes in the genome and are listed in **Supplementary Table 2**.

The genes in the constrained gene list show an enrichment for entries in the OMIM database, especially for entries associated with mental retardation and retinitis pigmentosa. 31% of the top 86 constrained genes – for which the observed number of missense rare variants is significant at $p < 10^{-6}$ – have entries in the Online Mendelian Inheritance in Man (OMIM) database with dominant or *de novo* inheritance patterns. None of them have recessive inheritance entries in OMIM. A comparison set was made of 111 genes for which the missense observations fell very closely around prediction ($-0.01 < Z < 0.01$). This set of genes had 2 OMIM entries (1.8%) with dominant or *de novo* inheritance and 11 (10%) with recessive inheritance.

Evaluating the *de novo* findings

In this study, we focus primarily on the *de novo* mutations found in five publications of autism spectrum disorders (ASD)[6-9] and two recent publications for intellectual disability[10,11]. We are also including 129 unpublished trios for ASD. We use the unaffected siblings sequenced as part of Iossifov *et al* as a control set[6].

Global excesses of mutations

To compare the observed rate of *de novo* mutations by mutation type to the expected rate, we summed the total probability of the given type of mutation and adjusted for the number of individuals in the study. Poisson distribution probabilities were invoked to determine the significance of the observation.

As presented in **Table 1a** in the main text, the rate of synonymous, missense, and loss-of-function (LoF) mutations in unaffected siblings match expectation fairly well. For the ASD cases, the only category that shows statistical significance is LoF. Roughly 0.09 *de novo* LoF mutations are expected per exome, and we observed ~0.13 across the 1,078 cases. This is a highly significant difference ($p=2.05x10^{-7}$). This recapitulates the excess in LoF mutations that each of the four recent studies reported, but with greater significance[6-9]. For intellectual disability, there is also significance for only the LoF category when compared to expectation (**Table 3a**, $p=6.49x10^{-7}$).

Number of genes with multiple mutations

Even though there was a global excess in LoF mutations in the ASD cases, the signal was spread over many genes, making it hard to determine which specific genes may be contributing to the etiology of ASD. One way to prioritize genes would be to focus on those genes that contain multiple *de novo* mutations. We wanted to evaluate whether there was an excess of genes that contained multiple *de novo* mutations. To do so, we simulated *de novo* events by extracting each gene's probability of mutation and then randomly drawing the expected number of *de novo* mutations based on weight (the probability). Using these simulations, we could determine an empirical p-value for the observed number of genes with multiple *de novo* mutations. Results are presented in **Supplementary Table 3** for the unaffected siblings, ASD cases, and intellectual disability cases. The "LoF+missense" category uses the combined probability of a LoF and missense mutation to evaluate genes that show two or more *de novo* mutations that are LoF, missense, or both. The lowest possible p-value is 0.001 since 1,000 simulations were run.

Both the ASD cases and the cases of intellectual disability show significantly more genes hit by multiple missense, LoF, and LoF+missense *de novo* mutations. The unaffected siblings show no such signal.

Single genes with multiple mutations

Since we have generated a per-gene probability of *de novo* mutations, we can directly evaluate genes that contain multiple *de novo* mutations for significance. To do so, each gene's probability of mutation is extracted and the predicted number of *de novo* mutations by mutation type is determined by adjusting for the number of individuals in the study. The observed and expected numbers of *de novo* mutations are compared and the Poisson is invoked to determine significance. We perform two comparisons: the LoF mutations alone and the LoF and missense

mutations together. The first comparison is only made for those genes that contain multiple LoF *de novo* mutations; the second is performed for genes that have a combination of missense and LoF *de novo* mutations. Here, we have set the significance threshold at 10^{-6} since it conservatively accounts for both the number of genes under study and the number of tests using the Bonferroni correction.

Table 2 in the main text lists the top genes for the ASD cases; **Table 4** in the main text lists all genes that have multiple *de novo* mutations in the cases of intellectual disability. **Supplementary Table 4** contains p-values for genes that had multiple *de novo* mutations in ASD cases. **Supplementary Table 5** lists all genes with multiple *de novo* mutations in the unaffected siblings.

Global mutation rates separated by IQ group

Due to the significant role of *de novo* mutation in intellectual disability, we wanted to investigate the overall rates of mutations for those ASD cases without intellectual impairments. Several intelligence tests were used to assess proband IQ across testing sites. The IQ analyses presented here include individuals whose IQ was measured using one of four standardized, commonly used tests to evaluate intelligence in children: the WISC-IV[12], the WASI[13], the WPPSI-III (preschool and primary school age)[14], and the DAS (early years and school age)[15]. These tests provide comparable assessments of full scale intelligence (g), using both verbal and nonverbal assessments[16]. Children who did not complete one of these four tests (n=95, 10.0%) were treated as missing without attempt. Probands who are missing IQ without attempt include those who were given an IQ test that does not assess intelligence comparably (n=78, 8.2%), specifically the Mullen Scales of Early Learning or the Leiter International performance scale, which are strongly weighted towards nonverbal assessment[17,18].

We had access to phenotypic information for 954 of the sequenced probands. Of these, 859 had taken an IQ test that could be compared to other tests. We removed those individuals that had a 30-point or greater difference between their verbal and nonverbal IQs to avoid inclusion of excess measurement error or learning disabilities. Verbal and nonverbal IQ were correlated strongly with each other ($r=0.70$, $p < 0.0001$) as well as with the full scale IQ score (verbal IQ: $r=0.89$, $p < 0.0001$; nonverbal IQ: $r=0.93$, $p < 0.0001$). We separated the remaining 801 probands into those with and without measured IQs above statistical average. It is common for individuals affected with ASDs to be unable to complete or be scored on an IQ test; this was the case for 14.3% (n=115) of probands for whom a test was attempted in the Simons sample. In the Simons Simplex Collection, probands who attempted to complete an eligible IQ test, but did receive a score, had significantly lower scores on the Vineland Scales of Adaptive Behavior (IQ test scored mean = 76.0, IQ test not scored mean = 60.3; $t = 15.9$, $p < 0.0001$). A Vineland composite standard score of 60 reflects adaptive behavior (overall functioning and self care skills) scores nearly three standard deviations below the mean, or approximately in the lowest 1% of the general population, controlling for age. As the inability to complete an IQ test is associated with case severity, we were specifically interested in estimating the de novo rate among individuals with both IQ above the general population mean and the behavioral capability to complete an IQ test—both indicators of higher functioning ASDs. The observed and expected *de novo* mutations per exome are listed in **Supplementary Table 6**. The individuals with full scale IQ $\geq$ 100 matched expectation for *de novo* mutations per exome. Those individuals without measured IQs over 100, on the other hand, showed a global excess in *de novo* LoF mutations. The results were

similar when verbal and nonverbal IQ were analyzed separately (**Supplementary Table 6c**). There was no excess of *de novo* LOF mutation in individuals with verbal (p=0.19) or nonverbal (p=0.48) IQ greater than 100.

Overlap between gene sets of interest and *de novo* containing genes

A number of gene sets have been proposed as relevant to autism or descriptive of an ASD biochemical pathway. Given the global excess of *de novo* LoF mutations, we wanted to evaluate whether or not the list of genes that contain such mutations overlap more than expected with several of the proposed gene sets.

In order to determine the significance of any observed overlap between a gene set of interest and the list of genes that contain *de novo* mutations, we first determine the total probability of mutation for all genes on the gene set of interest. The set total is compared to the total probability of mutation for all genes. This percentage becomes the expected overlap of *de novo* mutations with the gene set. Using the expected overlap and the number of mutations on the *de novo* list, we evaluate the observed overlap between the *de novo* list and the gene set of interest by invoking the binomial. All p-values are one-tailed. The *de novo* mutation list is broken down by mutation type (LoF, missense, and synonymous), as are the probabilities of mutation for the gene set of interest.

We evaluated the overlap between three *de novo* lists and four separate gene sets of interest (**Supplementary Figure 2**). The unaffected list comes from 647 unaffected siblings and control individuals that were sequenced across many studies[6,8,9,11,19]. The ASD cases and intellectual disability (ID) cases are the 1,078 and 151, respectively, discussed throughout this study[6-11,20]. Significance was conservatively set at 0.01. In the figure, the asterisk (*) indicates a p-value of less than 0.01, while the double asterisk (**) indicates a p-value of less than 10^{-4}. The gene sets of interest are a set of genes reported as disrupted in individuals with ASD or autistic features (Betancur)[21], the set of targets of FMRP identified by Darnell and colleagues (FMRP)[22], the set of significantly constrained genes that we defined earlier (Constrained), and the set of FMRP targets that are also constrained (Constrained FMRP).

The *de novo* mutations found in the unaffected individuals showed no significant overlaps with any of the input sets for any mutation type. For the *de novo* mutations found in the ASD cases, the overlaps for synonymous and missense mutations show no significant enrichments on any of the input sets. The *de novo* LoF mutations found in ASD cases, however, show significant overlap with the FMRP interactors, the constrained genes, and the set of constrained FMRP interactors ($p < 0.0001$ for all, 2.3 to 3.9-fold enrichment). The *de novo* LoF mutations in cases of intellectual disability have significant overlaps with all tested input sets ($p < 0.0001$ for all, 3.5 to 18.7-fold enrichment). The missense *de novo* mutations for intellectual disability are also significantly enriched on the set of FMRP interactors, the constrained genes, and the set of constrained FMRP interactors ($p < 0.0001$ for all, 2.6 to 5.5-fold enrichment). The *de novo* synonymous mutations in the intellectual disability cases have no significant enrichment with any of the gene sets of interest.

Phenotype of individuals with *de novo* LoF mutations in FMRP targets

Across the 1,078 individuals with ASD, there were 35 *de novo* LoF mutations in targets of FMRP spread across 34 individuals (referred to as FMRP-I here)[22]. For those individuals for which we had access to phenotypic information, we extracted IQ and sex. We found that the FMRP-I group had significantly fewer individuals with IQ ≥ 100 than the rest of the sample set (**Supplementary Table 7a**, Fisher's exact $p=4.01x10^{-4}$). As before, individuals who started an IQ test but were not given an IQ score due to being severely impaired are included in the IQ < 100 group. To ensure that the association was not driven by those probands with attempted but missing IQ values, we also tested the association using only those individuals with estimated full scale IQ scores (**Supplementary Table 7b**, Fisher's exact p=0.0021). The FMRP-I group also had a reduced male bias. Where the whole set of individuals is ~80% male, the FMRP-I group is only ~59%, which is a significant difference (**Supplementary Table 8**, Chi-square p=0.02).

Comparison of constraint metrics

Comparing the power of our method to that of NS:S ratio

The ratio of nonsynonymous (NS) substitutions per NS site to synonymous (S) substitutions per S site in a gene has been often used to determine if that gene has evidence of selection acting on it. A high NS:S ratio would indicate positive selection, while a low NS:S ratio would be evidence for purifying selection. Theoretically, our method of comparing observed NS variants to expectation should achieve greater statistical power than the NS:S comparison. To support this claim, we used the number of NS and S rare variants (minor allele frequency < 0.01%) found in the NHLBI's Exome Sequencing Project (ESP) dataset and determined each gene's deviation in terms of their ratio of S to NS sites compared to the genome-wide average.

We removed the 134 genes where the observed synonymous and nonsynonymous rates were both significantly elevated or significantly depressed from expectation as determined by our model (both $p < 0.001$). These poorly sequenced or mapped genes – as mentioned in the main text – were also removed from our analysis to define constrained genes. We then identified the remaining genes that were as deviant from the genome-wide average as the constrained genes we defined with our model were from expectation ($p < 0.001$). Compared to the 1,003 genes defined as constrained by our model, this approach only identified 377 genes that showed evidence of purifying selection, 237 (~63%) of which were also identified as constrained by our method. Included in the 766 genes considered constrained only by our metric were a number of genes – the top ten of which include *RYR2*, *MLL*, *MLL2*, and *SYNGAP1* – that have already been established as causes of autosomal or X-linked dominant forms of Mendelian disease (OMIM enrichment $p=5 \times 10^{-4}$).

Since our metric was able to identify more genes that showed evidence of selective constraint, and especially since some of those are known to be causes of Mendelian disease, we conclude that our method of identifying constrained genes adds substantial power to the traditional approach and is an appropriate metric.

Comparison of constrained genes to the RVIS metric

Recently, Petrovski *et al* published a similar method to search for genes that appeared to be intolerant of mutations[23]. Their method evaluates the shift in the allele frequency spectrum of variants identified in genes in the ESP dataset to identify genes that have more rare variation. Specifically, the number of common nonsynonymous variants found in each gene was regressed against the total number of variants to determine the intolerance score. Genes with an unusually high ratio of rare to common variation are more likely to be intolerant of mutations and are assigned a lower residual variation intolerance score (RVIS). This approach is orthogonal to our metric of constraint since we search for a deficiency of rare nonsynonymous variation.

We took the intersection between the two datasets to compare our metric with the scores provided in Petrovski *et al*[23]. This process eliminated some of the genes considered constrained by our metric, leaving 827 genes. Their score yielded a similar number of constrained genes (n=842), which were defined as those genes with a residual variation intolerance score in the top 5%. 231 genes were considered constrained by both metrics, which is far greater than expected (0.25%, ~41 genes). Using a Wilcoxon rank-sum test, we found that the genes defined as constrained by our metric had significantly lower (more intolerant) RVIS values ($p < 10^{-16}$). Similarly, the top 5% RVIS genes had significantly higher constraint scores (Wilcoxon rank-sum, $p < 10^{-16}$). We found a correlation of -0.35 between the two scores of constraint, which is illustrated in **Supplementary Figure 3**.

Confirming the association between constraint and *de novo* mutations

The power to determine if a gene is significantly constrained relies on gene size. As mentioned above, genes where we predicted fewer than 5 rare synonymous variants had to be removed. In order to confirm that the association we found between constraint and the *de novo* LoF mutations identified in ASD patients, we first investigated the relationship between constraint and the *de novo* mutations found in unaffected individuals. As depicted in **Supplementary Figure 2a**, we found no enrichment of *de novo* LoF mutations from unaffected individuals in constrained genes. Additionally, we included gene length as a covariate while performing regressions of ASD *de novo* LoF genes on constraint and found that the association remained. We also took the largest 10% of genes and performed the regression again; constraint was still significant, but the gene length – when included as a covariate – showed no association.

Our method of determining constraint generates the number of rare missense variants that are expected to be in each gene. As an alternative metric to constraint, we also evaluated the fraction of missense variation that was not seen, a metric that is completely independent of gene size. We found that, in a linear regression, the fraction of missing missense variation was significantly able to predict whether a gene was haploinsufficient ($p=2.13 \times 10^{-12}$).

For our final analyses to confirm that our enrichment analysis was not biased towards bigger genes, we created a list of the largest 5% of genes and queried the *de novo* loss-of-function mutations identified in unaffected individuals. We expect that there should be no significant relationship between *de novo* LoF mutations in unaffected individuals and these large genes. When we use a simple logistic regression to explain the *de novo* LoF genes in unaffected individuals, we find an odds ratio (OR) of about 5.5, which describes a highly significant

enrichment of big genes. Our method of determining enrichment, however, accounts for the expected mutation rate of each gene – thereby inherently incorporating gene size – and shows this set of mutations is not actually "enriched" at all (p=0.425; fold enrichment/OR = 1.1). These *de novo* LoF mutations in unaffected individuals are occurring in exactly the chance proportion they should be in larger genes. We therefore conclude that the enrichment analysis central to our interpretation of ASD events is not affected by gene lists being non-random with respect to size.

Comparison of three different metrics of constraint

Our metric is one way of searching for genes that appear to be relatively intolerant of mutations in the human population. One approach is the residual variation intolerance score (RVIS) created by Petrovski and colleagues[23], which evaluates the relative excess of rare variants to common ones in genes. Since Petrovski *et al* did not define a list of intolerant genes in their paper, we defined such a list by taking the top 5.5% most intolerant genes according to their metric. 5.5% was selected since that is the percentage of genes that we define as constrained using our metric. An additional alternative comes from Bustamante *et al*, who used both fixed and polymorphic synonymous and nonsynonymous sites to find genes that appear to be affected by selection, including 813 loci that appeared to be under negative selection[24].

We sought to compare both our constraint score and list of constrained genes with the results of these other approaches. To do this, we focused on the ability to predict known haploinsufficient genes (as defined in OMIM) and the enrichment of these genes with *de novo* LoF mutations identified in ASD patients. Our results are summarized in **Supplementary Table 9**. For the quantitative metrics (our constraint score and the RVIS metric), we performed a linear regression between haploinsufficient genes and the score with gene size as a covariate. While both metrics have significant predictive ability, our constraint score outperforms RVIS slightly (t-value = 10.011 for constraint, -9.561 for RVIS). For the list-based comparison, we used a logistic regression with gene length as a covariate. In this comparison, the top 5.5% intolerant genes according to RVIS had an odds ratio (OR) of ~5.5, while the constrained gene set that we defined had an OR of 4.9, both of which were significant. The genes identified by Bustamante and colleagues showed no significance (**Supplementary Table 9a**).

We also evaluated the fraction of these different sets of constrained genes that contained a *de novo* LoF in ASD cases. Our method, as explained above, determines the fraction of constrained genes that are expected to contain a *de novo* mutation by chance. We then evaluate the observed fraction and can determine both the fold enrichment and significance. When we evaluated the three previously mentioned lists of genes – our constrained, top 5.5% intolerant genes using RVIS[23], and the loci identified by Bustamante[24] – we found that our list of constrained genes had the greatest fold enrichment of genes that contained a *de novo* LoF in ASD cases ($p=3.58\times10^{-6}$; **Supplementary Table 9b**). The top 5.5% of genes identified using RVIS also performed well (fold enrichment of 1.9, $p=5.36\times10^{-5}$), but the loci from Bustamante *et al* showed no significant enrichment.

Autism spectrum disorder samples

As mentioned in the main text, almost all of the ASD samples used in this study were previously published: Neale *et al* (n=175 trios)[7], Iossifov *et al* (n=343)[6], O'Roak *et al* 2011 (n=17)[20], O'Roak *et al* 2012 (n=189)[8], and Sanders *et al* (n=225)[9]. Updated *de* novo calls were used for 400 of these trios[7,9]. In addition, we included 129 unpublished trios. Autism Consortium samples (n=78 trios) were collected in Boston under IRB approval from Harvard Medical School, Massachusetts General Hospital, Children's Hospital Boston, Tufts-NEMC, Boston University Medical Center with ADI and ADOS assessment. Finnish autism samples (n=51 trios) were collected under IRB approval at University of Helsinki with ADI and ADOS assessment and consented for autism research only. In both studies, all participants gave written informed consent, though as autism is classified as a childhood disorder, many subjects are children with informed consent provided by parents or guardians.

Supplementary References

1. Krawczak, M., Ball, E.V. & Cooper, D.N. Neighboring-Nucleotide Effects on the Rates of Germ-Line Single-Base-Pair Substitution in Human Genes. *The American Journal of Human Genetics* **63**, 474-488 (1998).
2. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies. *The American Journal of Human Genetics* **80**, 727-739 (2007).
3. Hellmann, I. *et al.* Why do human diversity levels vary at a megabase scale? *Genome Research* **15**, 1222-1231 (2005).
4. Stamatoyannopoulos, J.A. *et al.* Human mutation rate associated with DNA replication timing. *Nature genetics* **41**, 393-395 (2009).
5. Koren, A. *et al.* Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. *The American Journal of Human Genetics* **91**, 1033-1040 (2012).
6. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74**, 285-299 (2012).
7. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-245 (2012).
8. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).
9. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-241 (2012).
10. de Ligt, J. *et al.* Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *New England Journal of Medicine* **367**, 1921-1929 (2012).
11. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet* **380**, 1674-1682 (2012).
12. Weschler, D. *Weschler Intelligence Scale for Children--4th Edition (WISC-IV).* (Harcourt Assessment, San Antonio, Texas, 2003).
13. Weschler, D. *Weschler Abbreviated Scale of Intelligence (WASI)*, (Harcourt Assessment, San Antonio, Texas, 1997).
14. Weschler, D. *Weschler Primary and Preschool Scale of Intelligence--Third Edition*, (Harcourt Assessment, San Antonio, Texas, 2002).
15. Elliott, C.D. *DAS Administration and Scoring Manual*, (The Psychological Corporation, San Antonio, Texas, 1990).
16. Elliott, C.D. *DAS Introductory and Technical Handbook*, (The Psychological Corporation, San Antonio, Texas, 1990).
17. Mullen, E.M. *Mullen Scales of Early Learning*, (American Guidance Service Inc., Circle Pines, MN, 1995).
18. Roid, G.H. & Miller, L.J. *Leiter International Performance Scale- Revised*, (Stoelting Co., Wood Dale, Illinois, 1997).
19. Xu, B. *et al.* De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature genetics* **44**, 1365-1369 (2012).
20. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics* **43**, 585-9 (2011).
21. Betancur, C. Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. *Brain Research* **1380**, 42-77 (2011).
22. Darnell, J.C. *et al.* FMRP Stalls Ribosomal Translocation on mRNAs Linked to Synaptic Function and Autism. *Cell* **146**, 247-261 (2011).
23. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet* **9**, e1003709 (2013).
24. Bustamante, C.D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153-1157 (2005).

Supplementary Tables

Supplementary Table 1: Gene specific probabilities of mutation. The per-gene probabilities of mutation are listed for each gene (transcript specified) by mutation type. Probabilities of mutation are given per chromosome and have been transformed by $\log_{10}$. "NA" is listed when there is no probability of mutation due usually to low coverage.

Please see supplementary Excel document.

Supplementary Table 2. Top 1,003 constrained genes. The gene specific information listed includes transcript and identifier, chromosome, transcription start position, number of coding bases, probabilities of a synonymous and missense mutation (given per chromosome), the number of observed and expected synonymous and missense variants, the signed Z scores for the deviation for both synonymous and missense variants, and the ratio of missing missense variation ("ratio_missing").

Please see supplementary Excel document.

a) Unaffected siblings

Mutation Type	Observed genes with 2+ DNMs	Average expected genes with 2+ DNMs	Maximum expected genes with 2+ DNMs	p-value
Synonymous	0	0.5	2	1.0
Missense	5	2.5	5	0.1049
LoF	0	0.04	0	1.0
LoF+missense	6	3	6	0.0779

b) ASD cases

Mutation Type	Observed genes with 2+ DNMs	Average expected genes with 2+ DNMs	Maximum expected genes with 2+ DNMs	p-value
Synonymous	4	3.8	7	0.5186
Missense	33	21.4	29	0.0070
LoF	6	0.5	2	< 0.001
LoF+missense	48	27.2	36	< 0.001

c) Intellectual disability cases

Mutation Type	Observed genes with 2+ DNMs	Average expected genes with 2+ DNMs	Maximum expected genes with 2+ DNMs	p-value
Synonymous	1	0.09	1	0.0879
Missense	3	0.5	2	0.0090
LoF	2	0.01	0	< 0.001
LoF+missense	6	0.6	2	< 0.001

Supplementary Table 3. Evaluating genes with multiple *de novo* mutations. The observed number of genes with two or more *de novo* mutations (DNMs) in unaffected siblings (a), autism spectrum disorder (ASD) cases (b), and intellectual disability cases (c). The average and maximum expected number of such genes were determined by simulation.

Supplementary Table 4. Significance of genes with multiple *de novo* mutations (DNMs) in ASD cases

Gene	Mutations	# LoF	# Missense	# DNMs Expected	p-value	Test
DYRK1A	nonsense, frameshift, splice	3	0	0.0072	6.15E-08	LoF
SCN2A	frameshift, missense, missense, nonsense, nonsense	3	2	0.0177	9.20E-07	LoF
CHD8	missense, frameshift, nonsense, splice	3	1	0.0221	1.76E-06	LoF
KATNAL2	splice, splice	2	0	0.0049	1.19E-05	LoF
POGZ	frameshift, frameshift	2	0	0.0134	8.93E-05	LoF
ARID1B	frameshift, frameshift	2	0	0.0178	1.57E-04	LoF
SCN2A	frameshift, missense, missense, nonsense, nonsense	3	2	0.1334	3.15E-07	LoF+mis
CHD8	missense, frameshift, nonsense, splice	3	1	0.1724	3.20E-05	LoF+mis
SUV420H1	splice, missense, missense	1	2	0.0602	3.48E-05	LoF+mis
PLEKHA8	missense, missense	0	2	0.0302	4.46E-04	LoF+mis
TUBA1A	missense, missense	0	2	0.0338	5.59E-04	LoF+mis
SLCO1C1	missense, missense	0	2	0.0394	7.55E-04	LoF+mis
NTNG1	missense, missense	0	2	0.0413	8.29E-04	LoF+mis
TSNARE1	missense, missense	0	2	0.0498	1.20E-03	LoF+mis
TBR1	missense, frameshift	1	1	0.0541	1.41E-03	LoF+mis
MEGF11	missense, missense	0	2	0.0552	1.47E-03	LoF+mis
KRBA1	missense, missense	0	2	0.0642	1.98E-03	LoF+mis
SRBD1	missense, missense	0	2	0.0645	1.99E-03	LoF+mis
KIRREL3	missense, missense	0	2	0.0652	2.03E-03	LoF+mis
NR3C2	nonsense, missense	1	1	0.0655	2.05E-03	LoF+mis
UBE3C	missense, missense	0	2	0.0775	2.85E-03	LoF+mis
AGAP2	missense, missense	0	2	0.0825	3.22E-03	LoF+mis
ABCA13	missense, missense, missense	0	3	0.2890	3.24E-03	LoF+mis
ADCY5	missense, missense	0	2	0.1098	5.61E-03	LoF+mis
KIAA0182	missense, missense	0	2	0.1114	5.76E-03	LoF+mis
ZNF423	missense, missense	0	2	0.1131	5.94E-03	LoF+mis
ZNF638	frameshift, missense	1	1	0.1212	6.78E-03	LoF+mis
SCN1A	missense, missense	0	2	0.1352	8.36E-03	LoF+mis
LAMB2	missense, missense	0	2	0.1604	1.16E-02	LoF+mis
MYO7B	missense, missense	0	2	0.1616	1.17E-02	LoF+mis
KIAA0100	nonsense, missense	1	1	0.1619	1.18E-02	LoF+mis
PLXNB1	frameshift, missense	1	1	0.1718	1.32E-02	LoF+mis
CACNA1D	missense, missense	0	2	0.1732	1.34E-02	LoF+mis
ZFYVE26	frameshift, missense	1	1	0.1753	1.37E-02	LoF+mis
SBF1	missense, missense	0	2	0.1808	1.45E-02	LoF+mis

Supplementary Table 4 Continued. Significance of genes with multiple *de novo* mutations (DNMs) in ASD cases

Gene	Mutations	# LoF	# Missense	# DNMs Expected	p-value	Test
BRCA2	missense, missense	0	2	0.1928	1.64E-02	LoF+mis
TRIO	missense, missense	0	2	0.2374	2.41E-02	LoF+mis
ALMS1	missense, missense	0	2	0.2422	2.50E-02	LoF+mis
RELN	nonsense, missense	1	1	0.2429	2.51E-02	LoF+mis
ANK2	missense, nonsense	1	1	0.2591	2.83E-02	LoF+mis
MLL3	nonsense, missense	1	1	0.3159	4.05E-02	LoF+mis
DNAH5	frameshift, missense	1	1	0.3219	4.19E-02	LoF+mis
FAT1	missense, missense	0	2	0.3343	4.49E-02	LoF+mis
GPR98	missense, missense	0	2	0.3761	5.53E-02	LoF+mis
AHNAK2	missense, missense	0	2	0.4172	6.62E-02	LoF+mis
SYNE1	missense, missense	0	2	0.5931	1.20E-01	LoF+mis
TTN	missense, missense, missense, missense	0	4	2.1947	1.80E-01	LoF+mis
MUC5AC	missense, missense	0	2	.	.	LoF+mis
RFX8	missense, missense	0	2	.	.	LoF+mis
EFCAB8	missense, missense	0	2	.	.	LoF+mis

Supplementary Table 4. Significance of genes with multiple *de novo* mutations (DNMs) in autism spectrum disorder (ASD) cases. LoF mutations include nonsense, frameshift, and splice site-disrupting mutations. "# LoF Expected" refers to the expected number of *de novo* LoF mutations based on the probability of mutation for the gene as determined by our model. The genome-wide significance threshold is $1x10^{-6}$. "." = no data available.

Gene	Mutations	# LoF	# Missense	# DNMs Expected	p-value	Test
CSNK1G3	missense, frameshift	1	1	0.0098	4.78E-05	LoF+mis
UGT2B4	missense, missense	0	2	0.0102	5.12E-05	LoF+mis
USP34	missense, missense	0	2	0.0717	2.45E-03	LoF+mis
AHNAK2	missense, missense	0	2	0.1327	8.07E-03	LoF+mis
SYNE2	missense, missense	0	2	0.1369	8.56E-03	LoF+mis
TTN	missense, missense	0	2	0.6983	1.55E-01	LoF+mis

Supplementary Table 5. Significance of specific genes with multiple *de novo* mutations (DNMs) in unaffected siblings. LoF mutations include nonsense, frameshift, and splice site-disrupting mutations. "# LoF Expected" refers to the expected number of *de novo* LoF mutations based on the probability of mutation for the gene as determined by our model. The genome-wide significance threshold is $1x10^{-6}$.

a)

Mutation Type	Full Scale IQ scored above 100			
	Observed events per exome	Expected events per exome	p-value	
Synonymous	0.24	0.27	0.2346	2-tailed
Missense	0.66	0.62	0.4736	2-tailed
Loss-of-Function	0.08	0.09	0.5867	1-tailed

n = 229

b)

Mutation Type	Full Scale IQ not scored above 100			
	Observed events per exome	Expected events per exome	p-value	
Synonymous	0.22	0.27	0.0123	2-tailed
Missense	0.62	0.62	0.9946	2-tailed
Loss-of-Function	0.17	0.09	1.17E-10	1-tailed

n = 572

c)

Phenotypic Group	Number of samples	Observed *de novo* LoF events per exome	p-value
Verbal IQ ≥ 100	242	0.10	0.1903
Verbal IQ not scored above 100	712	0.15	2.43E-08
Nonverbal IQ ≥ 100	276	0.09	0.4829
Nonverbal IQ not scored above 100	678	0.16	1.09E-09

Supplementary Table 6. Investigating the rate of *de novo* mutation as a function of IQ. (a) The observed and expected rate of *de novo* mutations by mutation class for the autism spectrum disorder cases with full scale IQ ≥ 100. (b) The observed and expected rate of *de novo* mutations by mutation class for the autism spectrum disorder cases that did not have a full scale IQ above 100. (c) The observed rate of *de novo* loss-of-function (LoF) mutations split by verbal IQ and nonverbal IQ.

a) IQ Attempted but unscored individuals included

	FMRP-I	Rest of Cases
IQ ≥ 100	1 (3%)	254 (31%)
IQ not above100	29 (97%)	575 (69%)

Fisher's exact p-value = 4.01×10^{-4}

b) Only scored individuals

	FMRP-I	Rest of Cases
IQ ≥ 100	1 (5%)	254 (35%)
IQ not above100	20 (95%)	469 (65%)

Fisher's exact p-value = 0.0021

Supplementary Table 7. The number (and percentage) of individuals that have an IQ ≥ 100 or an IQ not scored above 100 split by containing a *de novo* loss-of-function mutation in a target of FMRP (FMRP-I) or not ("Rest of Cases"). In (a), individuals who started an IQ test but were not given an IQ score are included. Only individuals with IQ scores are included in (b).

	FMRP-I	Rest of Cases
Male	19 (63%)	658 (80%)
Female	11 (37%)	163 (20%)

Chi-square p-value = 0.02

Supplementary Table 8. The number (and percentage) of individuals that are male and female split by containing a *de novo* loss-of-function mutation in a target of FMRP (FMRP-I) or not ("Rest of Cases").

a) Linear and logistic regressions

		Quantitative Scores			List-Based		
		Constraint score	RVIS		Top Constrained	Top RVIS	Bustamante
OMIM Haplo-insufficiency	t-value	10.011	-9.561	OR	4.909	5.490	1.307
	p-value	$< 10^{-16}$	$< 10^{-16}$	p-value	$< 10^{-16}$	$< 10^{-16}$	0.191

b) Enrichment of genes with those containing a *de novo* LoF in ASD patients

		Top Constrained	Top RVIS	Bustamante
ASD *de novo* LoF	Fold enrichment	2.282	1.904	0.836
	p-value	3.58×10^{-6}	5.36×10^{-5}	0.718

Supplementary Table 9. Comparison of the predictive ability of different sets of constrained genes for known haploinsufficient genes and those disrupted by a *de novo* LoF mutation in ASD patients. In (a), the ability of both constraint scores and lists of constrained genes were tested for their ability to predict known haploinsufficient genes, as listed in OMIM. The quantitative scores (constraint and RVIS[23]) were used in a linear regression with gene size added as a covariate. The gene lists (constrained, top 5.5% most intolerant genes using RVIS[23], and the genes identified in Bustamante *et al*[24]) were evaluated with a logistic regression with gene size as a covariate. In (b), the three gene lists were evaluated for their enrichment of *de novo* LoF mutations identified in ASD patients. To do this, the expected fraction of constrained genes to contain one of these *de novo* mutations was determined and then used to establish the fold enrichment and significance of the observed fraction.