Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples

Jack A Kosmicki1–5, Kaitlin E Samocha1–3,5, Daniel P Howrigan1–3, Stephan J Sanders5, Kamil Slowikowski12–4,7,8, Monkol Lek1,2, Konrad J Karczewski1,2, David J Cutler9, Bernie Devlin10, Kathryn Roeder11, Joseph D Buxbaum12–17, Benjamin M Neale1–3, Daniel G MacArthur1,2, Dennis P Wall18, Elise B Robinson1–3 & Mark J Daly1–3

Recent research has uncovered an important role for de novo variation in neurodevelopmental disorders. Using aggregated data from 9,246 families with autism spectrum disorder, intellectual disability, or developmental delay, we found that ~1/3 of de novo variants are independently present as standing variation in the Exome Aggregation Consortium's cohort of 60,706 adults, and these de novo variants do not contribute to neurodevelopmental risk. We further used a loss-of-function (LoF)-intolerance metric, pLI, to identify a subset of LoF-intolerant genes containing the observed signal of associated de novo protein-truncating variants (PTVs) in neurodevelopmental disorders. LoF-intolerant genes also carry a modest excess of inherited PTVs, although the strongest de novo–affected genes contribute little to this excess, thus suggesting that the excess of inherited risk resides in lower-penetrant genes. These findings illustrate the importance of population-based reference cohorts for the interpretation of candidate pathogenic variants, even for analyses of complex diseases and de novo variation.

Autism spectrum disorders (ASDs) are a phenotypically heterogeneous group of heritable disorders that affect ~1 in 68 individuals in the United States1. Although estimates of the contribution of the common variant (heritable) to ASD risk are upward of 50% (refs. 2–4), few specific risk variants have been identified, partly because ASD sample sizes in genome-wide association studies remain limited to date. In contrast, the field has made substantial progress in understanding the genetic etiology of ASD via analysis of de novo (newly arising) variation through exome sequencing of parent–offspring trios5–10. Severe intellectual disability and developmental delay (ID/DD) are considerably less heritable than ASDs11 (though they are frequently comorbid) and have demonstrated a stronger contribution from de novo–frameshift, splice-acceptor, splice-donor, and nonsense variants (collectively termed PTVs)12–14. Furthermore, ASD cases with comorbid ID show a significantly higher rate of de novo PTVs than those with normal or above-average intellectual quotient (IQ)6,9,15–17, whereas higher-IQ cases have a stronger family history of neuropsychiatric disease15, thus suggesting a greater heritable contribution.

De novo variants are a unique component of the genetic architecture of human disease, because, having not yet passed through a single generation, any heterozygous variants with complete or near-complete elimination of reproductive fitness must reside almost exclusively in this category. Despite prior evidence of mutational recurrence18 (i.e., the same mutation occurring de novo in multiple individuals), most studies have implicitly assumed that each de novo variant is novel, in line with Kimura’s infinite sites model19, and thereafter have analyzed de novo variants genome wide without respect to their allele frequency in the population (Supplementary Note). However, the mutation rate is not uniform across the genome, and some regions and sites (e.g., CpG sites20) experience higher mutation rates than others. A classic example comes from achondroplasia, in which the same G>C and G>T variant at a CpG site has been observed de novo in 150 and 3 families, respectively18. As such, it should not be surprising to observe a de novo variant at a given site and to also observe the same variant (defined as one with the same chromosome, position, reference, and alternate allele) present as standing variation in the population.

Given the strong selective pressure on neurodevelopmental disorders21–23, we expect that most highly deleterious (high-risk-conferring) de novo PTVs will linger in the population for at most a few generations. Thus, the collective frequency of such variants in the population will approximate their mutation rate. Individual PTVs that are tolerated in relatively healthy adults, and more generally PTVs in genes that tolerate the persistence of such variants in the population, may be less likely to contribute significant risk to such phenotypes and are therefore permitted by natural selection to reach allele frequencies that are orders of magnitude greater than those of highly deleterious variants. Given the current size of the human population (~7 billion), and the expectation of one de novo variant per exome (1 in ~30 million bases), every nonembryonic lethal coding mutation is likely to be present as a de novo variant at least once in the human population. This reasoning, along with the availability of large exome-sequencing

A full list of affiliations appears at the end of the paper.

Received 12 May 2016; accepted 20 January 2017; published online 13 February 2017; doi:10.1038/ng.3789
reference databases, motivated our interest in searching for variants observed de novo in trio-sequencing studies that are also present as standing variation in the human population, thus indicating recurrent mutations. We herein refer to these de novo variants that are also observed as standing variation in the population as class 2 $de novo$ variants, and the remaining de novo variants are referred to as class 1 $de novo$ variants (i.e., observed only de novo, Fig. 1).

With the release of the Exome Aggregation Consortium (ExAC) data set of 60,706 adult individuals without severe developmental abnormalities$^{24}$, the rate and relative pathogenicity of recurrent variations in ASD and ID/DD, coupled with the large sample sizes published to date, led us to focus our evaluation on these phenotypes.

RESULTS

Class 2 de novo variation

We first asked how many of the 10,093 variants observed de novo in individuals with ID/DD$^{13}$, individuals with ASD, and unaffected ASD siblings might also be observed as standing variation in the 60,706 reference exomes from ExAC$^{24}$ (Fig. 1 and Online Methods). We found that 1,854 ASD (31.66%), 841 unaffected ASD sibling (33.05%), and 410 ID/DD (24.23%) $de novo$ variants were observed as standing variation in one or more ExAC individuals (class 2 $de novo$ variants) (Fig. 2a and Supplementary Tables 1–3). When we removed the 15,330 exomes originating from psychiatric cohorts (many of which were controls), the rate of class 2 $de novo$ variation decreased to 28.47% ($\pm$1.03%, 95% confidence interval [CI]), a rate statistically indistinguishable from the expected recurrence rate of 28.13% ($\pm$0.42%, 95% CI; two-sided binomial test; $P=0.45$; Fig. 2b, Supplementary Figs. 1 and 2, Supplementary Tables 4, and Online Methods). We found similar rates of class 2 $de novo$ variants in published trio studies of schizophrenia$^{25}$ and congenital heart disease$^{26,27}$ (Supplementary Tables 5 and 6). Although the presence of class 2 $de novo$ variants is not a novel observation$^{18,25}$, the rate is approximately three times greater than previous estimates$^{25}$, owing to substantially larger reference data sets (Fig. 2b and Supplementary Fig. 2).

We ran five secondary analyses to confirm the validity of the observed recurrence rate, ranging from evaluating the rate of CpG variants to ensuring a proper allele frequency distribution (Fig. 2c, Supplementary Note, and Supplementary Tables 7 and 8). We then sought to determine whether class 1 and class 2 $de novo$ variants contributed equally to ASD and ID/DD risk. As a control for the comparison of functional variants, the rates of both class 1 and class 2 $de novo$ synonymous variants were equivalent across ASD, ID/DD, and unaffected ASD siblings (Fig. 3a and Supplementary Table 9) and remained unchanged when we removed the psychiatric cohorts within ExAC (Supplementary Fig. 3a and Supplementary Table 10). Thus, collectively, neither class 1 nor class 2 $de novo$ synonymous variants demonstrated an association with ASD or ID/DD, in agreement with previous reports that, as a class, $de novo$ synonymous variants do not contribute to risk$^{5–10}$. Although previous reports have implicated $de novo$ PTVs as significant risk factors for ASD$^{6,15,16}$ and ID/DD$^{13}$, the class 2 $de novo$ subset of PTVs showed no such association for either ASD (0.015 per case versus 0.023 per unaffected ASD sibling; $P=0.98$; one-sided Poisson exact test$^{26}$) or ID/DD (0.016 per case versus 0.023 per unaffected ASD sibling; $P=0.94$; one-sided Poisson exact test), and there were slightly higher rates in unaffected ASD siblings (Fig. 3b and Supplementary Table 11). By contrast, after removal of class 2 $de novo$ PTVs, class 1 $de novo$ PTVs were significantly more enriched in individuals with ASD (0.13 per case) and ID/DD (0.19 per case) than in unaffected ASD siblings (0.07 per control) (ASD versus control, rate ratio (RR) = 1.83; $P=6.07 \times 10^{-13}$; ID/DD versus control, RR = 2.61; $P=6.31 \times 10^{-21}$; one-sided Poisson exact test). The lack of excess case burden in class 2 $de novo$ variants was consistent with what would be expected if such variants did not contribute to ASD and ID/DD risk. However, to ensure that we were not losing causal variants by removing all $de novo$ variants found in ExAC, we tested the class 2 $de novo$ PTVs at three ExAC allele frequency (AF) thresholds: singletons (one allele in ExAC), AF < 0.0001, and AF < 0.001. We found no significant differences at any threshold between the rates of class 2 $de novo$ PTVs in individuals with ID/DD or as compared with unaffected ASD siblings (Fig. 3c and Supplementary Table 12). Furthermore, these results remained consistent regardless of whether the psychiatric exomes in ExAC were retained or excluded, thus demonstrating that they were not driving the observed associations (Supplementary Fig. 3b and Supplementary Table 13). Thus, the data provided no evidence that these class 2 $de novo$ variants contributed to the previously observed enrichment of $de novo$ variation in ASD and ID/DD cases, and removing those variants present in ExAC increased the effect size for $de novo$ PTVs in ASD and ID/DD. Subsequently, we focused our analyses solely on variation absent from ExAC.

Gene-level analyses

Because observed risk to ASD or ID/DD was carried only by $de novo$ variants absent from the standing variation of ExAC, we next sought to extend this concept by evaluating whether the overall rate of PTVs per gene in ExAC might provide a similar guide to which ASD and ID/DD variants were relevant. Specifically, we investigated whether the gene-level constraint metric probability of loss-of-function intolerance ($pLI)^{16}$ might improve our ability to decipher which class 1 $de novo$ PTVs conferred the highest risk of ASD and ID/DD (Online Methods). Using the same threshold as Lek et al.$^{23}$, we used a threshold of $pLI \geq 0.9$ to define LoF-intolerant genes and investigated whether individuals with ASD had an increased burden of class 1 $de novo$ PTVs in such genes. When we restricted our search to solely class 1 $de novo$ PTVs in LoF-intolerant genes, we observed a significant excess in individuals with ASD (0.067 per exome) compared with their unaffected siblings (0.021 per exome; RR = 3.24; $P=3.14 \times 10^{-16}$, one-sided Poisson exact test). For individuals with ID/DD, the rate of class 1 $de novo$ PTVs in LoF-intolerant genes became more striking, with a rate of 0.139 per exome, thus resulting in a 6.70 RR relative to the control group of unaffected ASD siblings ($P=6.34 \times 10^{-38}$, one-sided Poisson exact test). By contrast, the rate of class 1 $de novo$ PTVs in LoF-intolerant genes was not a novel observation$^{18,25}$, the rate is approximately three times greater than previous estimates$^{25}$, owing to substantially larger reference data sets (Fig. 2b and Supplementary Fig. 2).
PTVs in LoF-tolerant genes (pLI <0.9) showed no difference between individuals with ASD (0.063 versus 0.051; P = 0.06; two-sided Poisson test) or individuals with ID/DD (0.048 versus 0.051; P = 0.75; two-sided Poisson exact test; Fig. 3d and Supplementary Table 14) compared with their unaffected ASD siblings. Again, the results remained unchanged when we excluded the ExAC psychiatric samples (Supplementary Fig. 3c and Supplementary Table 15). The same trend was observed in congenital heart disease26,27 and schizophrenia25 (Supplementary Note and Supplementary Tables 16–21). Hence, all detectable de novo PTV signal in these phenotypes was localized to 18% of genes with clear intolerance to PTVs in ExAC, and consequently the rate ratios were substantially amplified in this gene set.

Recent studies have inferred the presence of multiple de novo PTVs in the same gene as constituting evidence of contribution to ASD risk5–10. Of the 51 genes with two or more de novo PTVs, only 38 were absent in controls (Supplementary Table 22). This finding reinforced that the mere observation of multiple de novo PTVs in a gene is not sufficient to define the gene as being important5,16 and also allowed us to explore whether the pLI metric might be used to refine the identification of specific genes. In fact, 32 of the 38 case-only genes, but only 5 of the 13 control-only or case–control hit genes, were LoF intolerant, a highly significant difference (odds ratio (OR) = 8.07; P = 0.003; Fisher’s exact test) that greatly refined the list of genes to be pursued as likely ASD contributors.

Phenotypic associations for class 1 de novo PTVs in LoF-intolerant genes

Although enrichment of de novo PTVs is one of the hallmarks of ASD de novo studies5–10,15,16, another consistent finding is an increased burden of these variants among females with ASD5,15 and in individuals with ASD and low full-scale IQ (FSIQ)5,15,16. We investigated whether these hallmarks were present in the 6.55% of ASD cases with a class 1 de novo PTV in LoF-intolerant genes (pLI ≥0.9). Indeed, females were overrepresented in the subset (12.26% of females; 5.80% males; P = 1.75 × 10−5; Fisher’s exact test; Supplementary Table 23). Importantly, for the 6.86% of ASD cases with a class 2 de novo PTV or a class 1 de novo PTV in a LoF-tolerant gene (pLI <0.9), there were no differences between the sexes: 6.86% of females and 6.83% of males fell in this category (P = 1; Fisher’s exact test; Supplementary Table 24). Furthermore, class 2 de novo PTVs and class 1 de novo PTVs in LoF-tolerant genes showed no association with FSIQ (β = −0.001; P = 0.76; Poisson regression), whereas class 1 de novo PTVs in LoF-intolerant genes predominately explained the skewing toward lower FSIQ (β = −0.023; P = 7 × 10−8; Poisson regression; Fig. 4a). Given these observations, we divided the ASD class 1 de novo PTV signal in LoF-intolerant genes on the basis of sex and intellectual disability status (Online Methods). Females with comorbid ASD and intellectual disability had the highest rate of class 1 de novo PTVs in LoF-intolerant genes (RR = 8.71; P = 2.73 × 10−12; one-sided Poisson exact test). Despite the strong enrichment in females and individuals with comorbid ASD and intellectual disability, males with ASD without intellectual disability still showed enrichment of class 1 de novo PTVs in LoF-intolerant genes (RR = 2.95; P = 1.31 × 10−9; one-sided Poisson exact test; Fig. 4b and Supplementary Table 25). These secondary analyses strongly supported the implication from the primary analysis that, collectively, class 2 de novo PTVs and class 1 de novo PTVs in LoF-tolerant genes have little to no association with ASD or ID/DD and no observable phenotypic effects on individuals carrying them. By contrast, the class 1 de novo variants occurring in LoF-intolerant genes contained the association signal and phenotypic skewing observed to date.

Inherited variation

Because the effect size for de novo PTVs increased after removal of those variants present in ExAC, we postulated that a similar increase might be obtained from rare inherited PTVs. Under the assumption that risk-conferring variants should be transmitted more often to individuals with ASD, we tested for transmission disequilibrium in a cohort of 4,319 trios with an affected proband (Online Methods). Without filtering by pLI or presence/absence status in ExAC, singleton PTVs as a class showed no overtransmission (P = 0.31; two-sided binomial test). After removal of all of the variants present in ExAC or in a LoF-tolerant gene (pLI <0.9), we found a modest excess of transmitted
PTVs in ASD cases (RR = 1.16; $P = 9.85 \times 10^{-3}$; two-sided binomial test; Supplementary Table 26). As with all previous analyses, this result was virtually identical when the psychiatric cohorts in ExAC were removed (RR = 1.14; $P = 0.02$; two-sided binomial test). Although there were far more inherited PTVs than de novo PTVs, the inherited variant effect size (1.16 RR) was paradoxically minute relative to that of de novo PTVs (3.24 RR).

Despite the different effect sizes between de novo and inherited PTVs, the data did not suggest that the two classes of variation differed in penetrance. Instead, the excess of inherited PTVs appeared to reside in different genes from those implicated by de novo variation. Specifically, the largest de novo–variant excess resided in a limited and extremely penetrant set of genes that did not substantially contribute to inherited PTV counts. Of the 11 genes with three or more class 1 de novo PTVs in ASD cases and none in controls (47 de novo PTVs in total), all were intolerant of truncating variation (pLI ≥ 0.9) (Table 1 and Supplementary Table 22). These variants conferred risk of particularly severe outcomes: of the cases with available IQ data, 14 of the 29 individuals had IQs below 70 or were unable to complete a traditional IQ test, whereas only 27% of all individuals with ASD with available IQ data in this study fell into this group ($P = 0.008$; Fisher’s exact test). Across this same gene set, there were only four inherited PTVs, only the inherited frameshift in CHD8 bore evidence of mosaic transmission ($P = 5.49 \times 10^{-3}$; two-sided binomial test; Supplementary Table 27), thus suggesting...
that it may have arisen postzygotically and is not carried by a parent. This ratio—that 80–90% of the observed variants were de novo rather than inherited in ASD cases—indicated an enormous selective pressure against mutations in these genes that is far greater than the direct selection against ASD in general (Table 1). Indeed, despite ascertaining these 11 genes on the basis of those with the most class 1 de novo PTVs in ASD, we observed a higher rate of de novo PTVs in those same genes in the ID/DD studies (37 mutations in 1,284 cases). This result underscores that selection against these variants probably arises from more severe and widespread effects on neurodevelopment and cognition. Despite the minor contribution of inherited variation in these genes, some insights from studying families may be particularly useful. Unexpectedly, one of the four inherited PTVs, a nonsense variant in ANK2, was also observed de novo in an unrelated individual with ASD, thus providing a rare instance in which the same variant was observed both inherited and de novo in two unrelated individuals with ASD, yet was absent from 60,706 individuals in ExAC (Supplementary Note).

### Case–control analysis

Having observed a significant enrichment in both de novo and inherited PTVs absent from ExAC in LoF-intolerant genes (pLI ≥0.9), we applied this same methodology to case–control cohorts. Given that the variation present in a single individual is a combination of de novo (both somatic and germline) and inherited variation, we expected to see an effect size for PTVs intermediate between that of the de novo and inherited PTVs absent from ExAC in LoF-intolerant genes. Using a published cohort of 404 ASD cases and 3,654 controls

### Table 1 Top 12 genes with three or more class 1 de novo PTVs in individuals with ASD

| Gene | ASD de novo PTVs unaffected | ASD siblings | ID/DD | T | U | Case | Control | pLI | P value |
|------|----------------------------|--------------|-------|---|---|------|---------|-----|---------|
| CHD8 | 7                          | 0            | 0     | 1 | 0 | 0    | 0       | 1   | 3.70 × 10−13 |
| ARID1B | 5                         | 0            | 11    | 0 | 0 | 0    | 0       | 1   | 1.07 × 10−8  |
| DYSK1A | 5                         | 0            | 2     | 0 | 0 | 0    | 0       | 0.9996 | 2.46 × 10−11 |
| SYNGAP1 | 5                        | 0            | 9     | 0 | 0 | 0    | 0       | 0   | 2.47 × 10−10 |
| ADNP | 4                         | 0            | 4     | 0 | 0 | 0    | 1       | 0   | 3.93 × 10−9  |
| ANK2 | 4                         | 0            | 0     | 1 | 1 | 0    | 0       | 1   | 7.07 × 10−6  |
| DSCAM | 4                         | 0            | 2     | 0 | 0 | 0    | 1       | 0   | 3.62 × 10−7  |
| SCN2A | 4                         | 0            | 7     | 0 | 0 | 0    | 0       | 0   | 1.25 × 10−6  |
| ASH1L | 3                         | 0            | 0     | 0 | 0 | 0    | 0       | 0   | 1.67 × 10−4  |
| CHD2 | 3                         | 0            | 2     | 0 | 0 | 0    | 0       | 0   | 7.81 × 10−5  |
| KDM5B | 3                         | 0            | 2     | 0 | 0 | 0    | 0       | 0   | 3.12 × 10−5  |
| POGZ | 3                         | 0            | 2     | 0 | 0 | 0    | 0       | 0   | 5.09 × 10−5  |

Twelve genes with three or more class 1 de novo PTVs in 3,982 individuals with ASD. Additionally, for each gene, the number of class 1 de novo PTVs in 2,078 unaffected ASD siblings and in 1,284 individuals with ID/DD is shown, as well as the number of singleton, LOFTEES (see URLs) high-confidence PTVs absent from ExAC that were transmitted (T) or untransmitted (U) to 4,319 individuals with ASD and present in 404 cases of ASD and 3,654 population controls. P values represent the Poisson probability of observing more than the expected number of class 1 de novo PTVs (methodology described in ref. 16 and Online Methods). ID/DD, intellectual disability/developmental delay; ASD, autism spectrum disorder; PTV, protein-truncating variant; pLI, probability of loss-of-function intolerance.
from Sweden, we first analyzed the rate of singleton synonymous variants as a control for further analyses. We found no case-control difference among those variants present or absent from ExAC (P = 0.59; Fisher’s exact test; Supplementary Table 28). In the PTV category, we observed a slight excess in singleton PTVs in cases with ASD (917 PTVs in 404 cases) compared with controls (7,259 PTVs in 3,654 controls; OR = 1.16; P = 3.13 × 10−5; Fisher’s exact test; Supplementary Table 29). This signal increased after we removed all singleton PTVs present in ExAC or in LoF-tolerant genes, thus providing what is, to our knowledge, the first instance of an exome-wide excess of PTVs demonstrated in ASD without the use of trios (128 PTVs in 404 cases, 447 PTVs in 3,654 controls; OR = 2.63; P = 1.37 × 10−16; Fisher’s exact test; Supplementary Tables 30 and 31). In agreement with the previous de novo and inherited analyses, no signal existed for the remaining 7,601 singleton PTVs (OR = 1.06; P = 0.11; Fisher’s exact test; Supplementary Table 32). Finally, removing the psychiatric cohorts from ExAC resulted in a 2.42 OR for singleton PTVs absent from ExAC in LoF-intolerant genes (133 PTVs in 404 cases, 506 PTVs in 3,654 controls; P = 1.06 × 10−16; Fisher’s exact test; Supplementary Table 33).

DISCUSSION
Here we demonstrated that ~1/3 of the de novo variants identified in neurodevelopmental disease cohorts were also present as standing variation in ExAC, thus indicating widespread mutational recurrence. Reinforcing this notion, we demonstrated that these class 2 de novo variants were enriched in more mutable CpG sites. Most importantly, however, these class 2 de novo variants conferred no detectable risk of ID/DD and ASDs, and eliminating them from our analysis improved all genetic and phenotypic associations by removing the ‘noise’ of benign variation.

We further refined the class 1 de novo PTV association by using a gene-level intolerance metric (pLI) developed by using the ExAC resource and found that all detectable mutational excess resided in 18% of genes, which were among the most strongly and recognizably intolerant of truncating mutations. Specifically, 13.5% (±2.08%, 95% CI) of individuals with ID/DD and 6.55% (±0.88%, 95% CI) of individuals with ASD, but only 2.1% (±0.6%, 95% CI) of controls, had a de novo PTV absent from ExAC and present in a gene with a very low burden of PTVs in ExAC (pLI ≥ 20.9). ASD cases with such a variant were more likely to be female and/or have intellectual disability than the overall ASD population. For the remaining 93.45% of the ASD cohort, we did not observe any meaningful phenotypic differences (e.g., in IQ or sex) between the 6.86% of individuals with and the 86.59% of individuals without a class 2 de novo PTV or a class 1 de novo PTV in a LoF-tolerant gene. These results, together with the overall lack of excess case burden, suggest that, collectively, neither class 2 nor class 1 de novo PTVs in LoF-tolerant genes (pLI < 0.9) appear to confer significant ASD risk. Thus, we refined the role of de novo protein-truncating variation in ASD, confining the signal to a smaller subset of patients than previously described.

This analysis framework, operating at the variant level, also enabled careful examination of inherited variation in ASD. Although ASD is highly heritable, a few analyses have demonstrated specific inherited components. By removing inherited PTVs present in ExAC or in LoF-tolerant genes, we identified a modest signal of over-transmitted PTVs, in line with results from previous reports. The majority of inherited PTVs appeared to affect genes that have yet to show signal from de novo variation, and only 1% of these PTVs resided in the strongest-associated genes, thus indicating that the inherited variants reside in genes with a somewhat weaker selective pressure against them.

Ultimately, however, because these variants occurred in 15.4% of cases but carried only a 1.16-fold increased risk as a group, they explain little of the overall heritability (<1% of the variance in liability).

Given the current size of ExAC and the general scarcity of truncating variants, the pLI metric for constraint against loss-of-function variation does not yet provide precise resolution of the selection coefficient acting on PTVs in that gene. That is, even a pLI ≥ 0.9 does not guarantee a selection coefficient sufficiently high to ensure that the majority of variation is de novo rather than inherited. In fact, selection coefficients for genes with pLI ≥ 0.9 range from 0.1 to 0.5, in which the majority of variation is inherited, up to nearly 1, in which the variants are almost completely reproductively null. Only larger reference panels would enable these estimates to be refined, thereby articulating a gradient from the strongest genes currently flagged (for example, the 11 genes with three or more de novo PTVs in ASD and none in controls, whose contribution is thus almost entirely through penetrant, single-generation de novo variation) to those genes that remain to be clearly defined and that make their contribution largely through inherited, albeit less penetrant, variation. The substantial expansion of exome sequencing in ASD, alongside larger reference panels from which to draw more precise inferences about selective pressure acting on variation in each gene, would allow further elucidation of the genetic architecture of ASDs in the region of the effect-size spectrum from severe de novo variation at one end to common variation at the other.

ExAC currently comprises 15,330 individuals from psychiatric cohorts, the largest of which is the schizophrenia cohort. Given the shared genetics between ASD and schizophrenia, it is reasonable to hypothesize that the psychiatric cohorts within ExAC might have influenced our analyses. As we have shown, however, removing the psychiatric cohorts within ExAC did not change our results. In fact, of the 16 de novo PTVs in LoF-intolerant genes that were also variant in ExAC, only two resided solely in the 15,330 individuals from the psychiatric cohorts (CUX2 in ASD and LARP1 in unaffected ASD siblings). This small number is, in retrospect, not surprising because it is so unusual to observe a deleterious variant both de novo and as standing variation in individuals with the same ascertained phenotype, let alone in different ascertained phenotypes. The ANK2 nonsense variant was the only such instance of the same deleterious variant being de novo in one ASD trio and inherited in another.

Although we used ASDs and ID/DD here to explore this framework, it could certainly be applied toward any trait. However, this framework is optimally powered in traits governed by genes under strong selection, because it removes de novo variants that are more common in the context of a larger reference population. Our results reinforce that not all de novo variants are rare and contribute to risk, while highlighting the tremendous value of large population sequence resources, even for the interpretation of de novo variation and complex disease. This aspect is especially important in the case of clinical sequencing, in which the paradigm has unfortunately become that if a protein-altering de novo variant is present in the gene of interest, then it is often considered the causal variant. Clearly, not all de novo variants are equal, and not all de novo variants in a gene contribute to risk in the same manner.

METHODS
Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

URLs. Exome Aggregation Consortium (ExAC), http://exac.broadinstitute.org; LOFTEE, https://github.com/konradjk/loftee.
Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all of the members of the ATGU and the laboratory of D.P.W. for assistance in this endeavor. We thank the families who took part in the Simons Simplex Collection study and the Simons Variation in Individuals Project, as well as the clinicians who collected data at each of the study sites. The authors would like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found on the ExAC website (see URLs). We also greatly thank A. Byrnes, R. Fine, D. Frank, A. Martin, C. Nichols, N. Radd, K. Satterstrom, and E. Wigdor for their insightful contributions. Finally, we acknowledge G.A. Barnard for inspiring us to write in a more conversational tone similar to that in his seminal 1947 paper (Biometrika 34, 123–138, 1947). This work was supported by NIH grants U01MH100233, U01MH100209, U01MH100229, and U01MH100239 to the Autism Sequencing Consortium (ASC), and R56MH97849 and R01MH97849 to the Population-based Autism Genetics and Environment Study (PAGES). M.J.D., J.A.K., and K.E.S. were supported by grants from the Simons Foundation Autism Research Initiative (SFARI 342292 and a subaward from the Simons Center for the Social Brain at the Autism Sequencing Consortium (ASC), and the Human Genome Research Institute (HG002295/HG/NHGRI). E.B.R. was funded by National Institutes of Mental Health (MIT). M.L. and D.G.M.’s work on the ExAC project was funded by U54DK105566 (SFARI 342292 and a subaward from the Simons Center for the Social Brain at the Autism Sequencing Consortium (ASC), and R54MH097849 and R01MH097849 to the Population-based Autism Genetics and Environment Study (PAGES)).

AUTHOR CONTRIBUTIONS

J.A.K. and E.B.R. performed the analyses. J.A.K., D.P.W., B.R., and M.J.D. designed the experiment. J.A.K. and K.S. wrote the code. D.P.W., E.B.R., and M.J.D. supervised the research. J.A.K. and M.J.D. wrote the paper. J.A.K., K.E.S., D.P.H., S.I.S., M.L., K.J.K., D.G.M., and J.D.B. generated data. J.A.K., K.E.S., D.P.H., D.J.C., B.R., B.M.N., D.G.M., D.P.W., E.B.R., and M.J.D. contributed to analysis concepts and methods. J.A.K. was responsible for the remainder. All authors revised and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Autism and Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators. Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2010. MMWR Surveill. Summ. 63, 1–21 (2014).
2. Lee, S.H. et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat. Genet. 45, 986–994 (2013).
3. Gaugler, T. et al. Most genetic risk for autism resides with common variation. Nat. Genet. 46, 881–885 (2014).
4. Klein, L. et al. Common genetic variants, acting additively, are a major source of risk for autism. Mech. Med. Dev. 3, 9 (2012).
5. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515, 209–215 (2014).
6. Iossifov, I. et al. De novo coding mutations contributing to autism spectrum disorder. Nature 515, 216–221 (2014).
7. Iossifov, I. et al. De novo gene disruptions in children on the autistic spectrum. Neuron 74, 285–299 (2012).
8. Neale, B.M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485, 242–245 (2012).

1Analytic and Translational Genetics Unit (ATGU), Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.
2Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.
3Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.
4Program in Bioinformatics and Integrative Genomics, Harvard University, Cambridge, Massachusetts, USA.
5Program in Genetics and Genomics, Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA.
6Department of Psychiatry, University of California, San Francisco, San Francisco, California, USA.
7Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, Massachusetts, USA.
8Partners Center for Personalized Genetic Medicine, Boston, Massachusetts, USA.
9Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, USA.
10Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA.
11Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.
12Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, New York, USA.
13Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA.
14Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York, USA.
15Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA.
16Mindhich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York, USA.
17Department of Pediatrics (Systems Medicine), Biomedical Data Science, and Psychiatry (by courtesy), Stanford University, Stanford, California, USA.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.
**ONLINE METHODS**

**Data sets and data processing.** Two versions of the Exome Aggregation Consortium (ExAC) database were used in this analysis: the full version of ExAC (n = 60,706) and the nonpsychiatric version of ExAC (n = 45,376). The nonpsychiatric version of ExAC has the following cohorts removed: Bulgarian trios (n = 461), sequencing in Szumi (n = 948), Swedish schizophrenia as bipolar studies (n = 11,219), schizophrenia trios from Taiwan (n = 1,505), and Tourette syndrome association international consortium for genomics (n = 297). We used a combined set of 8,401 published de novo variants from 3,982 probands with ASD and 2,078 of their unaffected siblings from two recent large-scale exome sequencing studies: de Rubeis et al.\(^\text{a}\) (n\(_{\text{ASD}}\) = 1,474, \(n_{\text{unaffected,sib}}\) = 267) and Lissikov et al.\(^5\) (n\(_{\text{ASD}}\) = 2,508, \(n_{\text{unaffected,sib}}\) = 1,911) (Supplementary Table 1). We also used 1,692 de novo variants from 1,284 probands published in studies of ID (de Ligt et al.\(^{12}\), n = 100; Rauch et al.\(^{14}\), n = 51) and DD (ref. 13, n = 1,133) (Supplementary Table 2). De novo variants from congenital heart disease\(^{26,27}\) and schizophrenia\(^5\) were also downloaded for additional confirmation of the recurrent mutation rate (Supplementary Tables 5 and 6). Details of the sequencing and de novo calling can be found in the referenced publications.

To ensure uniformity in variant representation and annotation across data sets and with respect to the ExAC reference database\(^4\), we created a standardized variant representation through a Python implementation of vt normalize\(^35\) and reannotated all variants with Variant Effect Predictor (VEP)\(^{36}\) version 81 with GENCODE v19 on GRCh37. VEP provided the Ensembl Gene IDs, gene symbols, the Ensembl transcript IDs for use in determining canonical transcripts, as well as PolyPhen2 and SIFT scores. We used the canonical transcript when possible for cases in which the variant resided in multiple transcripts. When the variant resided in multiple canonical transcripts, the canonical transcript with the most deleterious annotation was used. If no canonical transcript was available, the most deleterious annotation was used. As such, variants in Supplementary Tables 1 and 6 may differ from their respective publications, owing to standardizing variant representation and annotation.

**Determining class 1 or class 2 de novo variants.** De novo variants were classified as class 1 or class 2 on the basis of their absence or presence, respectively, in ExAC. Presence or absence in ExAC was defined if the variant had the same chromosome, position, reference, and alternate allele in both files. Owing to the heterogeneous nature of ExAC, and the different capture arrays used in the original exome sequencing studies incorporated into ExAC, we elected to use all of the variants in ExAC, not just those with a PASS status in the GATK variant-calling filter. For insertions/deletions, we used a conservative stance requiring the variants to match exactly (i.e., a subset was not sufficient). For example, if a de novo variant on chromosome 5 at position 77,425,626 were to have a reference allele of AGATG and a alternate allele in which 4 nt were deleted (AGATG to A), we would not say that the variant is present in ExAC if another variant was present at the same genomic position in ExAC and had only the first two of these nucleotides deleted (AGA to A). Finally, for variants outside of the proportion of the genome covered by ExAC, we considered them to be class 1 de novo variants; as expected, none of these variants resided in the coding region (Supplementary Table 34).

**Variant calling for transmission and case–control analysis.** We used the Genome Analysis Toolkit (GATK v3.1-144) to re-call a data set of 22,144 exomes from the Autism Sequencing Consortium (ASC)\(^{37}\) and Simons Simplex Collection (SSC)\(^{38}\) sequencing efforts. This call set contained 4,319 complete trios (including all those from which the published and validated de novo mutations were identified), which we used to evaluate inherited variation, and a published case–control data set of individuals of Swedish ancestry (404 individuals with ASD and 3,564 controls)\(^5\). We applied a series of quality-control filters on the genotype data, using the genome-wide transmission rate as a guide for filter inclusion/exclusion. More specifically, we calibrated various genotyping filters such that synonymous singleton variants (in which the alternate allele was seen in only one parent in the data set) were transmitted at a rate of 50%, because we expected synonymous variants, as a class, to be transmitted 50% of the time. As with the ExAC analysis\(^4\), we found GATK’s default Variant Quality Score Recalibration (VQSR) to be too restrictive, owing to the bias toward common sites. To decrease the number of singleton variants being filtered out, we recalibrated the Variant Quality Score Log Odds (VQSLOD) threshold from ~1.49 to ~1.75, dropping the singleton synonymous transmission rate from 51.1% to 49.9998% (Supplementary Fig. 4). Additional filtering was performed at the individual level, in which we required a minimum read depth of 10 and a minimum GQ and PL of 25 for each individual’s variant call. We also applied an allele balance filter specific for each of the three genotypes (homozygous reference, heterozygous, and homozygous alternate), in which allele balance was defined as the number of alternate reads divided by the total number of reads. We required the allele balance for homozygous reference individuals to be less than 0.1, the allele balance for heterozygous individuals to be between 0.3 and 0.7, and the allele balance for homozygous alternate individuals to be greater than 0.9. Calls that did not pass these filters were set to missing. Finally, for the transmission analysis, we removed variants in which more than 20% of families failed one of our filters. For the case–control analysis, we removed variants in which more than 5% of the individuals failed one of our filters.

**On the use of the Poisson exact test for comparing rates of de novo variation between two samples.** As has been done in many other papers\(^6–8,39–41\), we sought to test whether the rate of a given class of de novo variation was significantly different between our cohorts of individuals with ASD or ID/DD as compared with unaffected ASD siblings. Because the number of de novo variants per individual follows a Poisson distribution\(^6\), we tested \(H_0: \lambda_1 \not= \lambda_2\) versus \(H_0: \lambda_1 = \lambda_2\), where \(\lambda_i\) is the rate of a given class of de novo variation in group \(i\), on the basis of the Poisson exact test (also known as the C test)\(^28\). We were unable to compare the observed rates against the expected rate, because the expectations published in Samocha et al.\(^18\) are for all de novo variants, not just de novo variants present/absent from ExAC. An important consequence of our hypothesis test is that effect sizes are reported as rate ratios (RR), which is simply the quotient of the two rates. Although they are more commonly reported, odds ratios require Bernoulli random variables (for example, an individual either has or does not have a de novo variant), and as such would be incorrect, given the hypothesis being tested. Had we sought to test for a significant difference between the proportion of individuals with a de novo PTV, then an odds ratio would have been appropriate (and Fisher’s exact test would have sufficed). Thus, only by using the Poisson exact test could we reject the null hypothesis that the rate of de novo PTVs is the same between individuals with ASD and their unaffected siblings and find evidence that individuals with ASD have a higher rate of de novo PTVs than their unaffected siblings. The difference between the two tests is a subtle but important one.

**On the use of pLI (probability of loss-of-function intolerance).** Using the observed and expected number of PTVs per gene in the ExAC data set, we developed a metric to evaluate a gene’s apparent intolerance to such variation\(^24\). Briefly, the probability of loss-of-function intolerance (pLI) was computed with an EM algorithm that assigned genes to one of three categories: fully tolerant (in which PTVs were presumed to be neutral and, like synonymous variants, to occur at rates proportional to the mutation rate), ‘recessive-like’ (showing PTV depletion similar to that in known severe autosomal recessive diseases) and ‘haploinsufficient-like’ (showing PTV depletion similar to that in established severe haploinsufficiencies). pLI is the posterior probability that a gene resides in the last category which is, least tolerant of LoF (more severe haploinsufficiencies).}

Phenotype analysis. Full-scale deviation IQ scores were measured with several tests including differential ability scales, the Wechsler intelligence scale for children, and the Wechsler abbreviated scale of intelligence. IQ has previously been associated with the de novo PTV rate in the SSC\(^{25,26,27}\). In this analysis, we used Poisson regression to estimate the relationship between the rate of each of class 1 and class 2 PTVs and proband full-scale deviation IQ.

Calculating the expected number of class 2 de novo variants in a reference database. For a set of \(r\) de novo variants, each with the same allele count, \(K\), in ExAC, we can estimate the number of those variants still observed at least once in a subset of size \(n\), by using the hypergeometric distribution. That is,
how many of those same sites will still be present as standing variation in a
downsampling version of ExAC? Specifically,
\[
\text{expected count} = r (1 - P(k = 0)) \approx r \left( 1 - \frac{\binom{K}{n} \binom{N-K}{n-K}}{\binom{N}{n}} \right) = r \left( 1 - \frac{\binom{N-K}{n}}{\binom{N}{n}} \right)
\]

where \(k\) is approximately hypergeometric \((N,K,n)\), and \(N\) is the number of chromosomes in the current version of ExAC, which at this point is 121,412.

This assumption holds only when each downsampled set of ExAC preserves the ancestry proportions of the total sample.

Calculating mutation rates for class 1 and class 2 de novo PTVs. Samocha et al.\(^{16}\) calculated per-gene mutation rates for all synonymous variants, missense variants, and PTVs, not for those present/absent in ExAC. The rate of class 1 depth-corrected mutation rate for class 1 PTVs can be roughly calculated for comparison with the expected de novo mutation rate for PTVs, denoted \(\hat{\mu}_{\text{PTV}}\). With equation (1):

\[
\hat{\mu}_{\text{PTV}} = \hat{\mu}_{\text{class 1 PTV}} + \hat{\mu}_{\text{class 2 PTV}}
\]

In case the logic behind equation (1) is not completely clear, the number of class 1 and class 2 PTVs is equal to the total number of PTVs. Samocha et al. provided \(\hat{\mu}_{\text{PTV}}\), so only \(\hat{\mu}_{\text{class 1 PTV}}\) and \(\hat{\mu}_{\text{class 2 PTV}}\) need to be calculated.

Given all of the PTVs in ExAC, and the probability of each trinucleotide-to-trinucleotide mutation, we can calculate \(\hat{\mu}_{\text{class 2 PTV}}\) with equation (2):

\[
\hat{\mu}_{\text{class 2 PTV}} = \sum_{i}^{\text{PTVs}} \hat{\mu}_{\text{SNPi}}
\]

where \(i\) indexes the \(n\) PTVs for a given gene present in ExAC, and \(\hat{\mu}_{\text{SNPi}}\) is the mutation rate of that specific trinucleotide substitution that creates a PTV. With \(\hat{\mu}_{\text{class 2 PTV}}\) calculated, \(\hat{\mu}_{\text{class 1 PTV}}\) follows from equation (1). However, these per-gene \(\hat{\mu}_{\text{PTV}}\) calculations do not account for sequencing depth. Correcting for depth of sequencing becomes tricky, because the depth of sequencing varies among studies and will not necessarily be the same as the depth of sequencing for ExAC. However, we can roughly approximate the depth-corrected \(\hat{\mu}_{\text{class 2 PTV}}\) for each gene by using the following equation under the assumption that the fraction of the raw mutability from class 2 is equal to the fraction of the class 2 depth-corrected mutability:

\[
\hat{\mu}_{\text{class 2 PTV}, \text{depth corrected}} = \frac{\hat{\mu}_{\text{class 2 PTV}}}{\hat{\mu}_{\text{PTV}}}
\]

The depth-corrected \(\hat{\mu}_{\text{class 1 PTV}}\) follows according to the same logic as in equation (1).

Data availability. Data included in this manuscript have been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession number phs000298.v2.