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

Jack A. Kosmicki[1-4], Kaitlin E. Samocha[1-3,5], Daniel P. Howrigan[1-3], Stephan J. Sanders[6], Kamil Slowikowski[2,4,7,8], Monkol Lek[1,2], Konrad J. Karczewski[1,2], David J. Cutler[9], Bernie Devlin[10], Kathryn Roeder[11], Joseph D. Buxbaum[12-17], Benjamin M. Neale[1-3], Daniel G. MacArthur[1,2], Dennis P. Wall[18], Elise B. Robinson[1-3], Mark J. Daly[1-3]

[1] Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA
[2] Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA
[3] Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[4] Program in Bioinformatics and Integrative Genomics, Harvard University, Cambridge, MA 02138, USA
[5] Program in Genetics and Genomics, Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02446, USA
[6] Department of Psychiatry, University of California, San Francisco, San Francisco, California 94158, USA
[7] Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02446, USA
[8] Partners Center for Personalized Genetic Medicine, Boston, MA 02446, USA
[9] Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA
[10] Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA
[11] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[12] Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[13] Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[14] Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[15] Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[16] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[17] Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[18] Department of Pediatrics and Psychiatry (by courtesy), Division of Systems Medicine, Stanford University, CA, 94305, USA

Address for Correspondence:

Mark J. Daly
Chief, Analytic and Translational Genetics Unit
Massachusetts General Hospital

617-643-3290
mjdaly@atgu.mgh.harvard.edu

Abstract
Recent research has established the significant role of newly arising (*de novo*) variation in neurodevelopmental disorders. Using aggregated data from 5603 families affected by ASD, intellectual disability (ID), or developmental delay (DD), we show that overall ~1/3 of *de novo* variants are independently observed as standing variation in the Exome Aggregation Consortium's (ExAC) cohort of 60,706 adults without severe neurodevelopmental disease, and that these *de novo* variants do not contribute substantively to neurodevelopmental risk. We further use a loss-of-function intolerance metric derived from ExAC to identify a small subset of genes that contain the entire signal of associated *de novo* protein truncating variants (PTVs) in neurodevelopmental disorders (ASD: rate ratio (RR) = 3.24, $P < 10^{-16}$; ID/DD: RR = 6.70, $P < 10^{-38}$). These same genes carry a moderate excess of PTVs inherited by individuals with ASD from their unaffected parents (RR = 1.16, $P = 0.00985$), and a larger excess in 404 ASD cases and 3654 controls (OR = 2.66; $P = 6.34 \times 10^{-19}$), with again no association seen outside these genes. These findings illustrate the importance of population-based cohorts as reference samples for the interpretation of candidate pathogenic variants, even for the analysis of *de novo* variation.

Introduction
Autism spectrum disorders (ASD) are a phenotypically heterogeneous group of heritable disorders that affect approximately 1 in 68 individuals in the US[1]. While estimates of the common variant (heritable) contribution toward ASD liability are upwards of 50%[2-4], few specific risk variants have been identified as ASD GWAS sample sizes to date are far smaller than those required to make the type of progress seen in schizophrenia[5] and other neuropsychiatric phenotypes. Conversely, the field has made substantial progress via analysis of *de novo* (newly arising) variation using exome sequencing of parent-offspring trios[6-11]. Severe intellectual disability and developmental delay (ID/DD) are considerably less heritable than ASDs[12] (though frequently comorbid) and have demonstrated, perhaps unsurprisingly, a much stronger contribution from *de novo* frameshift, essential splice site, and nonsense variants (collectively termed protein truncating variants [PTVs])[13-15]. Furthermore, ASD cases with comorbid ID show a significantly higher rate of *de novo* PTVs than those with normal or above average IQ[7,10,16-18], while higher IQ cases have a stronger family history of neuropsychiatric disease[16], suggesting a greater heritable contribution.

De novo variants comprise a unique component of the genetic architecture of human disease since, not having yet passed through a single generation, any heterozygous variants with complete or near-complete elimination of reproductive fitness must reside exclusively or almost exclusively in this category. Despite prior evidence of mutational recurrence[19] (i.e., the same mutation occurring *de novo* in multiple individuals), most studies have implicitly assumed each *de novo* variant was novel, in line with Kimura's infinite sites model[20] and thereafter analyzed *de novo* variants genome-wide without respect to their allele frequency in the population (Supplementary Note). However, we know that the mutation rate is not uniform across the genome, with some portions experiencing higher mutation rates than others (e.g., CpG sites[21]). A classic example comes from Achondroplasia, in which the same G-to-C and G-to-T variant at a CpG site was observed *de novo* in 150 and 3 families, respectively[19]. As such, it should not be surprising to observe a *de novo* variant at a given site, and also to 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 disorders (e.g., ASD, intellectual disability, and schizophrenia)[22-24], we expect most highly deleterious (high-risk conferring) *de novo* PTVs will linger in the population for at most a few generations. Thus, these variants' average allele frequency in the population will approximate their mutation rate. Individual PTVs tolerated enough to be seen in relatively healthy adults, and more generally PTVs in genes that clearly tolerate the survival of such variants in the population, might be less likely to contribute significant risk to such phenotypes with an allele frequency orders of magnitude larger than that 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), logic (i.e., the pigeonhole principle) dictates that every non-embryonically lethal coding mutation is present as a *de novo* variant in the human population. This reasoning, along with the availability of large exome sequencing 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, indicating the presence of a recurrent mutation. We will herein refer to these *de novo* variants that are also observed as standing variation in the population as **class 2 *de novo* variants** (observed at least twice), with the remaining *de novo* variants referred to as **class 1 *de novo* variants** (i.e., observed only once) (**Fig. 1**).

With the release of the Exome Aggregation Consortium's (ExAC) rich dataset of 60,706 adult individuals without severe developmental abnormalities[25], we can now investigate this claim in detail. While there have been many studies examining *de novo* variation in human disease, the success in ASD and ID/DD, coupled with the large sample sizes published to date, led us to use these phenotypes in this evaluation.

Results

Class 2 *de novo* variation

We first asked how many of the 10,093 variants observed *de novo* in ID/DD cases, ASD cases, and unaffected ASD siblings are also seen as standing variation in the 60,706 reference exomes from ExAC[25] (**Fig. 1**; Online Methods). We found that 1854 ASD (31.66%), 841 unaffected ASD sibling (33.05%), and 409 ID/DD (24.27%) *de novo* variants are observed as standing variation in one or more ExAC individuals (class 2 *de novo* variants) (**Fig. 2A**; **Supplementary Tables 1** and **2**). Similar rates of class 2 *de novo* variation were also seen in published studies of *de novo* variants in schizophrenia[26] and congenital heart disease[27] (**Supplementary Tables 3** and **4**; Online Methods). While the presence of class 2 *de novo* variants is not a novel observation[19,26], the rate is approximately three times larger than previous estimates[26] owing to significantly larger reference datasets (**Supplementary Fig. 1**; **Supplementary Note**).

We first sought confirmation that the observed recurrence rate – the proportion of variants observed both *de novo* and as standing variation in the population – was technically sound and not the result of some undetected contamination or missed heterozygote calls in parents (i.e., false *de novo* variants). Five secondary analyses strongly support the technical validity of this work. 1) In line with previous publications[26], class 2 *de novo* variants, regardless of their functional impact, are enriched at CpG sites as compared to class 1 *de novo* variants ($P < 1 \times 10^{-20}$; two-sided Fisher's exact test; **Supplementary Table 5**). 2) As the exomes used to call *de novo* variants in De Rubeis, et al. (2014) were used in the joint calling of ExAC, and many were

sequenced at the same center as the majority of ExAC samples, it is possible that false class 2 *de novo* variants could be elevated in this dataset via contamination or joint calling artifact. However, no difference in the rate of class 2 *de novo* variation between De Rubeis, et al., (2014) and Iossifov, O'Roak, Sanders, Ronemus, et al., (2014) was observed (P = 0.10; Fisher's exact test; **Supplementary Table 6**). 3) The frequency distribution of class 2 *de novo* variants should be skewed dramatically upward towards common variation if contamination or missed parental heterozygotes were contributing – however the ExAC frequency of class 2 *de novo* variants at CpGs were indistinguishable from all such variants in ExAC (comparing to variants drawn at random with the same annotation and CpG content) (P = 0.14; Wilcoxon rank sum test; **Fig. 2B**). 4) A subset of synonymous variants specifically experimentally validated in the ASD studies showed nearly the same recurrence rate as the overall set, most definitively establishing that these mutations are indeed arising independently in the ASD sample set (P = 0.60; Fisher's exact test). 5) Lastly, it is well known that mutation rate increases with paternal age[28-31], thus rates of both class 1 and class 2 *de novo* variants should show association with paternal age if both classes were indeed genuine *de novo* variants. Indeed, for the 1861 unaffected ASD sibling with available paternal age information, rates of both class 1 and class 2 *de novo* variants are significantly associated with increasing paternal age (class 1: $\beta = 0.002$, $P = 4.11 \times 10^{-9}$; class 2: $\beta = 0.0009$, $P = 1.06 \times 10^{-5}$; linear regression).

Having confirmed a substantial portion of *de novo* variants were observed independently as standing variation in other ExAC samples, we then sought to determine whether class 1 and class 2 *de novo* variants contribute equally to ASD and ID/DD risk. As a control for the comparison of functional variants, rates of both class 1 and class 2 *de novo* synonymous variants are equivalent across ASD, ID/DD, and unaffected ASD siblings (**Fig. 3A**; **Supplementary Table 7**). Thus regardless of whether a *de novo* synonymous variant is class 1 or class 2, collectively neither show association with ASD or ID/DD, consistent with our expectation based on previous reports that as a class, *de novo* synonymous variants do not contribute to risk[6-11]. While *de novo* PTVs have been previously implicated as significant risk factors for ASD[6,7,16,17] and ID/DD[14], the class 2 *de novo* subset of PTVs show no such association for either ASD (0.015/case vs. 0.023/unaffected ASD sibling; P = 0.98; one-sided Poisson exact test[32]) or ID/DD (0.016/case vs. 0.023/unaffected ASD sibling; P = 0.94; one-sided Poisson exact test), with slightly higher rates in unaffected ASD siblings (**Fig. 3B**; **Supplementary Table 8**). By contrast, after removing class 2 *de novo* PTVs, class 1 *de novo* PTVs are even more highly enriched in individuals with ASD (0.13/case) and ID/DD (0.19/case) as opposed to unaffected ASD siblings (0.07/control) (ASD vs. control: rate ratio (RR) = 1.83; $P = 6.07 \times 10^{-12}$, ID/DD vs. 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 is consistent with what would be expected if such variants were neutral and unassociated. The data provides no evidence to suggest these class 2 *de novo* variants contribute to the previously observed enrichment of *de novo* variation in ASD and ID/DD cases, and removing those variants present in ExAC increases the rate ratio for *de novo* PTVs in ASD and ID/DD (**Fig. 3B**; **Supplementary Table 8**). **Thus for subsequent analyses, we focus on variation absent from ExAC**.

Gene level analyses

The analyses above focused solely on single variants, asking whether a given *de novo* variant was present or absent in the standing variation of ExAC, and whether that variant was more or

less likely to confer risk for ASD or ID/DD. Extending this concept, we sought to evaluate whether the overall distribution of PTVs in a gene in ExAC provided a similar guide to which ASD and ID/DD variants were relevant. Specifically, we investigated whether the gene-level constraint metric, pLI[16] (probability of being loss-of-function intolerant), can improve our ability to decipher which class 1 *de novo* PTVs confer the most risk to ASD and ID/DD (Online Methods). Using the same threshold as used in Lek et al., 2015, we defined a loss-of-function (LoF) intolerant gene as one with a pLI ≥0.9. Under the assumption that genes sensitive to PTVs are enriched in ASD, we investigated whether affected ASD probands had an increased burden of such class 1 *de novo* PTVs as compared to their unaffected siblings. When we restricted to solely class 1 *de novo* PTVs in LoF intolerant genes (pLI ≥0.9), we observed significant excess in affected ASD probands (0.067 events per exome) compared to their unaffected siblings (0.021 events 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 becomes more striking, with a rate of 0.139 per exome, resulting in an RR of 6.70 when compared to the control group of unaffected ASD siblings ($P = 6.34 \times 10^{-38}$; one-sided Poisson exact test). By contrast, the remaining *de novo* PTVs (class 1 *de novo* PTVs in tolerant genes [pLI <0.9] and class 2 *de novo* PTVs) show no difference between individuals with ASD and their unaffected siblings (0.078 vs. 0.073; $P = 0.26$; one-sided Poisson test), or individuals with ID/DD and unaffected ASD siblings (0.063 vs. 0.073; $P = 0.87$; one-sided Poisson exact test). Hence, all detectable *de novo* PTV signal in both ASD and ID/DD can be localized to 18% of genes with clear intolerance to PTVs in ExAC, with, consequently, substantially amplified rate ratios in this gene set (**Fig. 3C**; **Supplementary Table 9**).

Recent studies inferred the presence of multiple *de novo* PTVs in the same gene as evidence of a contribution to ASD risk[6-11]. Thirty-eight genes had two or more *de novo* PTVs in ASD cases and none in their unaffected siblings, while 13 genes had multiple *de novo* PTVs among controls or in cases and controls (**Supplementary Table 10**). This not only reinforces previous analyses that established the mere observation of multiple *de novo* PTVs is not sufficient to define a gene as important[6,17], but also provides an opportunity to explore whether the pLI metric can refine the implication of specific genes. In fact, 32 of the 38 case-only genes, but only 5 of 13 control-only or case-control hit genes, are in the pLI ≥ 0.9 set – a highly significant difference (OR > 10, $P < 0.001$) that greatly refines the list of genes to be pursued as likely ASD contributors.

Phenotypic associations for *de novo* PTVs absent in ExAC and in loss-of-function intolerant genes

While enrichment of *de novo* PTVs is one of the hallmarks of ASD *de novo* studies[6-11,16,17], another consistent finding is an increased burden of these variants among females with ASD[7,16] and in ASD individuals with low full-scale IQ (FSIQ)[7,16,17]. We investigated whether these hallmarks were present in the 6.6% of affected probands with a class 1 *de novo* PTV in LoF intolerant genes (pLI ≥0.9). Indeed, females are overrepresented in the subset (12.26% of females; 5.80% males; $P = 1.75 \times 10^{-5}$; two-sided Fisher's exact test; **Supplementary Table 11**). Importantly, for the 6.86% of affected ASD probands with a class 2 *de novo* PTV or a class 1 *de novo* PTV in a tolerant gene (pLI < 0.9), there is no difference between the sexes, with 6.86% of females and 6.83% of males falling in this category ($P = 1$; two-sided Fisher's exact test; **Supplementary Table 12**). This same pattern is observed with average FSIQ (Online Methods): individuals with a class 1 *de novo* PTV in a LoF intolerant gene have an average FSIQ of 82.21 ±

18.34, lower than that of both the remaining individuals with a class 2 *de novo* PTV or tolerated class 1 *de novo* PTV (FSIQ = 91.63 ± 20.09) and individuals without a *de novo* PTV (FSIQ = 92.28 ± 20.20). Furthermore, class 2 *de novo* PTVs and class 1 *de novo* PTVs in pLI < 0.9 genes show no association with IQ (β = -0.001; P = 0.76; Poisson regression), while class 1 *de novo* PTVs in LoF intolerant genes predominately explain the skewing towards lower IQ (β = -0.023; $P = 7 \times 10^{-8}$; Poisson regression; **Fig. 4**). These secondary analyses strongly support the implication of the primary analysis, that class 2 *de novo* variants and class 1 *de novo* variants in LoF tolerant genes (pLI < 0.9) have little to no association to ASD or ID/DD and no observable phenotypic impact on those cases carrying them. By contrast, the class 1 *de novo* variants occurring in LoF intolerant genes contain the association and phenotypic skewing observed to date.

Inherited variation

As the association signal for *de novo* PTVs increased after removing those variants present in ExAC, we postulated a similar increase could be obtained from rare inherited PTVs. With the bulk of ASD liability residing in inherited space[3,18], risk-conferring variants should be transmitted more often to individuals with ASD. We tested for transmission disequilibrium in a cohort of 4319 trios with an affected proband (Online Methods). Without filtering by pLI or presence/absence status in ExAC, singleton PTVs, as a class, showed no sign of over-transmission (12,231 transmitted, 12,073 untransmitted; P = 0.31; binomial test). After removing all of the variants present in ExAC or in a LoF tolerant gene (pLI <0.9), we found a modest, but statistically significant, excess of transmitted PTVs in individuals with ASD (667 transmitted, 576 untransmitted; RR = 1.16; P = 0.00985; binomial test; **Supplementary Table 13**). While this constitutes a much larger number of variants than the number of *de novo* observations in this category, the effect size is paradoxically minute by comparison with the 3.24 RR seen in the *de novo* PTV category.

Rather than suggesting a mysterious difference between the penetrance of the same variants depending on *de novo* status, the data suggest that different genes are likely responsible for the excess of transmitted alleles than those implicated from *de novo* variation. Specifically, the largest *de novo* variant excesses reside in a limited and extremely penetrant set of genes that do not contribute substantially to inherited PTV counts. If we consider the 11 genes with 3 or more class 1 *de novo* PTVs in ASD cases and none in controls (47 *de novo* PTVs in total), all 11 are intolerant of truncating variation (pLI ≥ 0.9) (**Table 1**; **Supplementary Table 10**). These variants confer risk to particularly severe outcomes as, of the cases with available IQ data, nearly half have IQ < 70 or were unable to complete a traditional IQ test[16] (14 individuals / 29 individuals with available data) while overall only 27% of ASD individuals with with available IQ data in this study fall into this group (P = 0.008; two-sided Fisher's exact test). Across this same gene set, there are only 6 inherited PTVs (from a total of 7 observed in the parents of the 4319 ASD trios). This ratio – in which 80-90% of the observed variants are *de novo* rather than inherited in ASD cases – indicates enormous selective pressure against mutations in these genes (far greater than the direct selection against ASD in general). Indeed, despite selecting these genes based on being the most often mutated in ASD, we observe a higher rate of *de novo* PTVs in these genes in the ID/DD studies (37 mutations / 1284 cases) underscoring that selection against these variants likely arises from more severe and widespread impact on neurodevelopment and cognition (**Table 1**). Despite the minor contribution of inherited

variation in these genes, some insights from studying families may be particularly useful. To our surprise, one of the 6 inherited PTVs, a nonsense variant in *ANK2*, was also observed *de novo* in an unrelated ASD proband, providing a rare instance in which the same variant was observed inherited and *de novo* in two unrelated individuals with ASD and also absent in 60,706 individuals from ExAC (**Supplementary Note**).

Case-control analysis

Having observed a significant enrichment in both *de novo* and inherited PTVs absent from ExAC in pLI ≥0.9 genes, we applied this same methodology to case-control cohorts. Given that the variation present in a single individual will be a combination of *de novo* (both somatic and germline) and inherited variation, we expect to see an association for PTVs intermediate between the *de novo* and inherited analyses for PTVs absent from ExAC in LoF-intolerant genes. Using a published cohort of 404 ASD cases and 3654 controls from Sweden[6], we first analyzed singleton synonymous variants as a control for further analyses. Following the analysis of *de novo* synonymous variants, we first tested all synonymous variants, then split the data by whether the synonymous variants were present in or absent from ExAC. While all synonymous variants are nominally significant (OR = 1.05, P = 0.01; Cochran-Mantel-Haenszel test), neither synonymous variants present in ExAC nor absent from ExAC show association after using a Bonferroni correction for three tests (in ExAC: P = 0.02; not in ExAC: P = 0.29; Cochran-Mantel-Haenszel test with α=0.0167). In contrast, we observed a highly significant excess of singleton PTVs absent from ExAC in LoF intolerant genes (pLI ≥0.9) among ASD cases (128 PTVs / 404 cases) when compared to controls (447 PTVs / 3654 controls) resulting in a 2.66 OR ($P = 6.34 \times 10^{-19}$; Cochran-Mantel-Haenszel test; **Supplementary Table 14**).

Discussion

Here we have shown that ~30% of *de novo* variants identified in neurodevelopmental disease patients are also present as standing variation in a large panel of population reference samples (ExAC), indicating the presence of widespread mutational recurrence. Furthermore, we demonstrated these class 2 *de novo* variants are enriched for more mutable CpG sites, and in some cases represent intermediate-frequency polymorphisms in populations under-represented in previous reference sequencing panels. But most importantly, these class 2 *de novo* variants confer no detectable contribution to risk of ID/DD and ASDs, and their elimination in 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 using a gene-level intolerance metric (pLI) developed using the ExAC resource and identified that all detectable mutational excess was found in 18% of genes most strongly and recognizably intolerant of truncating mutation. Specifically, 13.5% of ID/DD patients and 6.6% of individuals with ASD, but only 2% of controls, have a *de novo* PTV absent from ExAC and present in a gene with a very low burden of PTVs in ExAC (pLI ≥0.9). ASD cases with such a variant are more likely to be female and have intellectual disability than the overall ASD population. For the remaining 93.4% of the ASD cohort, neither class 2 nor class 1 *de novo* PTVs in LoF tolerant genes (pLI <0.9) appear to play a role in their ASD diagnosis, nor do we observe any meaningful phenotypic distinction between them. Thus we have refined the role of *de novo* protein truncating variation in ASD, confining the signal to a smaller subset of patients than previously described[7,33].

This analysis framework, operating at the variant level, also enabled a careful examination of inherited variation in ASD. While ASD is highly heritable[3], few analyses[34] have demonstrated specific inherited components. By removing inherited PTVs present in ExAC or in a loss-of-function tolerant gene, we discovered a modest signal of over-transmitted PTVs, in line with previous reports[34]. The vast majority of inherited PTVs appear to affect genes that have yet to show signal from *de novo* variation, with only 1% residing in the strongest associated genes, indicating the inherited variants are hitting genes with somewhat weaker selective pressure against them. Ultimately, however, as these variants occur in 15.4% of cases but carry only 1.16-fold increased risk as a group, they explain little of the overall heritability (<1% of the variance in liability).

The pLI metric for constraint against loss-of-function variation, given the current size of ExAC and general scarcity of truncating variants, does not yet provide precise resolution of the selective coefficient acting on PTVs in that gene. That is, a pLI of ≥ 0.9 is consistent with coefficients as low as 0.1, in which the vast majority of variants will be inherited, as well as those which are so damaging as to be almost completely reproductively null. Further increases in the size of reference panels such as ExAC will allow us to refine these estimates, articulating a gradient from the strongest genes we currently flag (e.g., the 11 genes with ≥ 3 *de novo* PTVs in ASD and none in controls that will make their contribution almost entirely through penetrant *de novo* variation that will most often not survive beyond a single generation) to those genes we have yet to define clearly that will make their contribution largely through inherited, albeit less penetrant, variation. The significant expansion of exome sequencing in ASD, alongside the growth of reference panels from which to draw more precise inferences about selective pressure against variation in each gene, will allow us to begin to access these genes and fill in the genetic architecture of ASDs in the region of the effect size spectrum between severe *de novo* variation at one end and common variation at the other.

While we use ASDs and ID/DD here to explore this framework, it can certainly be applied toward any trait. However, this framework is optimally powered in traits governed by genes under strong selection, as it will remove *de novo* variants that are more common when examined in the context of a larger reference population (e.g., ExAC). Our results reinforce the obvious point that **not** all *de novo* variants identified in neurodevelopmental disease patients are rare and contribute to risk, while highlighting the tremendous value of large population sequence resources even in the interpretation of *de novo* variation. This 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 very likely the causal variant[35,36]. Clearly, not all *de novo* variants are equal, and not all *de novo* variants in a gene contribute to risk in the same way.

Online Methods

Datasets and data processing.

We used a combined set of 8401 published *de novo* variants from 3982 probands with ASD (2651 of which were administered an IQ test) and 2078 of their unaffected siblings from two recent large-scale exome sequencing studies[6,7] (**Supplementary Table 1**). We also used 1692 *de novo* variants from 1284 probands published in studies of intellectual disability (ID)[13,15] and

developmental delay (DD)[37] (**Supplementary Table 2**). Details of the sequencing and *de novo* calling can be found in the referenced publications. To ensure uniformity in variant representation and annotation across datasets and with respect to the ExAC reference database, we created a standardized variant representation through a Python implementation of vt normalize[38] and re-annotated all variants analyzed in this paper using Variant Effect Predictor (VEP)[39] version 81 with Gencode v19 on GRCh37. VEP provided the Ensembl Gene IDs, gene symbol, the Ensemble Transcript ID for use in determining canonical transcripts, as well as PolyPhen2 and SIFT scores. We used the annotation from the canonical transcript when possible for cases when the variant resided in multiple transcripts, and the most deleterious annotation in cases of multiple canonical transcripts. If no canonical transcript was available, the most deleterious annotation was used.

Determining whether a *de novo* variant is class 1 or class 2.
De novo variants were classified as class 1 or class 2 based on their respective absence or presence in ExAC. Presence or absence in ExAC was defined if the variant had the same chromosome, position, reference, and alternate allele in both files. For insertions/deletions, we took a conservative stance that they must match exactly (i.e., a subset was not sufficient; **Supplementary Note**). Lastly, for variants outside of proportion of the genome covered by ExAC, we considered them to be class 1 *de novo* variants – as expected, none of these variants reside in the coding region (**Supplementary Table 15**).

Variant calling for transmission analysis.
We used the Genome Analysis Toolkit (GATK v3.1-144) to call a dataset of 22,144 exomes from the Autism Sequencing Consortium (ASC) & Simons Simplex Collection (SSC)[40] sequencing efforts. This call set contained 4319 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 case-control dataset of individuals of Swedish ancestry (404 individuals with ASD and 3564 controls). 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 – where the alternative allele was seen in only one parent in the dataset – was transmitted at a rate of 50%, because we expect, as a class, synonymous variants to be transmitted 50% of the time (**Supplementary Note**; **Supplementary Fig. 2**). Lastly, we removed variants in which more than 20% of families failed one of our filters.

On the use of pLI (probability of loss-of-function intolerance).
Using the observed and expected number of PTVs per gene in the ExAC dataset, we developed a metric to evaluate a gene's apparent intolerance to such variation[25]. Briefly, the probability of loss-of-function intolerance (pLI) was computed using an EM algorithm that assigned genes to one of three categories: fully tolerant (in which PTVs are presumed neutral and occur, like synonymous variants, at rates proportional to the mutation rate), "recessive like" (showing PTV depletion similar to known severe autosomal recessive diseases) and "haploinsufficient" (showing PTV depletion similar to established severe haploinsufficiencies). pLI is the posterior probability that a gene resides in the last, most loss-of-function intolerant, category. See the supplement of Lek, et al., for more details.

Phenotype Analysis
Full-scale deviation IQs were available from 2232 individuals from the SSC[40]. Full-scale deviation IQ scores in the SSC were measured using several tests including the Differential Ability Scales, the Wechsler Intelligence Scale for Children, and the Wechsler Abbreviated Scale of Intelligence. IQ has previously been associated with *de novo* PTV rate in the SSC[7,16,41]. In this analysis, we used poisson regression to estimate the relationship between rate of each of class 1 and class 2 PTVs and proband full-scale deviation IQ in the SSC.

Acknowledgements
We would like to thank all of the members of the ATGU and the Wall Lab for their assistance in this endeavor. We would also like to greatly thank Andrea Byrnes, Alicia Martin, Kyle Slatterstrom, and David Fronk for their insightful contributions. Lastly, we would like to acknowledge G. A. Barnard, for inspiring us to write in a more conversational tone, similar to his seminal 1947 paper, *Significance Tests For 2x2 Tables*. ML and DGM's work on the ExAC project was funded by U54DK105566 and R01GM104371 from the National Institutes of Health. KS was funded by T32 HG002295/HG/NHGRI. EBR was funded by National Institutes of Mental Health Grant 1K01MH099286 and NARSAD Young Investigator grant 22379.

Author Contributions
JAK and EBR performed the analyses. JAK, DPH, EBR, MDJ designed the experiment. JAK and KS wrote the code. DPW, EBR, MJD supervised the research. JAK and MJD wrote the paper. JAK, KES, DPH, SJS, ML, KJK, DGM, and JDB generated data. JAK was responsible for the remainder. All authors revised and approved the final manuscript.

Competing Financial Interests
The authors declare no competing financial interests.

Citations

1. Developmental, D.M.N.S.Y. & Investigators, P. Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002)* **63**, 1 (2014).
2. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* **45**, 984-94 (2013).
3. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat Genet* **46**, 881-5 (2014).
4. Klei, L. *et al.* Common genetic variants, acting additively, are a major source of risk for autism. *Mol Autism* **3**, 9 (2012).
5. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427 (2014).
6. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-215 (2014).
7. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-221 (2014).
8. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-99 (2012).
9. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).
10. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-50 (2012).
11. Sanders, S.J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237-41 (2012).
12. Reichenberg, A. *et al.* Discontinuity in the genetic and environmental causes of the intellectual disability spectrum. *Proceedings of the National Academy of Sciences* **113**, 1098-1103 (2016).
13. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* **367**, 1921-9 (2012).
14. Deciphering Developmental Disorders, S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-8 (2015).
15. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674-82 (2012).
16. Robinson, E.B. *et al.* Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proc Natl Acad Sci U S A* **111**, 15161-5 (2014).
17. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).
18. Robinson, E.B. *et al.* Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat Genet* **advance online publication**(2016).
19. Bellus, G.A. *et al.* Achondroplasia is defined by recurrent G380R mutations of FGFR3. *Am J Hum Genet* **56**, 368-73 (1995).
20. Kimura, M. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893-903 (1969).
21. Coulondre, C., Miller, J.H., Farabaugh, P.J. & Gilbert, W. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274**, 775-780 (1978).

22. Haukka, J., Suvisaari, J. & Lonnqvist, J. Fertility of patients with schizophrenia, their siblings, and the general population: a cohort study from 1950 to 1959 in Finland. *Am J Psychiatry* **160**, 460-3 (2003).
23. Laursen, T.M. & Munk-Olsen, T. Reproductive patterns in psychotic patients. *Schizophr Res* **121**, 234-40 (2010).
24. Power, R.A., Kyaga, S., Uher, R. & et al. FEcundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* **70**, 22-30 (2013).
25. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv* (2015).
26. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-84 (2014).
27. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-3 (2013).
28. Francioli, L.C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **47**, 822-6 (2015).
29. Genome of the Netherlands, C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**, 818-25 (2014).
30. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-5 (2012).
31. Michaelson, J.J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431-42 (2012).
32. Przyborowski, J. & Wilenski, H. Homogeneity of Results in Testing Samples from Poisson Series: With an Application to Testing Clover Seed for Dodder. *Biometrika* **31**, 313-323 (1940).
33. Picoraro, J. & Chung, W. Delineation of New Disorders and Phenotypic Expansion of Known Disorders Through Whole Exome Sequencing. *Current Genetic Medicine Reports* **3**, 209-218 (2015).
34. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat Genet* **47**, 582-588 (2015).
35. Akawi, N. *et al.* Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nat Genet* **47**, 1363-1369 (2015).
36. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405-24 (2015).
37. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-8 (2015).
38. Tan, A., Abecasis, G.R. & Kang, H.M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202-4 (2015).
39. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70 (2010).
40. Fischbach, G.D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192-195 (2010).
41. Sanders, S.J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-33 (2015).

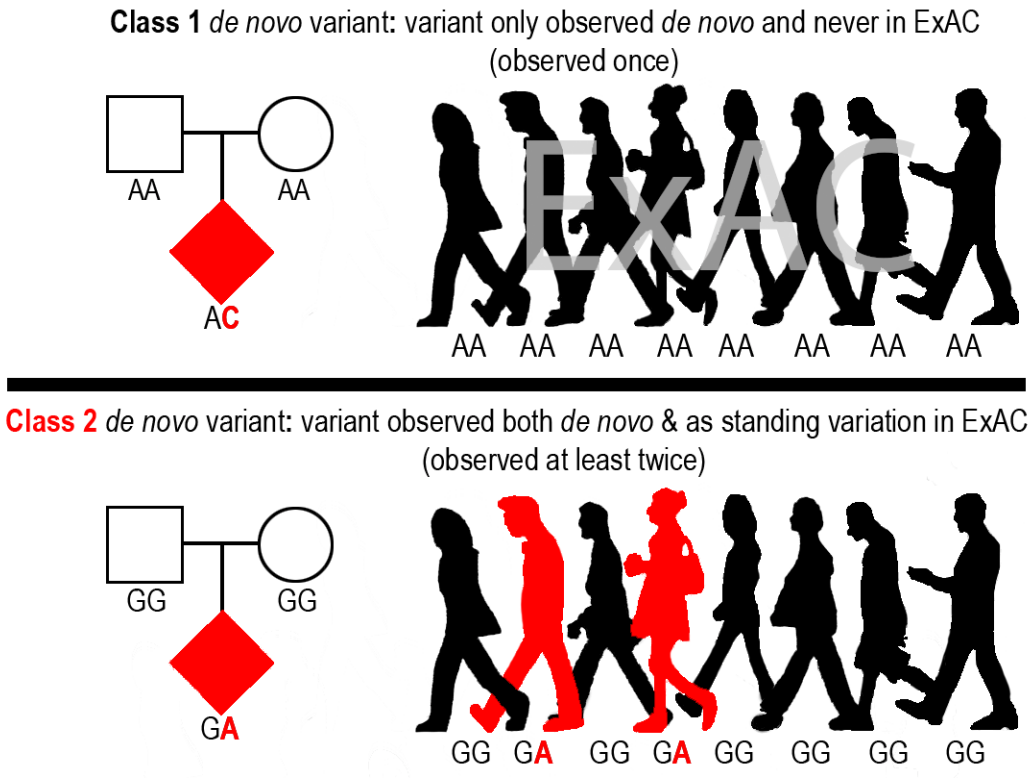

Figure 1 Illustration of class 1 and class 2 *de novo* variants with the genotypes of each variant for 8 of the 60,706 individuals in ExAC.

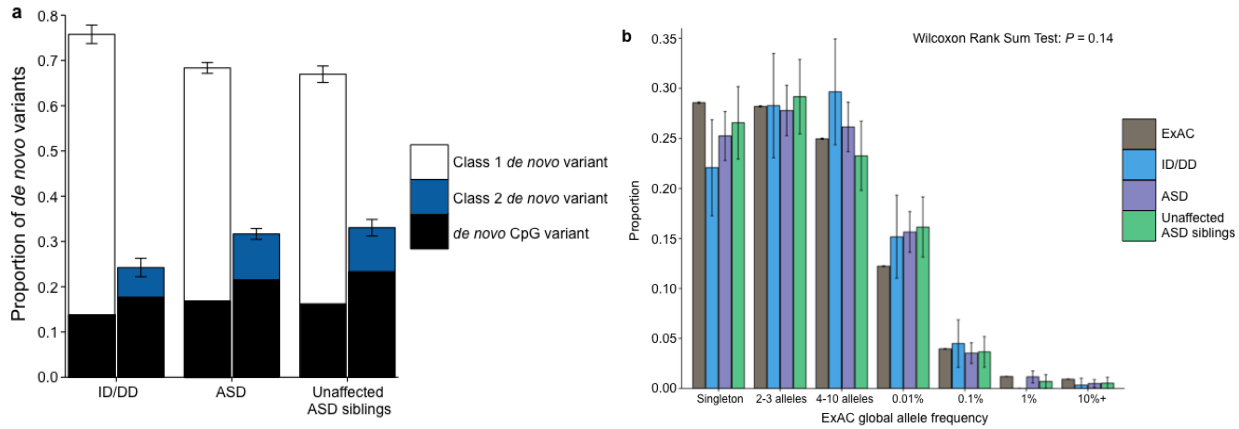

Figure 2 Properties of class 2 *de novo* variants. (**a**) The proportion of *de novo* variants across each cohort split between class 1 (left) and class 2 (right). Class 2 *de novo* variants are strongly enriched for CpG variants ($P < 10^{-20}$). (**b**) Allele frequency distribution of class 2 *de novo* CpGs by cohort with the matching distribution of CpG variants in ExAC for comparison. Allele frequency distributions do not significantly differ ($P = 0.14$; Wilcoxon rank sum test). Error bars represent 95% confidence intervals throughout **a** and **b**. ID/DD, intellectual disability / developmental delay; ASD, autism spectrum disorder

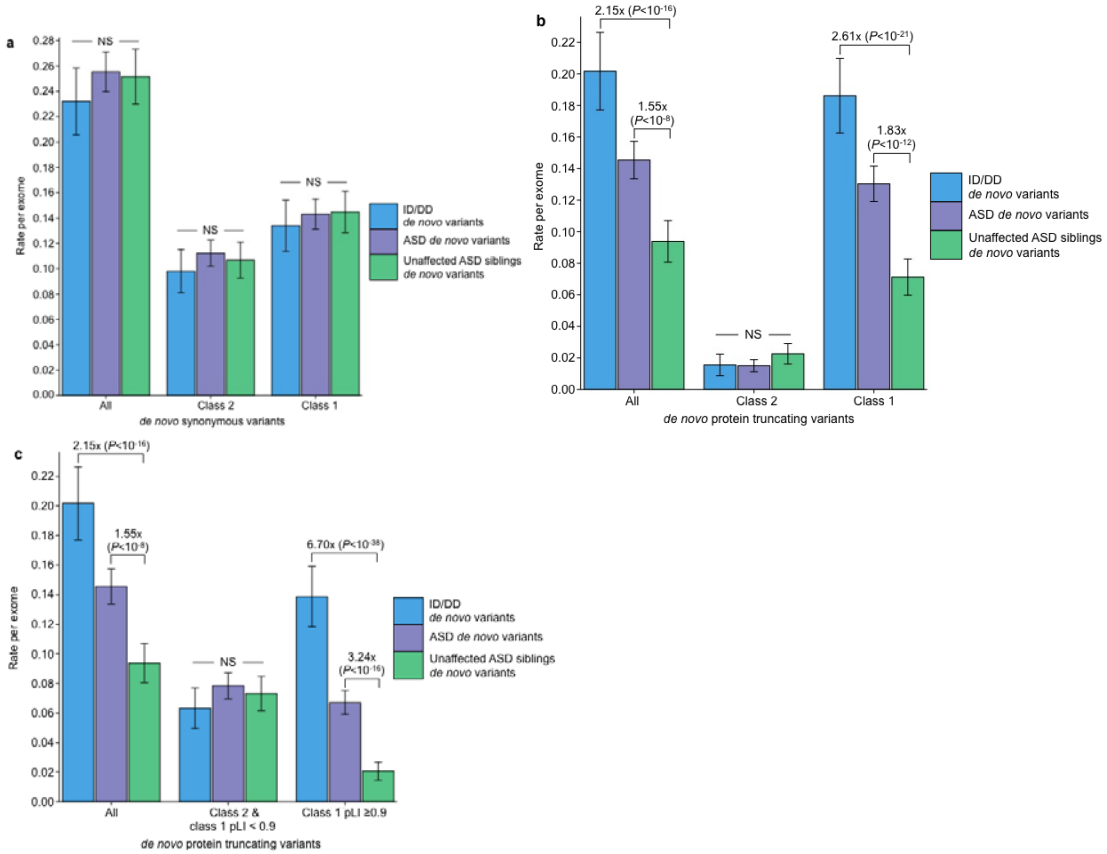

Figure 3 Partitioning the rate of *de novo* variants per exome by class 1, class 2, and pLI. Within each grouping, the rate is shown for ID/DD (left), ASD (middle), and unaffected ASD siblings (right). (**a**) Rate of *de novo* synonymous variants per exome partitioned into class 2 and class 1. No significant difference was observed for any grouping of *de novo* synonymous variants. (**b**) Rate of *de novo* PTVs per exome partitioned into class 2 (middle) and class 1 (right). Only class 1 *de novo* PTVs in ID/DD and ASD show association when compared to unaffected ASD siblings. (**c**) Rate of *de novo* PTVs partitioned into class 1 *de novo* PTVs in pLI ≥0.9 genes (right), and the remaining *de novo* PTVs (class 2 *de novo* PTVs and class 1 *de novo* PTVs in pLI <0.9 genes; middle). The entire observed association for *de novo* PTVs resides in class 1 *de novo* PTVs in pLI ≥0.9 genes. For all such analyses, the rate ratio and significance were calculated by comparing the rate for ID/DD and ASD to the rate in unaffected ASD siblings using the Poisson exact test[32] (**Supplemental Note**). Error bars represent 95% confidence intervals throughout (**a**)–(**c**). ID/DD, intellectual disability / developmental delay; ASD, autism spectrum disorder; PTV, protein truncating variant; pLI, probability of loss-of-function intolerance; NS, not significant

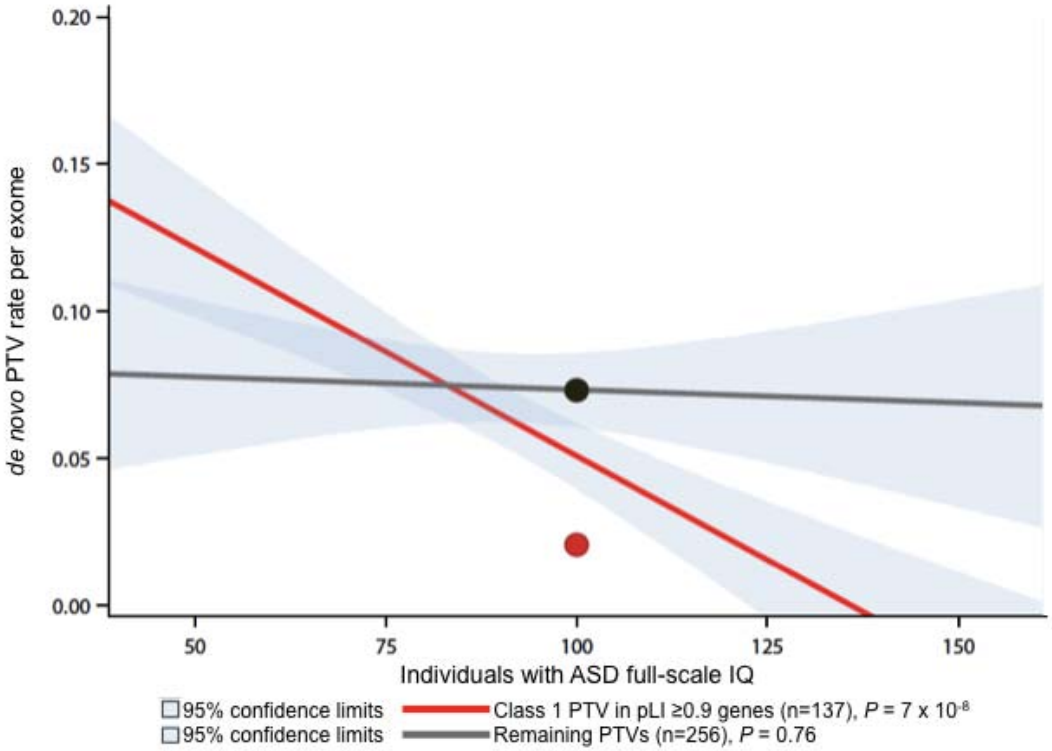

Figure 4 IQ distribution of class 1 *de novo* PTVs in pLI ≥0.9 genes (red) and remaining *de novo* PTVs (class 2 and class 1 pLI < 0.9) (grey) in 393 individuals with ASD with a measured full-scale IQ. Dots indicate the rate in unaffected ASD siblings for their respective categories of *de novo* PTVs. Significance calculated using Poisson regression. Only class 1 *de novo* PTVs show association with full-scale IQ. ASD, autism spectrum disorder; PTV, protein truncating variant; pLI, probability of loss-of-function intolerance

Table 1 Top 12 genes with ≥3 class 1 *de novo* PTVs in individuals with ASD

	Class 1 *de novo* PTVs			Inherited		Case-control		
Gene	ASD	Unaffected ASD siblings	ID/DD	T	U	Case	Control	pLI
CHD8	7	0	0	1	0	0	0	1
ARID1B	5	0	11	0	0	0	0	1
DYRK1A	5	0	2	0	0	0	0	0.9996
SYNGAP1	5	0	9	0	0	0	0	1
ADNP	4	0	4	2	0	1	0	0.9989
ANK2	4	0	0	1	1	0	0	1
DSCAM	4	0	0	2	0	1	0	1
SCN2A	4	0	7	0	0	1	0	1
ASH1L	3	0	0	0	0	0	0	1
CHD2	3	0	2	0	0	0	0	1
KDM5B	3	2	0	5	1	0	0	5.09E-05
POGZ	3	0	2	0	0	2	0	1

12 genes with ≥3 class *1 de novo* PTVs in 3982 individuals with ASD. Additionally, for each gene, we have listed the number of class 1 *de novo* PTVs in 2078 unaffected ASD siblings and in 1284 individuals with ID/DD, as well as the number of singleton PTVs absent from ExAC that were transmitted (T) or untransmitted (U) to 4319 individuals with ASD and present in 404 cases of ASD and 3654 population controls. ID/DD, intellectual disability / developmental delay; ASD, autism spectrum disorder; PTV, protein truncating variant; pLI, probability of loss-of-function intolerance