nature genetics

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

Kaitlin E Samocha[1–4], Elise B Robinson[1–3], Stephan J Sanders[5,6], 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[1,2,8,15–17], Aarno Palotie[8,11,12], Eric Boerwinkle[7,18], Joseph D Buxbaum[15–17,19–21], 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[1–3] & Mark J Daly[1–3]

Spontaneously arising (*de novo*) mutations have an important role in medical genetics. For diseases with extensive locus heterogeneity, such as autism spectrum disorders (ASDs), the signal from *de novo* mutations is distributed across many genes, making it difficult to distinguish disease-relevant mutations from background variation. Here we provide a statistical framework for the analysis of excesses in *de novo* mutation per gene and gene set by calibrating a model of *de novo* mutation. We applied this framework to *de novo* mutations collected from 1,078 ASD family trios, and, whereas we affirmed a significant role for loss-of-function mutations, we found no excess of *de novo* loss-of-function mutations in cases with IQ above 100, suggesting that the role of *de novo* mutations in ASDs might reside in fundamental neurodevelopmental processes. We also used our model to identify ~1,000 genes that are significantly lacking in functional coding variation in non-ASD samples and are enriched for *de novo* loss-of-function mutations identified in ASD cases.

Exome sequencing has enabled the identification of *de novo* (newly arising) mutations and has already been effectively used to identify causal variants in rare mendelian diseases. In the case of Kabuki syndrome, the observation of a *de novo* mutation in *MLL2* (*KMT2D*) in 9 of the 10 cases analyzed strongly implicated loss of *MLL2* function as causal[1]. The conclusion that *MLL2* is important in Kabuki syndrome etiology based on the *de novo* mutation findings relies upon the unlikely accumulation of independent and infrequently occurring events in the vast majority of these unrelated cases. By contrast, *de novo* mutations have a smaller role in the pathogenesis of heritable complex traits, such as ASDs, and associated *de novo* mutations are spread across multiple genes. These differences in the etiologic architecture of complex traits make the task of identifying 'causal' genes considerably more challenging. For example, recent exome sequencing studies demonstrated a significant excess of *de novo* loss-of-function mutations in ASD cases but lacked the ability to directly implicate more than a very small number of genes[2–6].

The main complicating factor for interpreting the number of observed *de novo* mutations for a particular gene is the background rate of *de novo* mutation, which can vary greatly between genes. As more individuals are sequenced, multiple *de novo* mutations will inevitably be observed in the same gene by chance. However, if *de novo* mutation has a role in a given disease, we would expect to find

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [3]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [4]Program in Genetics and Genomics, Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA. [5]Department of Psychiatry, Yale University School of Medicine, New Haven, Connecticut, USA. [6]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, USA. [7]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. [8]Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [9]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. [10]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA. [11]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [12]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK. [13]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [14]Synapdx, Lexington, Massachusetts, USA. [15]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [16]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [17]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [18]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas, USA. [19]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [20]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [21]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [22]Department of Psychiatry, University of Illinois at Chicago, Chicago, Illinois, USA. [23]Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [24]Center for Molecular Neuroscience, Vanderbilt University, Nashville, Tennessee, USA. [25]Department of Psychiatry, University of Pittsburgh Medical School, Pittsburgh, Pennsylvania, USA. [26]Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. [27]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. Correspondence should be addressed to M.J.D. (mjdaly@atgu.mgh.harvard.edu).

Received 10 December 2013; accepted 9 July 2014; published online 3 August 2014; doi:10.1038/ng.3050

that genes associated with the disease would contain more *de novo* mutations than expected by chance.

Here we develop a statistical model of *de novo* mutation to evaluate the findings from exome sequencing data. With this model, we establish a statistical framework to evaluate the rates of *de novo* mutation, not only on a per-gene basis (in a frequentist manner analogous to that used in common genome-wide association analysis) but also globally and by gene set. We further use this model to predict the expected amount of rare standing variation per gene and to detect those genes that are significantly and specifically deficient in functional variation, likely reflecting processes of selective constraint. Consequently, as selection has reduced standing functional variation in these genes, it is reasonable to hypothesize that mutations in these genes are more likely to be deleterious.

We used the mutational model along with our list of highly constrained genes to evaluate the relationship between *de novo* mutation and ASDs. Most of the families in these analyses were also included in a set of previous studies of *de novo* mutation, which reported an overall excess of *de novo* loss-of-function mutations in ASD cases, as well as multiple *de novo* mutations in specific genes[2–5]. We build on those studies to examine the aggregate rates of *de novo* mutation, the excess of multiply mutated genes and the overlap of *de novo* mutations with gene sets, which highlights the complex relationship between intellectual functioning and the genetic architecture of ASDs.

RESULTS

Basis of the mutational model

Accurate estimation of the expected rate of *de novo* mutation in a gene requires a precise estimate of each gene's mutability. Although gene length is an obvious factor in a gene's mutability, local sequence context is also a well-known source of differences in mutation rate[7]. Accordingly, we extended a previous model of *de novo* mutation based on sequence context and developed gene-specific probabilities for different types of mutation: synonymous, missense, nonsense, essential splice site and frameshift (Online Methods, **Supplementary Fig. 1** and **Supplementary Table 1**)[3]. Underscoring the importance of the sequence context factors in the model, this genome-wide rate yields an expected mutation rate of 1.67×10^{-8} mutations per base per generation for the exome alone. Using counts of rare (minor allele frequency < 0.001) synonymous variants identified in the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP), we found that our per-gene probabilities of mutation were significantly more correlated ($r = 0.940$) with these counts than with gene length alone ($P < 1 \times 10^{-16}$; Online Methods).

Having established accurate per-gene probabilities of mutation, we could then investigate the rates and distribution of *de novo* mutations found in sequencing studies. Specifically, we wished to systematically assess (i) whether cases had genome-wide excesses of certain functional categories of *de novo* mutation; (ii) whether individual genes could be associated via *de novo* mutation with genome-wide statistical significance; (iii) whether specific sets of genes collectively showed significant enrichment of *de novo* mutations; and (iv) whether there were genome-wide excesses of genes with multiple *de novo* mutations. Below we demonstrate the usefulness of the statistical framework in addressing all of these questions with respect to recently generated family exome sequencing data for autism and intellectual disability.

Identifying genes under selective constraint

There has been a long-standing interest in identifying genes in the human genome that are sensitive to mutational changes, as these genes would be the most likely to contribute to disease. Recent work made use of ESP data to create a metric evaluating the proportion of common functional variation in each gene, thereby identifying genes that appeared to be intolerant of mutation[8]. Along these lines, we correlated our calculated per-gene probabilities of mutation with the observed counts of rare missense variants in the ESP data set. In contrast to the high consistency between predicted synonymous mutation rates and observed synonymous counts (expected if the category is under no specific selection), we observed a significant number of genes with a severe deficit in missense variants compared to the expectation generated from predicted mutation rates ($P < 1 \times 10^{-16}$). Such a deficit is consistent with strong evolutionary constraint: when damaging mutations arise, they are quickly removed from the population by purifying selection. To avoid erroneously identified constrained genes, we removed 134 genes with either significantly elevated or decreased synonymous and nonsynonymous rates (both $P < 0.001$; Online Methods).

Comparing both the synonymous and missense mutation predictions of our model to the ESP data set, we identified a list of excessively constrained genes (missense Z score > 3.09; corresponding to $P < 0.001$) that represented roughly 5% of all genes (**Supplementary Table 2**). A high proportion of the most significantly constrained genes (missense constraint $P < 1 \times 10^{-6}$) were associated with autosomal or

Table 1 Evaluation of the rates of *de novo* mutation in ASD cases and unaffected siblings

Genome-wide excesses of mutational events

	Unaffected siblings (n = 343 families)			ASD cases (n = 1,078 families)		
Mutation type	Observed events per exome	Expected events per exome	P value	Observed events per exome	Expected events per exome	P value
Synonymous	0.21	0.27	0.0218[a]	0.25	0.27	0.1065[a]
Missense	0.61	0.62	0.8189[a]	0.64	0.62	0.5721[b]
Loss of function	0.09	0.09	0.4508[b]	0.13	0.09	**2.05×10^{-7}**[a]

Genome-wide excesses of multiply mutated genes

	Unaffected siblings (n = 343 families)			ASD cases (n = 1,078 families)		
Mutation type	Observed genes with ≥2 DNMs	Average expected genes with ≥2 DNMs	P value	Observed genes with ≥2 DNMs	Average expected genes with ≥2 DNMs	P value
Synonymous	0	0.49	1.0	4	3.8	0.5186
Missense	5	2.5	0.1049	33	21	**0.0070**
Loss of function	0	0.039	1.0	6	0.5	**<0.001**
Loss of function + missense	6	3.0	0.0779	48	27	**<0.001**

The top half of the table shows the observed and expected rates of mutation by type per exome for unaffected siblings[2] and ASD cases, including some unpublished US and Finnish trios[2–6]. The bottom half of the table shows the number of genes with multiple *de novo* mutations in unaffected siblings and ASD cases across studies. The average number of expected genes with multiple *de novo* mutations was determined by simulation. DNMs, *de novo* mutations. Significant P values are shown in bold.
[a]Two-tailed. [b]One-tailed.

Table 2 Individually significant genes identified from the analysis of *de novo* mutations in ASD cases

Gene	Mutations	Number of observed loss-of-function mutations	Number of expected loss-of-function mutations	*P* value
DYRK1A	Nonsense, splice site, frameshift	3	0.0072	**6.15×10^{-8}**
SCN2A	Nonsense, nonsense, frameshift	3	0.018	**9.20×10^{-7}**
CHD8	Nonsense, splice site, frameshift	3	0.022	1.76×10^{-6}
KATNAL2	Splice site, splice site	2	0.0049	1.19×10^{-5}
POGZ	Frameshift, frameshift	2	0.013	8.93×10^{-5}
ARID1B	Frameshift, frameshift	2	0.018	1.57×10^{-4}

Shown are genes with multiple *de novo* loss-of-function mutations across 1,078 ASD cases. Loss-of-function mutations include nonsense, frameshift and splice site–disrupting mutations. Number of expected loss-of-function mutations refers to the expected number of *de novo* loss-of-function mutations based on the probability of mutation for the gene as determined by our model. The genome-wide significance threshold is 1×10^{-6}. Significant *P* values are shown in bold.

X-linked dominant, largely sporadic mendelian disease entries in the Online Mendelian Inheritance in Man database (OMIM; $n = 27/86$). By contrast, a set of genes for which the missense constraint was very close to the expectation ($n = 111$; $-0.01 < Z < 0.01$) had only 2 *de novo* or dominant disease inheritance entries in OMIM, a number significantly different from that for the highly constrained set ($P < 1 \times 10^{-8}$). For the 86 most highly constrained genes, no autosomal recessive mendelian disorders have been documented. However, 11 of the 111 genes with average levels of constraint have been identified as causal in autosomal recessive mendelian disorders. The significant excess of recessive disease-causing genes in the middle part of the distribution in comparison to the constrained set ($P < 0.003$) underscores the idea that recessive inheritance models do not induce strong constraint.

Mutation rates for ASDs and intellectual disability

We applied the model to two primary data sets: published results from ASD sequencing studies[2–6] with a collection of additional unpublished ASD family trios and published results from individuals with severe intellectual disability[9,10]. Comparisons of the predicted number of mutations per exome and the observed data from the 1,078 ASD cases as well as the 343 sequenced unaffected siblings[2–6] are shown in **Table 1**. The model's predictions matched the observed data for the unaffected siblings well, but the cases showed a significant excess of *de novo* loss-of-function mutations ($P = 2.05 \times 10^{-7}$), consistent with the findings of the individual sequencing studies. Using our model to simulate null *de novo* mutation sets, we found that there were significantly more genes with two or more *de novo* loss-of-function mutations than would be expected by chance ($P < 0.001$; six observed when less than one was expected; **Supplementary Table 3**). Notably, although we did not observe a global excess of *de novo* missense mutations, we did observe an excess of genes with 2 or more functional (loss-of-function or missense) *de novo* mutations (48 such genes were observed when the average number expected was 27; $P < 0.001$) and genes with 2 or more *de novo* missense mutations alone (33 such genes were observed when the average number expected was 21; $P = 0.007$ for missense variants; **Table 1**). No such excess of genes containing multiple *de novo* mutations was seen in the unaffected siblings (**Table 1**). Of note, our framework also supports the assessment of many other weightings and combinations of alleles—such as missense variants only (optimal for pure gain-of-function disease models), predicted damaging missense variants only and exact probability estimates for specific combinations of loss-of-function and missense variants—beyond those shown above.

Some of the genes that had 2 or more *de novo* loss-of-function mutations across the 1,078 subjects with ASD are listed in **Table 2**. The results for all genes can be found in **Supplementary Table 4**. A conservative significance threshold of $P = 1 \times 10^{-6}$ was used, correcting for 18,271 genes and 2 tests. Considering this set of 1,078 trios as a single experiment, 2 genes (*DYRK1A* and *SCN2A*) exceeded this conservative genome-wide significance threshold for more *de novo* loss-of-function mutations than predicted. *SCN2A* also had significantly more functional *de novo* mutations than expected. *CHD8*, with three *de novo* loss-of-function mutations and one missense mutation, was very close to the significance threshold in these studies ($P = 1.76 \times 10^{-6}$ for loss-of-function mutations; $P = 3.20 \times 10^{-5}$ for functional mutations). However, a recent targeted sequencing study found 7 additional *de novo* loss-of-function mutations in *CHD8* in ASD cases[11], bringing the total number of *de novo* loss-of-function mutations in *CHD8* to 10, a number that was highly significant ($P = 8.38 \times 10^{-20}$ when accounting for the total number of trios ($n = 2,750$) examined in the combination of the targeted and exome-wide studies). These results offer the encouraging point that, as with genome-wide association studies (GWAS), larger collaborative exome sequencing efforts for trios will define unambiguous risk factors. It is important to note, however, that not all genes with a large number of *de novo* mutations had significant *P* values. For example, *TTN* had four missense *de novo* mutations in ASD cases but had a *P* value that was not even nominally significant ($P = 0.18$), owing to the enormous size of the gene. Even having two *de novo* loss-of-function mutations was on occasion not enough to provide compelling significance (*POGZ*; two frameshift mutations; $P = 8.93 \times 10^{-5}$). In comparison, none of the genes found to contain multiple *de novo* mutations in the unaffected siblings crossed the significance threshold (**Supplementary Table 5**).

These analyses were also applied to the results from the sequencing studies of moderate to severe (IQ < 60) intellectual disability[9,10].

Table 3 Evaluation of the rates of *de novo* mutation in cases with intellectual disability

Genome-wide excesses of mutational events

Mutation type	Intellectual disability cases: Observed events per exome	Intellectual disability cases: Expected events per exome	*P* value
Synonymous	0.19	0.27	0.0267[a]
Missense	0.70	0.62	0.2380[a]
Loss of function	0.24	0.09	**6.49×10^{-7}**[b]

Genome-wide excesses of multiply mutated genes

Mutation type	Intellectual disability cases: Observed genes with ≥2 DNMs	Intellectual disability cases: Average expected genes with ≥2 DNMs	*P* value
Synonymous	1	0.092	0.0879
Missense	3	0.47	**0.0090**
Loss of function	2	0.011	**<0.001**
Loss of function + missense	6	0.60	**<0.001**

The top half of the table shows the observed and expected rates of mutation by type per exome for cases of intellectual disability ($n = 151$ families)[9,10]. The bottom half of the table shows the number of genes with multiple *de novo* mutations in intellectual disability cases across studies. The average number of expected genes with multiple *de novo* mutations was determined by simulation. DNMs, *de novo* mutations. Significant *P* values are shown in bold. [a]Two-tailed. [b]One-tailed.

Table 4 Individually significant genes identified from the analysis of *de novo* mutations in individuals with intellectual disability

Gene	Mutations	Number of loss-of-function mutations	Number of missense mutations	Number of DNMs expected	*P* value	Test
SYNGAP1	Splice site, frameshift, frameshift	3	0	0.0017	**8.15×10^{-10}**	Loss of function
SCN2A	Missense, nonsense, frameshift, frameshift	3	1	0.0025	**2.56×10^{-9}**	Loss of function
SCN2A	Missense, nonsense, frameshift, frameshift	3	1	0.019	**5.01×10^{-9}**	Loss of function + missense
STXBP1	Missense, missense, splice site	1	2	0.0071	**5.87×10^{-8}**	Loss of function + missense
TCF4	Missense, missense	0	2	0.0069	2.39×10^{-5}	Loss of function + missense
GRIN2A	Missense, missense	0	2	0.016	1.34×10^{-4}	Loss of function + missense
TRIO	Missense, missense	0	2	0.033	5.60×10^{-4}	Loss of function + missense

Shown are genes with multiple functional *de novo* mutations across 151 cases of intellectual disability[9,10]. Loss-of-function mutations include nonsense, frameshift and splice site–disrupting mutations. The genome-wide significance threshold is 1×10^{-6}. The number of mutations is either compared to the expected number for loss-of-function mutations only or for both loss-of-function and missense mutations, as indicated by the number of DNMs expected and test columns. Significant *P* values are shown in bold.

Intellectual disability, like ASD, showed a significant excess of *de novo* loss-of-function mutations ($P = 6.49 \times 10^{-7}$; **Table 3**). Even with a much smaller sample size ($n = 151$), there were genes with significantly more loss-of-function and functional *de novo* mutations than predicted by the model (**Table 4**). The data for intellectual disability also showed significantly more genes with multiple missense, loss-of-function and functional *de novo* mutations than predicted ($P = 0.009$ for missense mutations; $P < 0.001$ for loss-of-function and functional mutations).

In our ASD sample, we then investigated the rate of *de novo* events as a function of IQ; roughly 80% of this sample had an IQ assessment attempted. We found that the rate of *de novo* loss-of-function mutation in ASD cases with a measured IQ above average was no different than the expectation (IQ ≥ 100; $n = 229$; 0.08 *de novo* loss-of-function mutations per exome in comparison to the expectation of 0.09; $P = 0.59$). By contrast, the rate in the rest of the sample was substantially higher than the expectation ($n = 572$; rate of 0.17 *de novo* loss-of-function mutations per exome; $P = 1.17 \times 10^{-10}$). Furthermore, when directly compared (rather than being compared to our expectation), these two groups were significantly different from each other ($P < 0.001$), confirming a difference in genetic architecture among ASDs as a function of IQ (**Supplementary Table 6**). These conclusions were unchanged in separate analyses of nonverbal and verbal IQ as well as full-scale IQ (**Supplementary Table 6**).

Gene set enrichment

Given the significant global excess of *de novo* loss-of-function mutations in ASD cases, we wanted to evaluate whether the set of genes harboring *de novo* loss-of-function mutations had significant overlap with several sets of genes proposed to be relevant to autism or describing biochemical pathways. We used the probabilities of mutation to determine the fraction of loss-of-function mutations expected to fall into the given gene set. We then used the binomial distribution to evaluate the number of observed loss-of-function mutations overlapping with the set in comparison to the established expectation. When we applied this analysis to a set of 112 genes reported to be disrupted in individuals with ASDs or autistic features, we observed no enrichment of *de novo* loss-of-function mutations (**Fig. 1**, Betancur)[12]. By contrast, we applied this analysis to a recent study of 842 genes found to interact with the fragile X mental retardation protein (FMRP) *in vivo* and found a highly significant overlap (2.3-fold enrichment; $P < 0.0001$; **Fig. 1**)[2,13]. This enrichment with the targets of FMRP held even when we removed the *de novo* mutations identified in the study by Iossifov *et al.*[2], which initially reported an enrichment of *de novo* mutations in ASD cases in FMRP-associated genes (2.5-fold enrichment; $P < 0.0001$).

We then evaluated the group of individuals from the ASD studies who had a *de novo* loss-of-function event in one of the targets of FMRP. On average, these cases were enriched for having a measured IQ of <100 (Fisher's exact test $P = 4.01 \times 10^{-4}$; **Supplementary Table 7**) as well as a significantly reduced male/female ratio ($P = 0.02$; **Supplementary Table 8**) as compared to the remaining sequenced cases (**Supplementary Note**). These individuals represented about 3% of the total sample, when, at most, a 1% overlap would be expected. The estimated odds ratio (OR) of *de novo* loss-of-function events in the set of FMRP target genes was around 6, very similar to the ORs estimated for large copy number variants (CNVs) that disrupt multiple genes[14]. In addition, the OR for the published cases of moderate to severe intellectual disability noted above (IQ < 60; not ascertained for ASDs) having a *de novo* loss-of-function event in the set of FMRP targets was roughly 10.

The same analysis was applied to the list of *de novo* loss-of-function events from the unaffected siblings of ASD cases and additional control individuals ($n = 647$)[2,4,5,15]. There was a significant enrichment when evaluating overlap with the set of autism-related genes ($P = 0.0095$; **Fig. 1**). However, no significance was observed for overlap with the *in vivo* targets of FMRP. The list of *de novo* loss-of-function mutations from the individuals with intellectual disability, on the other hand, was significant for both sets ($P < 1 \times 10^{-4}$ for both sets; **Supplementary Fig. 2**). Even the *de novo* missense mutations found in the intellectual disability cases showed significant overlap with both

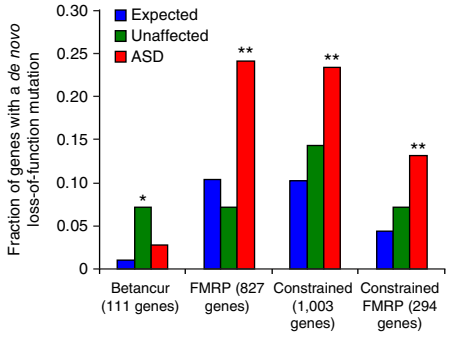

Figure 1 The expected and observed fraction of genes with a *de novo* loss-of-function mutation in ASD cases and unaffected controls for four gene sets of interest. ASD cases ($n = 1,078$) and unaffected controls ($n = 647$) were sequenced across various studies (refs. 2–6,10,15). "Betancur" refers to a set of genes reported to be disrupted in individuals with ASDs or autistic features; of the 112 on the list[12], we could evaluate 111. "FMRP" refers to the genes whose mRNAs are bound and regulated by the fragile X mental retardation protein, as identified by Darnell *et al.*[13]. The "constrained" category is a set of 1,003 genes that we defined as significantly lacking rare missense variation, indicating intolerance to mutation. The targets of FMRP that are also considered constrained by our metric make up the "constrained FMRP" category. $^*P < 0.01$, $^{**}P < 1 \times 10^{-4}$, binomial test.

sets under study ($P = 0.02$ for autism-related genes and $P < 0.0001$ for the targets of FMRP; **Supplementary Fig. 2**).

Evaluating constrained genes

We further applied the enrichment analysis to our set of constrained genes and found that they contained more *de novo* loss-of-function mutations than expected by chance (2.3-fold enrichment; $P < 0.0001$; **Fig. 1**). We observed a greater fold enrichment when focusing on the subset of constrained genes that were also identified in the FMRP study (3.0-fold enrichment; $P < 0.0001$; **Fig. 1**)[13]. We note that the FMRP targets showed significant overlap with the constrained set of genes (OR = 1.29; $P < 0.0001$), which is consistent with the report that the targets of FMRP are under greater purifying selection than expected[2]. All enrichments were demonstrated to be independent of gene length (**Supplementary Note**).

The genes that contained a *de novo* missense or loss-of-function mutation in the intellectual disability cases also showed a significant enrichment for both the constrained gene set and the set of constrained targets of FMRP ($P < 0.0001$ for all lists). In comparison, no enrichment was found with either set for the list of genes that had a *de novo* loss-of-function mutation in unaffected siblings and control individuals.

In addition to treating constraint as a dichotomous trait, we also evaluated the missense *Z* score for each of the genes with a *de novo* loss-of-function mutation. We found that the distribution of missense *Z* scores for genes with a *de novo* loss-of-function mutation in unaffected individuals was no different than the overall distribution of scores (Wilcoxon $P = 0.8325$; **Fig. 2**). By contrast, both the genes with a *de novo* loss-of-function mutation in ASD and intellectual disability cases had values significantly shifted toward high constraint (Wilcoxon $P < 1 \times 10^{-6}$ for both). Furthermore, we compared the distribution of *Z* scores among each of the three groups. Both the ASD and intellectual disability distributions were significantly different from the distribution of missense *Z* scores for unaffected individuals ($P = 0.0148$ and 0.0012, respectively). The intellectual disability missense *Z* scores were also significantly higher than the corresponding ASD values ($P = 0.0319$).

When evaluating the ASD cases split by IQ group, we found no enrichment of genes with *de novo* loss-of-function mutations with either constrained genes or targets of FMRP in the group with IQ of ≥100 ($P > 0.5$ for both sets of genes), but we found very strong enrichment in the set with IQ of <100 ($P < 0.0001$ for both sets of genes). These results underscore the idea that phenotypically distinct subsets of ASD cases may have significantly different contributions from *de novo* mutation.

Comparison of constraint metric with existing methods

Identifying constrained genes by comparing observed nonsynonymous sites to the expectation is conceptually similar to the traditional approach of detecting selective pressure by comparing observed nonsynonymous sites to observed synonymous sites (for example, d_N/d_S) that has been used extensively. Our approach should in principle achieve greater statistical power to detect constrained genes; comparison of an observation to an expectation is statistically more powerful than contrasting that observation with a generally smaller second observation (the number of observed synonymous variants). To investigate this claim, we identified genes that had significant evidence for selective constraint using the d_N/d_S metric (their ratio of synonymous to nonsynonymous sites deviated from the genome-wide average at $P < 0.001$; **Supplementary Note**). There were only 377 of these genes, over half of which overlapped with the constrained gene list defined by our method ($n = 1{,}003$; overlap of 237 genes). The genes identified as significantly constrained by only our metric (the top 10 of which included *RYR2*, *MLL* (*KMT2A*), *MLL2* and *SYNGAP1*) were still significantly enriched for known causes of autosomal and X-linked dominant forms of mendelian disease ($P = 5 \times 10^{-4}$). We therefore conclude that the model-based approach to identifying constrained genes adds substantial power to traditional approaches. The importance of this increased power to detect constraint is further articulated in the ASD and intellectual disability analyses below.

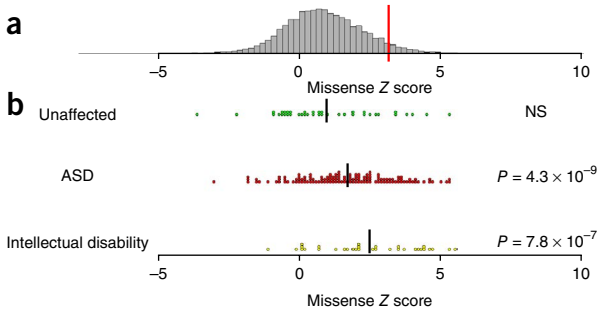

Figure 2 Distributions of missense *Z* scores and *Z* scores for genes containing *de novo* loss-of-function mutations identified in unaffected individuals, ASD cases and intellectual disability cases. (**a**) Distribution of missense *Z* scores. The red bar indicates a *Z* score of 3.09, or the threshold for inclusion in the set of 1,003 constrained genes. (**b**) Missense *Z* scores for genes containing *de novo* loss-of-function mutations in unaffected individuals, ASD cases and intellectual disability cases[2–6,9,10,15]. Black bars indicate the mean *Z* score of each group: 0.94, 1.68 and 2.46 for unaffected individuals, ASD cases and intellectual disability cases, respectively. Although the missense *Z* scores of the *de novo* loss-of-function mutations found in unaffected siblings matched the overall distribution (Wilcoxon $P = 0.8325$; NS, not significant), *de novo* loss-of-function mutations found in both ASD and intellectual disability cases were significantly shifted toward more extreme constraint values ($P < 1 \times 10^{-6}$ for both). All *P* values for deviation from the overall distribution are listed on the right side of the figure. In addition, the distributions of missense *Z* scores for each of the three *de novo* lists were all individually significant at $P < 0.05$.

Several groups have previously published approaches and specific gene sets from these that are also aimed at identifying genes under excessive purifying selection or generally intolerant of functional mutation. Bustamante *et al.*[16] expanded on the McDonald-Kreitman framework[17], contrasting fixed differences in the primate lineage to polymorphic differences in humans to identify a set of genes under weak negative selection, while more recently Petrovski *et al.*[8] used the excess of rare versus common missense variation within humans to flag genes intolerant of functional variation. We found a reasonable correlation between our metric of constraint and the residual variation intolerance score (RVIS) of Petrovski *et al.* (**Supplementary Fig. 3**)[8]. A comparison of these approaches as applied to the prioritization of known haploinsufficient genes, as well as to the *de novo* loss-of-function mutations in autism described here, is provided in the **Supplementary Note** and demonstrates that the two human-only approaches (constraint and RVIS) perform better on these tasks of identifying medical genetics lesions of severe effect in modern humans (**Supplementary Table 9**). Intriguingly, both of these other approaches use independent information from each other and from our approach (which uses the absence of rare functional variation in comparison to the expectation within humans), raising the possibility that composite scores employing all three sources of information could add further value in highlighting which genes are most sensitive to heterozygous mutation.

DISCUSSION

We have developed a framework for evaluating excesses of *de novo* mutations identified through exome sequencing. Even though this framework can be leveraged to evaluate excesses of mutation across a study and in gene sets, the key focus is on evaluating the significance for individual genes. Given the small number of observed *de novo* events per gene, simple case-control comparisons cannot achieve any meaningful level of significance. For example, observing 3 *de novo* loss-of-function mutations in a small gene in 1,000 case trios is perhaps quite compelling, especially if no such mutations were identified in 1,000 control trios. However, a simple three-to-zero case-control comparison in this situation would yield no compelling statistical evidence (one-tailed $P = 0.125$). Incidence of such extremely rare events, however, can be evaluated if the expected rate of such events is known. Sequencing large numbers of control trios to gather empirical rate estimates on a per-gene basis that are accurate is infeasible and inefficient. The calibrated model and statistical approach described here can achieve a close approximation of this ideal. Our method, therefore, offers the ability to evaluate the rate of rare variation in individual genes in situations where burden tests would fail.

Other groups have developed similar statistical frameworks[11,18]; notably, the Epi4K Consortium[18] used the same base model we began with[3] to interpret event rates. Our model, however, has two primary strengths. First, our model of *de novo* mutation incorporates additional factors beyond sequence context that affect mutation rate. Both the depth of coverage (how many sequence reads were present on average) for each base and the regional divergence around the gene between humans and macaques independently and significantly improve the predictive value of our model (**Supplementary Note**). Second, given the high correlation between the number of rare synonymous variants in ESP and the probability of a synonymous mutation determined by our full model, we have a metric to evaluate the extent to which genes in the human genome show evidence of selective constraint. The list of 1,003 genes that we define as constrained contains an enrichment of genes known to cause severe human disease—an observation analogous to that recently made in using empirical comparison of common and rare rates of functional variation to evaluate intolerance to mutation[8]. In fact, site count deficits and shifts in site frequency each contribute independent information to the definition of constraint and can in principle be combined in a composite test.

The results of our metric were compared to both the scores created by Petrovski *et al.*[8] and the loci identified as being under negative selection by Bustamante *et al.*[16]. Overall, our metric and the RVIS metric defined by Petrovski *et al.* worked similarly well, reinforcing the benefits that could come from combining the two approaches. It is unsurprising that these methods outperform the evolutionary ones on the specific matter of genes intolerant to heterozygous mutation. Evolutionary methods examining differences between polymorphism and fixed differences, which are more sensitive to weaker negative selection, require that mutations be tolerated well enough to become polymorphic in the first place. By contrast, approaches measuring the complete absence of variation will pick up the most strongly intolerant genes.

Ideally, we can conceptualize defining two metrics of genic constraint, one based on missense variants and the other based on loss-of-function variants. With only 6,503 individuals in ESP, we are underpowered to determine significant deviations for most genes with respect to loss-of-function variants. As sample size increases, our ability to calculate constraint improves. For example, if the sample size were to increase by an order of magnitude, we would be able to evaluate approximately 66% of genes using loss-of-function variants. We therefore view the constrained gene list as a work in progress, to be updated when larger exome sequencing data sets become available.

Applying our statistical framework to *de novo* mutations from 1,078 ASD cases shows that, although there is no global excess in *de novo* missense mutations, there are significantly more genes that contain multiple *de novo* missense mutations than expected. We also see significant overlap between the list of genes with a *de novo* loss-of-function mutation in ASD cases and the set of constrained genes that we defined. In addition, there is significant overlap between the genes with a *de novo* loss-of-function mutation and the targets of FMRP, as reported in Iossifov *et al.*[2]. All of the significant signals in ASD—the global excess of *de novo* loss-of-function mutations, the excess of genes with multiple functional *de novo* mutations, the overlap between the genes with *de novo* loss-of-function mutation and both constrained genes and the targets of FMRP—are not found in the subset of ASD cases with IQ of ≥100. The lack of signal in this subset indicates that genetic architecture among ASDs varies as a function of IQ. Overall, the probabilities of mutation defined by our full model and list of constrained genes can be used to critically evaluate the observed *de novo* mutations from sequencing studies and to aid in the identification of variants and genes that have a critical role in disease.

URLs. Online Mendelian Inheritance in Man (OMIM), http://omim.org/; Exome Variant Server, http://evs.gs.washington.edu/EVS/; site to query constraint information and *de novo* mutations from published studies, http://atgu.mgh.harvard.edu/webtools/gene-lookup/; Picard, http://picard.sourceforge.net/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. New data included in this manuscript have been deposited in the database of Genotypes and Phenotypes (dbGaP), merged with our published data under accession phs000298.v1.p1.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

All data from published studies are available in the respective publications. All newly generated data and computational tools used in this paper will be available online as downloadable material. We have also constructed a website to query genes that provides information on constraint and the *de novo* mutations found in the specified gene across published studies of *de novo* mutation. We would like to thank E. Daly and M. Chess for their contributions to data analysis and the construction of the website, respectively. We acknowledge the following resources and families who contributed to them: the National Institute of Mental Health (NIMH) repository (U24MH068457); the Autism Genetic Resource Exchange (AGRE) Consortium, a program of Autism Speaks (1U24MH081810 to C.M. Lajonchere); The Autism Simplex Collection (TASC) (grant from Autism Speaks); the Simons Foundation Autism Research Initiative (SFARI) Simplex Collection (grant from the Simons Foundation); and The Autism Consortium (grant from the Autism Consortium). This work was directly supported by US National Institutes of Health (NIH) grants R01MH089208 (M.J.D.), R01MH089025 (J.D.B.), R01MH089004 (G.D.S.), R01MH089175 (R.A.G.) and R01MH089482 (J.S.S.) and was supported in part by US NIH grants P50HD055751 (E.H.C.), R01MH057881 (B.D.) and R01MH061009 (J.S.S.). We acknowledge partial support from grants U54HG003273 (R.A.G.) and U54HG003067 (E. Lander). We thank T. Lehner (NIMH), A. Felsenfeld (National Human Genome Research Institute) and P. Bender (NIMH) for their support and contribution to the project. E.B., J.D.B., B.D., M.J.D., R.A.G., K. Roeder, A.S., G.D.S. and J.S.S. are lead investigators in the ARRA Autism Sequencing Collaboration (AASC). We would also like to thank the NHLBI GO Exome Sequencing Project (ESP) and its ongoing studies that produced and provided exome variant calls on the web: the Lung GO Sequencing Project (HL-102923), the

Women's Health Initiative (WHI) Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010).

AUTHOR CONTRIBUTIONS

K.E.S., B.M.N. and M.J.D. conceived and designed the mutational model and constraint methods. K.E.S. and E.B.R. executed the analyses. K.E.S., E.B.R., L.M.M., J.A.K., S.M., A.K., D.P.W., D.G.M., S.M.P., J.D.B., B.D. and K. Roeder contributed to analysis concepts and methods. K.E.S., S.J.S., C.S., A.S., K. Rehnström, S.B.G., M.D., A.P., E.B., J.D.B., E.H.C., R.A.G., G.D.S., J.S.S., B.D., K. Roeder, B.M.N. and M.J.D. contributed autism sequencing, evaluation and manuscript comments. K.E.S., E.B.R., B.M.N. and M.J.D. performed the primary writing.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Ng, S.B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat. Genet.* **42**, 790–793 (2010).
2. Iossifov, I. *et al. De novo* gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (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. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
5. Sanders, S.J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
6. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nat. Genet.* **43**, 585–589 (2011).
7. Antonarakis, S.E. CpG dinucleotides and human disorders. in *Encyclopedia of Life Sciences* (John Wiley & Sons, Chichester, UK, 2006).
8. 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).
9. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
10. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
11. O'Roak, B.J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
12. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380**, 42–77 (2011).
13. Darnell, J.C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
14. Sanders, S.J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).
15. Xu, B. *et al. De novo* gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.* **44**, 1365–1369 (2012).
16. Bustamante, C.D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
17. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
18. Epi4K Consortium & Epilepsy Phenome/Genome Project. *De novo* mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).

ONLINE METHODS

***De novo* mutation information.** Published *de novo* mutations were collected for both ASD[2–6] and severe intellectual disability[9,10]. Updated *de novo* calls were provided from two of the ASD studies[3,5]. Details about sample collection, sequencing and variant processing can be found in the separate studies.

Additional sequencing. Exome sequencing of the additional families ($n = 129$) was performed at the Broad Institute. Exons were captured using Agilent 38Mb SureSelect v2. After capture, a round of ligation-mediated PCR was performed to increase the quantity of DNA available for sequencing. All libraries were sequenced using an Illumina HiSeq 2000 instrument. Data were processed with Picard, which uses base quality score recalibration and local realignment at known indels[19] and Burrows-Wheeler Aligner (BWA)[20] to map reads to hg19. SNPs were called using the Genome Analysis Toolkit (GATK) for all trios jointly[19,21]. The variable sites that we have considered in analysis were restricted to those that passed GATK standard filters. From this set of variants, we identified putative *de novo* mutations and validated them as previously described[3]. Autism Consortium samples ($n = 78$ trios) were collected in Boston under institutional review board (IRB) approval from Harvard Medical School, Massachusetts General Hospital, Children's Hospital Boston, Tufts–New England Medical Center and Boston University Medical Center with ADI and ADOS assessment. Finnish autism samples ($n = 51$ trios) were collected under IRB approval at the University of Helsinki with ADI and ADOS assessment and consented for autism research only. In both studies, all participants gave written informed consent, although, as autism is classified as a childhood disorder, many subjects are children, with informed consent provided by their parents or guardians.

Mutational model. We wanted to create an accurate model of *de novo* mutation for each gene. To do so, we extended a previous sequence context–based model of *de novo* mutation to derive gene-specific probabilities of mutation for each of the following mutation types: synonymous, missense, nonsense, essential splice site and frameshift[3]. In brief, local sequence context was used to determine the probability of each base in the coding region mutating to each other possible base and then to determine the coding impact of each possible mutation. These probabilities of mutation were summed across genes to create a per-gene probability of mutation for the aforementioned mutation types (see the **Supplementary Note** for more details). Here we applied the method to exons and immediately flanking essential splice sites, but note that the framework is applicable to non-genic sequences. While fitting the expected rates of mutation to observed data, we added a term for local primate divergence across 1 Mb (to capture additional unmeasured sources of regional mutational variability) and another for the average depth of sequence of each nucleotide (to capture inefficiency of variant discovery at lower sequencing depths); both terms significantly improved the fit of the model to observed data (details in the **Supplementary Note**). We also investigated a regional replication timing term[22] but found no evidence for it significantly improving the model (**Supplementary Note**).

To evaluate the predictive value of the model of *de novo* coding mutations, we extracted synonymous variants that were seen 10 times or fewer in the 6,503 individuals in ESP and compared the number of these rare variants in each gene to (i) the length of the gene and (ii) the probability of a synonymous mutation for that gene as determined by our model. Although gene length alone showed high correlation ($r = 0.880$), our full model showed significantly greater correlation ($r = 0.940$; $P < 1 \times 10^{-16}$). Of note, the stochastic variability of counts from ESP is such that, if the model were perfect, the correlation to any instance of these data would be 0.975, indicating that little additional gene-to-gene variability remains to be explained. The relative rates of different types of coding mutation were quite similar to those in previous work based on primate substitutions[23]. With this calibrated model of relative mutability, we determined the absolute expected mutation rate per gene by applying a genome-wide mutation rate of 1.2×10^{-8} mutations per base pair per generation (**Supplementary Note**)[24,25].

Removing potential false positive constrained genes. To identify genes that appeared to be significantly constrained, we used our probabilities of mutation to predict the expected amount of synonymous and nonsynonymous variation in ESP data. Those genes that had the expected amount of synonymous variation but were significantly ($P < 0.001$) deficient for missense variation were labeled as constrained. To ensure that genes were not nominated as being constrained erroneously, we excluded from all analyses 134 genes in which the observed synonymous and nonsynonymous rates were both significantly elevated or significantly decreased (both $P < 0.001$). Upon inspection, this list contained a number of genes that contained an internal duplication (for example, *FLG*), a nearby pseudogene (for example, *AHNAK2*) and a number of cases where recent duplications and/or annotation errors have led to the same sequence being assigned to two genes (for example, *SLX1A* and *SLX1B*). These are all scenarios where standard exome processing pipelines systematically undercall variation (reads are unmapped owing to uncertainty on which gene to assign them to) or overcall false variants owing to read misplacement. This further suggests that a byproduct of this analysis framework is the identification of a residual set of challenging genes for current exome sequencing pipelines.

19. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
20. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
21. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
22. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012).
23. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
24. Campbell, C.D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44**, 1277–1281 (2012).
25. Conrad, D.F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43**, 712–714 (2011).

doi:10.1038/ng.3050