Transcript expression-aware annotation improves rare variant discovery and interpretation

Beryl B. Cummings1,2,3, Konrad J. Karczewski1,2, Jack A. Kosmicki1,2,4, Eleanor G. Seaby1,2, Nicholas A. Watts1,2, Moriel Singer-Berk2, Jonathan M. Mudge6, Juha Karjalainen1,2,5, F. Kyle Satterstron1,2,5, Anne O’Donnell-Luria1,2,7, Timothy Poterba1,2,5, Cotton Seed1,2,5, Matthew Solomonson1,2, The Genome Aggregation Database Production Team, The Genome Aggregation Database Consortium, Mark J. Daly1,2,5, Daniel G. MacArthur1,2

Correspondence: danmac@broadinstitute.org

Author Affiliations
1. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA.
2. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
3. Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA.
4. Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA 02115, USA
5. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
6. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB101SD, UK
7. Division of Genetics and Genomics, Boston Children’s Hospital, Boston, MA 02115, USA

Abstract
The acceleration of DNA sequencing in patients and population samples has resulted in unprecedented catalogues of human genetic variation, but the interpretation of rare genetic variants discovered using such technologies remains extremely challenging. A striking example of this challenge is the existence of disruptive variants in dosage-sensitive disease genes, even in apparently healthy individuals. Through manual curation of putative loss of function (pLoF) variants in haploinsufficient disease genes in the Genome Aggregation Database (gnomAD)1, we show that one explanation for this paradox involves alternative mRNA splicing, which allows exons of a gene to be expressed at varying levels across cell types. Currently, no existing annotation tool systematically incorporates this exon expression information into variant interpretation. Here, we develop a transcript-level annotation metric, the proportion expressed across transcripts (pext), which summarizes isoform quantifications for variants. We calculate this metric using 11,706 tissue samples from the Genotype Tissue Expression project (GTEx) and show that it clearly differentiates between weakly and highly evolutionarily conserved exons, a proxy for functional importance. We demonstrate that expression-based annotation selectively filters 22.8% of falsely annotated pLoF variants found in haploinsufficient disease genes in gnomAD, while removing less than 4% of high-confidence pathogenic variants in the same genes. Finally, we apply our expression filter to the analysis of de novo variants in patients with autism spectrum disorder (ASD) and developmental disorders and intellectual disability (DD/ID) to show that pLoF variants in weakly expressed regions have effect sizes similar to those of synonymous variants, while pLoF variants in highly expressed exons are most strongly enriched among cases versus controls. Our annotation is fast, flexible, and generalizable, making it possible for any variant file to be annotated with any isoform expression dataset, and will be valuable for rare disease diagnosis, rare variant burden analyses in complex disorders, and curation and prioritization of variants in recall-by-genotype studies.
A primary challenge in the use of genome and exome sequencing to predict human phenotypes is that our capacity to identify genetic variation exceeds our ability to interpret their functional impact (3, 4). One underappreciated source of variability for variant interpretation involves differences in alternative mRNA splicing, which enables exons to be expressed at different levels across tissues. These expression differences mean that variants in different regions of a gene can have different phenotypic outcomes depending on the isoforms they affect. For example, variants occurring in an exon differentially included in two isoforms of CACNA1C with diverse tissue expression patterns result in distinct types of Timothy syndrome (5). Pathogenic variants in the isoform that exhibits multi-tissue expression result in a multi-system disorder (5-7), whereas those on the isoform predominantly expressed in heart result in more severe and specific cardiac defects (8). In addition, Mendelian variants have been found on tissue-specific isoforms (9, 10) and isoform expression levels in TTN have been used to show that pLoF variants found in healthy controls occur in exons that are absent from dominantly expressed isoforms, whereas those in dilated cardiomyopathy patients occur on constitutive exons (11), emphasizing the utility of exon expression information for variant interpretation.

We find that isoform diversity is also a contributor to the paradoxical finding of disruptive variants in dosage-sensitive disease genes in ostensibly healthy individuals. In the gnomAD database, we identify 401 high-quality pLoF variants that pass both sequencing and annotation quality filters in 61 haploinsufficient disease genes where heterozygous pLoF variants are established to cause severe developmental delay phenotypes with high penetrance (Methods). Given the severity of these phenotypes and their extremely low worldwide prevalence, ranging from 1 in 10,000 to less than 1 in a million, very few, if any true pLoF variants would be expected to be found in the gnomAD population. As such, most or all of these observed pLoF variants are likely to be sequencing or annotation errors (12). Manual curation of these variants reveals common error modes that result in likely misannotation of pLoFs, with diversity of transcript structure, mediated by variants falling on low-confidence transcripts, emerging as a major consideration (Figure 1, Supplementary Figure 1, Supplementary Tables 1-3). However, no existing tools systematically incorporate information on transcript expression into variant interpretation.

![Figure 1: Curation of pLoF variants in haploinsufficient disease genes found in gnomAD reveals transcript errors as a major confounding error mode in variant annotation](image-url)

We identified and manually curated 401 pLoF variants in the gnomAD dataset in 61 haploinsufficient severe developmental delay genes and flagged any reason the pLoF may not be a true LoF variant. Top plot shows the frequency of each error mode present in the 306 variants classified as unlikely to be a true LoF. Transcript errors emerge as a major putative error mode in the annotation of these pLoF variants. Beeswarm plot on bottom shows the average pext score across GTEx tissues presented in the manuscript for each variant in the error categories. This shows that pext values are discriminatedally lower for variants that are annotated as possible transcript errors.

CC-BY 4.0 International license
The advent of large-scale transcriptome sequencing datasets, such as GTEx (2), provides an opportunity to incorporate cross-tissue exon expression into variant interpretation. However, the current formats of these databases do not readily allow for unbiased estimation of exon expression. The GTEx web browser offers information on exon-level read pileup across tissues, but this approach is confounded by technical artifacts such as 3' bias (13) (preferential coverage of bases close to the 3' end of a transcript; Supplementary Figure 2A). Such systematic biases mean that simple exon-level coverage in a transcriptome dataset cannot be used as a reliable proxy for exon expression, especially in longer genes (Figure 2A, Supplementary Figure 2B).

Isoform quantification tools provide estimates of isoform expression levels that correct, albeit imperfectly (13, 14), for confounding by 3' bias as well as other technical artifacts such as isoform length, isoform GC content, and transcript sequence complexity (15-17). Here, we utilize isoform-level quantifications from 11,706 tissue samples from the GTEx v7 dataset to derive an annotation-specific expression metric. For each tissue, we annotate each variant with the expression of every possible consequence across all transcripts, which can be used to summarize expression in any combination of tissues of interest. We first compute the median transcript expression across samples per tissue, and define the expression of a given variant as the sum of the expression of all transcripts for which the variant has the same annotation (Figure 2A, Supplementary Figure 3A).

Figure 2: Summary of transcript-expression based annotation method A. Overview of transcript aware annotation. Most genes have many annotated isoforms, which can have varying expression patterns across tissues. Utilizing number of reads aligning to exonic regions in transcriptome datasets as a proxy for exon expression (top panel black) has confounding effects due to 3' bias. In this example, while exon 3 and 8 may have different expression levels in Brain – Cortex, the number of reads aligning to the two exons are similar, masking exon usage differences. Transcript-aware annotation defines the expression of every variant as the sum of transcripts that have the same annotation. The resulting transcript-level expression plots do not exhibit 3' bias, and reveal exon usage differences across tissues. B. Example of utility of transcript-expression based annotation. There are 20 high quality pLoF variants in the
haploinsufficient developmental delay gene \textit{TCF4} in gnomAD, annotated as dashed lines and arrows. All 20 variants have no evidence of expression in the GTEx dataset, suggesting functional TCF4 protein can be made in the presence of these variants.

The pext metric allows for quick visualization of the expression of exons across a gene. Figure 2B shows \textit{TCF4}, a haploinsufficient gene in which heterozygous variants result in Pitt-Hopkins syndrome \cite{18}, a highly penetrant disorder associated with severe developmental delay. This gene harbors 20 unique high quality pLoF mutations across 56 individuals in the gnomAD database. All 20 variants lie on exons with no evidence of expression across the GTEx dataset (Figure 2B, Supplementary Figure 4) indicating that functional TCF4 protein can be made in the presence of these variants. This visualization is now available for all genes in the gnomAD browser (gnomad.broadinstitute.org), and can aid in rapid identification of variants occurring on exons with little to no evidence of expression in GTEx.

To explore whether expression-based annotation marks functionally important regions, we compared the distribution of the pext metric in evolutionarily conserved and unconserved regions using phyloCSF \cite{19}. Exons with patterns of multi-species conservation consistent with coding regions have higher phyloCSF scores, and should exhibit detectable expression patterns, whereas regions with lower scores will be enriched for incorrect exon annotations, which are expected to have little evidence of expression in a population transcriptome dataset. As expected, we observe significantly lower expression for unconserved regions, and near-constitutive expression in highly conserved regions (Figure 3A, Supplementary Figure 5A). This difference remains statistically significant after correcting for exon length (logistic regression $p < 1.0 \times 10^{-100}$), which can influence both phyloCSF scores and isoform quantifications, indicating that transcript expression-aware annotation marks functionally relevant exonic regions.

While the metrics are associated, we find that pext provides orthogonal information to conservation for variant interpretation. For example, regions with low evidence of conservation but high expression (in Figure 3A) are enriched for genes in immune-related pathways (Methods), which are selected for diversity but represent true coding regions. In addition, the pext value is higher for pLoF variants annotated as high confidence (HC) by the Loss of Function Transcript Effect Estimator\cite{1} (LOFTEE) with no additional flags than those flagged as having found on unlikely open reading frames or weakly conserved regions (Figure 3B, Supplementary Figure 5B). However, LOFTEE-HC variants with no flags can also have low pext values, suggesting transcript-expression aware annotation adds additional information to the currently available interpretation toolkit.

We undertook manual evaluation of 128 regions marked as unexpressed (mean pext $< 0.1$ in all tissues and in GTEx brain) in 61 haploinsufficient genes following the GENCODE manual annotation workflow\cite{20} to evaluate the annotation quality in these coding sequence (CDS) regions. A third of flagged regions were associated with low quality models that have been removed or switched to non-coding biotypes in subsequent GENCODE releases (Supplementary Figure 6) while 70% of the remaining regions correspond to models that satisfy only minimum criteria for inclusion in the gene set, corresponding to ‘putative’ annotations that lack markers for CDS functionality (Supplementary Table 4). Nonetheless, we find support for some highly conserved CDS regions, several of which show evidence of transcription in fetal tissues, underlining the importance of incorporating multiple isoform expression datasets for interpretation (Supplementary Figure 6D).

Nonsynonymous variants found on constitutively expressed regions would be expected to be more deleterious than those on regions with no evidence of expression. To test this, we defined expression bins based on the average pext value across GTEx tissues where an average pext value less than 0.1 was defined as low (or unexpressed), above 0.9 as high (or near-constitutive) and intermediate values as medium expression. We compared the mutability-adjusted proportion singleton (MAPS), a measure of negative selection on variant classes\cite{21}, partitioned on the LoF Observed Upper-bound Fraction (LOEUF) decile, a measure of constraint against pLoF variants in the gnomAD dataset\cite{1} in each of these expression bins. MAPS scores differed substantially between pLoF variants found on low-expressed and high-expressed regions in genes intolerant to pLoF variation (Figure 3C, Supplementary Figure 5C). This information is complementary to existing variant prioritization tools such as PolyPhen-2\cite{22} (Supplementary Figure 5D). This skew of nonsynonymous variation in high-expressed regions suggests that variation arising in such exons tends be more deleterious, whereas nonsynonymous variants on regions with low expression are similar to missense variants in their inferred deleteriousness.
We define highly conserved and unconserved regions and compared the expression status of these regions across GTEx. Highly conserved regions are enriched for having near-constitutive expression whereas unconserved regions are enriched for having little to no usage across GTEx. This difference is significant after correcting for gene length (logistic regression p value < 1 x 10^{-100}). We note that unconserved regions with high levels of expression (pext > 0.9) are enriched for immune-related genes, which are selected for diversity and thus have low conservation, but represent true coding regions.

B. Transcript-expression based annotation recapitulates, and adds information to, existing interpretation tools. LOFTEE-HC pLoF variants in gnomAD with no flags are enriched for higher pext values, whereas HC pLoF variants falling on low phyloCSF or unlikely ORF regions are enriched for low expression. However, HC-pLoF variants can also be filtered based on a low pext score. Red dots represent median pext value across GTEx tissues.

C. Nonsynonymous variants found on near-constitutive regions tend to be more deleterious. We compared the mutability adjusted proportion singleton (MAPS) score for variants with low (<0.1), medium (0.1 ≤ pext ≤ 0.9) and high (pext > 0.9) expression. Variants with near-constitutive expression have a higher MAPS score, indicating higher deleteriousness than those with little to no evidence of expression. Dashed grey and orange line represent MAPS values for all gnomAD missense and all synonymous variants, respectively.

To evaluate the utility of transcript expression-based annotation in Mendelian variant interpretation, we assessed the number of variants that would be filtered based on a pext cutoff of <0.1 (low expression) across GTEx tissues for three gene sets. Firstly, we evaluated high-quality pLoF variants in the 61 manually curated haploinsufficient genes in gnomAD and ClinVar(23). The low pext expression bin resulted in filtering of 22.8% of pLoF variants in haploinsufficient developmental delay genes in gnomAD, but only 3.8% of high-quality pathogenic variants in ClinVar (Figure 4A; p = 4.7 x 10^{-35} Methods). We next compared pLoF variants in autosomal recessive disease genes found in a homozygous state in at least one individual in gnomAD and any pLoF variant in these genes in ClinVar and observed similar results: expression-based annotation filters 30.0% of variants in gnomAD while only filtering 3.2% of variants in ClinVar (Figure 4B; p = 3.5 x 10^{-61}).

Finally, we evaluated gnomAD pLoF variants in genes that are constrained against pLoF variation(1) (LOEUF score < 0.35). Given that these genes are depleted for loss-of-function variation in the general population, we expect the observed pLoF variants in these genes to be enriched for annotation errors. We compared the proportion filtered to synonymous variants in the same genes, which we expect to be randomly distributed. Our metric removes 16.8% of pLoF variants in constrained genes, but only 5.2% of synonymous variants (Figure 4C; p < 1.0 x 10^{-100}). In all cases, the vast majority of filtered variants were otherwise high-confidence with no LOFTEE annotation flags, suggesting again that pext provided additional information to existing variant prioritization tools in removing annotation errors (Supplementary Figure 7).
Figure 4: Transcript-expression based annotation aids Mendelian variant interpretation. A. Comparison of proportion of high quality pLoF variants filtered in a curated list of 61 haploinsufficient developmental delays genes in gnomAD vs ClinVar with a cutoff of average pext across GTEx ≤ 0.1 (low expression). Expression-based filtering results in removal of 22.8% of gnomAD pLoFs and 3.8% of confidently curated set of pLoFs in ClinVar. B. Expression-based annotation filters 30% of pLoF variants found in gnomAD in a homozygous state in at least one individual, and 3.2% of any pLoF variants found in the same genes in ClinVar. C. We extended this filtering approach to pLoF and synonymous variants in gnomAD pLoF-intolerant genes (defined by LOEF < 0.35). This filters 16.8% LoF and 5.2% of synonymous variants. Numbers below bar plots indicate the total number of high-quality variants considered in each group. For pLoFs only LOFTEE-HC variants were considered, p-values calculated from fisher’s exact test for counts.

To explore the benefits of this approach for rare variant analyses, we applied pext binning to burden testing of de novo variants in patients with developmental delay/intellectual disability (DD/ID) or autism spectrum disorder (ASD) using a set of 23,970 de novo variants collated from several studies including the Deciphering Developmental Disorders (DDD) project and the Autism Sequencing Consortium (ASC)(24-29). We find that de novo pLoF variants in patients with DD/ID in low-expressed regions have effect sizes similar to those of synonymous variants (rate ratio, denoted as RR, of low-expressed pLoFs = 1.08, p = 0.90) whereas pLoF variants in highly expressed regions have much larger effect sizes (RR = 4.64, p = 3.74 x 10^{-38}; Figure 5A). This observation is consistent for de novo variants in autism (RR for low-expressed pLoFs = 0.80, p = 0.47; RR for high-expressed pLoFs = 2.11, p = 8.2 x 10^{-8}, Figure 5B) and congenital heart disease with co-morbid neurodevelopmental delay (Supplementary Figure 8A) as well as rare variants (AC ≤ 10) identified in highly constrained genes in the large iPSYCH case/control study of Danish patients with autism spectrum disorder and attention-deficit/hyperactivity disorder (Supplementary Figure 8B). Overall, we consistently observe low-expressed pLoFs to have effect sizes similar to those of synonymous variants, with pLoF variants in constitutive regions having larger effect sizes, suggesting that incorporating transcript expression-aware annotation in rare variant studies can boost power for gene discovery.

Figure 5: Application of transcript-expression based annotation to de novo variant analyses in A. developmental delay and/or intellectual disability (DD/ID) and B. autism spectrum disorder (ASD). We find that de novo pLoF variants found on near constitutively expressed regions in GTEx brain tissues have larger effect sizes than de novo LoF variants in weakly expressed regions in both disorders. Strikingly, de novo pLoF variants found on regions with little evidence for expression are equally distributed in cases vs controls as de novo synonymous variants, suggesting such variants can be removed from gene burden testing analyses to boost discovery power. The high pext expression bin contains 46.1% and 42.3%, and the low expression bin contains 4.0% and 6.0% of pLoF variants found in DD/ID and ASD patients, respectively. Rate ratio represents estimate from the poisson exact test.
We have described the development and validation of a transcript expression-based annotation framework to integrate results from transcriptome sequencing experiments into clinical variant interpretation. While our initial analysis utilizes GTEx, our method can be used with any isoform expression dataset to annotate any variant file rapidly in the scalable software framework Hail (https://hail.is). For example, annotation of >120,000 gnomAD individuals with GTEx takes under an hour using 60 cores, at a cost of about $5 on public cloud compute, which can be further scaled to larger datasets. In addition, the annotations we provide are flexible: while we have described the use of average transcript-level expression across many tissues, alternative approaches such as using maximum expression across any tissue, may prove useful depending on variant interpretation goals (Supplementary Figures 9 & 10).

We note that while this metric successfully discriminates between near-constitutive and low expression levels, which are useful for prioritizing and filtering variants, respectively, regions with intermediate expression levels are more challenging to interpret. However, we hypothesize directed analyses of intermediate expression levels may help elucidate the role of alternative splicing in phenotypic diversity (30, 31). In addition, while we have binned average pext scores across GTEx tissues into low, medium and high expression, different genes will likely have varying optimal tissues and thresholds for variant interpretation. Regions tagged as low expression are often corroborated by expert opinion of CDS curation, but domain knowledge of a gene will outperform this summary metric.

An important caveat in our approach is the imprecision of isoform quantification methods using short-read transcriptome data. However, we note that repeating key analyses in the manuscript with a different isoform quantification tool showed consistent results (Methods, Supplementary Figure 11), suggesting robustness to the precise pipeline used. The utility of this framework will increase as our ability to quantify isoform expression across tissues improves, including refinement of methods and gene models, as well as availability of long-read RNA-seq data from human tissues. In addition, improvement of single-cell RNA-seq technologies and generation of data across human tissues will provide insight into cell type-specific exon usage for incorporation into variant interpretation.

The code used to generate pext is available as open source software (https://github.com/macarthur-lab/tx_annotation). In addition, we provide a precomputed file of the transcript expression value for every possible single nucleotide variant in the human genome. This metric has already proven useful in variant curation for drug target identification (32) and for filtering variants for identification of human knockouts(1). Overall, our metric can be incorporated into variant interpretation in Mendelian disease pipelines, rare variant burden analyses, and the prioritization of variants for recall-by-genotype studies.

Methods

Data and code availability
We utilized the gnomAD v.2.1.1 sites Hail 0.2 (https://hail.is) table which is accessible publicly at gs://gnomad-public/release/2.1.1 and at https://gnomad.broadinstitute.org. The GTEx v7 gene and isoform expression data were downloaded from the GTEx portal (gtexportal.org). The GTEx pipeline for isoform quantification is available publicly (https://github.com/broadinstitute/gtex-pipeline/) and briefly involves 2-pass alignment with STAR v2.4.2a(33), gene expression quantification with RNA-SeQC v1.1.8(34), and isoform quantification with RSEM v1.2.22. The LOEUF constraint file was downloaded from gs://gnomad-resources/lof_paper/. Variants used in all gnomAD analyses in the manuscript passed random forest filtering, and all pLoF variants were annotated as high confidence (HC) by LOFTEE v.1.0, which is described in an accompanying manuscript(1). All files used in the analyses in the manuscript are available in gs://gnomad-public/papers/2019-tx-annotation/. Scripts to QC the gnomAD dataset are available at https://github.com/macarthur-lab/gnomad_qc and the scripts to generate files for the analyses are available at https://github.com/macarthur-lab/tx_annotation.

Curation of pLoF variants in haploinsufficient developmental disease genes
For identification of haploinsufficient developmental delay genes, we selected genes curated by the ClinGen Dosage Sensitivity Working Group(35); 58 of the 61 genes had a score of 3 with sufficient evidence for pathogenicity, while two genes (CHAMP1, CTCF) had a score of 2 (some evidence) and one gene (HERE) was not yet scored. The penetrance of pathogenic variants in each gene was reviewed in the literature, and only genes with >75% reported penetrance were included. These conditions are those too severe to expect to see an individual in gnomAD (likely unable to consent for a study without guardianship). The 61 genes include...
We extracted pLoF variants, defined as essential splice acceptor, essential splice donor, stop gained, and frameshift variants, identified in the 61 haploinsufficient disease genes from the gnomAD v2.1.1 exome and genome sites tables, and considered only those pLoF variants that passed random forest filtering in the gnomAD dataset, and were annotated with high confidence (HC) by LOFTEE v1.0. Of 61 genes, 55 had at least one high quality pLoF available in gnomAD. We performed manual curation of 401 pLoF variants using a web-based curation portal to identify any reason a pLoF may have been a variant calling or annotation error, and categorized the likelihood of each variant being a true LoF.

Technical errors comprised genotyping errors, strand biases, reference errors, and repetitive regions that could be detected by visual inspection of reads in the Integrative Genomics Viewer(36) (IGV) and from the UCSC genome browser(37). Genotyping errors comprised skewed allele balances (conservative cutoff of \(\leq 35\%\)), low complexity sequences, GC rich regions, homopolymer tracts (\(\geq 6\) base pairs or \(\geq 6\) trinucleotide repeats) and low quality metrics (genotype quality, or GQ, \(< 20\)). Strand bias was flagged when a variant was skewed preferentially on the forward or reverse strand, or when the majority (\(>90\%\)) of a given strand covered a region; this was often observed around intron/exon boundaries. Strand biases despite balanced coverage of the forward and reverse strands were weighted towards likely not LoF, whereas a strand bias due to skewed strand coverage were weighted alongside other genotyping errors. Reference errors were uncommon, but identified by a small deletion in a given exon, posing as a \(<5\) base pair intron. Most genotyping errors and strand biases in isolation were not deemed critical in deciding whether a variant was likely not LoF or not LoF, with the exception of allele balance \(\leq 25\%\). Mapping errors were often identified by an enrichment of complex variation surrounding a variant of interest. Furthermore, the UCSC browser was used to highlight mapping discrepancies, such as self-chain alignments, segmental duplications, simple tandem repeats, and microsatellite regions.

In-frame multi-nucleotide variants (MNVs), essential splice site rescue, and frame-restoring insertion-deletions are rescue events that are predicted to restore gene function. MNVs were visualized in IGV and cross checked with codons from the UCSC browser; in frame MNVs that rescued stop codons were scored as not LoF. Essential splice site rescue occurs when an in frame alternative donor or acceptor site is present, which likely has a minimal effect on the transcript. Thirty-six base pairs upstream and downstream of the splice variant were assessed for splice site rescue. Cryptic splice sites within 6 base pairs of the splice variant were considered a complete rescue, rendering the variant not LoF. Rescue sites > 6 base pairs away but within +/- 20 base pairs were weighted with less confidence, scoring as likely not LoF. All potential splice site rescues were validated using Alamut v.2.11 (https://www.interactive-biosoftware.com/alamut-visual/). Frame-restoring indels were identified by scanning approximately +/- 80 base pairs from the annotated indel and counting any insertions/deletions to assess if the frame would be restored.
and/or partial exon conservation. Re-initiation events were flagged when a methionine downstream of the variant in the first coding exon was predicted to restart transcription, and were predicted to be likely not LoF. Variants occurring after a stop codon in the last coding exon were considered not LoF, particularly across the region of the exon or transcript in question. Error categories were grouped for Figure 1 as follows: Minority of transcripts and weak exon conservation were grouped as transcript errors, genotyping errors and homopolymers as sequencing errors, essential splice rescue and MNV grouped as rescue and strand bias was included in other annotation errors.

The criteria above were strictly adhered throughout and manual curation was performed by two independent reviewers to ensure maximum consistency and minimize human error. Any discordance in curation was re-curated by both curators together and resolved. Full results of manual curation are available in Supplementary Table 3.

Calculation of transcript-expression aware annotation

We first imported the GTEx v7 isoform quantifications into Hail and calculated the median expression of every transcript per tissue. This precomputed summary isoform expression matrix is available for GTEx v7 in gs://gnomad-public/papers/2019-tx-annotation/data/GRCH37_hg19/. We also import and annotate a variant file with the Variant Effect Predictor (VEP) version 85 against Gencode v19, implemented in Hail with the LOFTEE v1.0 plugin.

We use the transcript consequences VEP field to calculate the sum of isoform expression for variant annotations, i.e. the annotation-level expression across transcripts (ext). For variants that have multiple consequences for one transcript (for example, a SNV that is both a missense and a splice region variant on one transcript) we use the worst consequence, ordered by VEP (in this example, missense takes precedence over splice region). We filter the consequences to those only occurring on protein coding transcripts. Full ordering of the VEP consequences is available at: useast.ensembl.org/info/genome/variation/prediction/predicted_data.html

We then sum the expression of every transcript per variant, for every combination of consequence, LOFTEE filter, and LOFTEE flag for every tissue (Supplementary Figure 3A). For example, if a SNV is synonymous on ENST1, a LOFTEE HC stop-gained on ENST3 and ENST4, and LOFTEE low-confidence (LC) stop gained variant on ENST 5 and ENST6, the ext values will be synonymous: ENST1, stop-gained HC: ENST 3 + ENST4, and stop-gained LC: ENST5 + ENST6 per tissue. This can be computed with the tx_annotate() function by setting the tx_annotation_type to “expression”. We foresee the non-normalized ext values to be useful when only considering one tissue of interest.

To allow for taking average expression values across tissues of interest, we normalize the expression value for a given value to the total expression of the gene on which the variant is found. This is carried out by dividing the ext value with the sum of the expression of all transcripts per tissue in transcripts-per-million (TPM) v1.1.8(34) (Supplementary Figure 3B). The resulting pext (proportion expression across transcript) value can be interpreted as the proportion of the total transcriptional output from a gene that would be affected by the given variant annotation in question. If the gene expression value (and thus the denominator) in a given tissue is 0, the pext value will not be available for that tissue.

We note that for a minority of genes, when RSEM assigns higher relative expression to non-coding transcripts, the sum of the value of coding transcripts can be much smaller than the gene expression value for the transcript, resulting in low pext scores for all coding variants in the gene, and thus resulting in possible filtering of all variants for a given gene. In many cases this appears to be the result of spurious non-coding transcripts with a high degree of exon overlap with true coding transcripts. To prevent this artifact from affecting our analyses, we first calculated the maximum pext score for all variants across all protein coding genes, and removed any gene where the maximum pext score was below 0.2. This resulted in the filtering of 668 genes, representing 3.3% of all genes analyzed. We note that there is no overlap with the 668 genes and the haploinsufficient gene list, 97 of the filtered genes are present in OMIM (representing 1.5% of the OMIM gene list) and 42 filtered genes are considered constrained (representing 1.4% of LOEUF < 0.35, or constrained, genes) thus having low impact on variant interpretation in the context of disease associations.

When taking averages across tissues, such unavailable pext values are not considered (ie. when taking the mean across tissues, we remove NAs). This value can be computed with the tx_annotate() function by setting
the tx_annotation_type to "proportion". For the analyses in this manuscript, we remove reproduction-associated GTEx tissues (endocervix, ectocervix, fallopian tube, prostate, uterus, ovary, testes, vagina), cell lines (transformed fibroblasts, transformed lymphocytes) and any tissue with less than one hundred samples (bladder, brain Cervicalc-1 spinal cord, brain substantia nigra, kidney cortex, minor salivary gland) resulting in the use of 38 GTEx tissues.

The full transcript-expression aware annotation pipeline, implemented in Hail 0.2, is fully available at https://github.com/macarthur-lab/tx_annotation with commands laid out for analyses in the manuscript. Passing a Hail table through the tx_annotate() function returns the same table with a new field entitled “tx_annotation” which provides either the ext or pext value per variant-annotation pair, depending on parameter choice. We provide a helper function to extract the worst consequence and the associated expression values for these annotations. All analyses in the manuscript are based on the worst consequence of variant, ordered by VEP(39).

Functional validation of transcript-expression aware annotation
Conservation analysis was performed using phyloCSF scores using the same file utilized for the LOFTEE plugin, available publicly in gs://gnomad-public/papers/2019-tx-annotation/data/other_data/phylocsf_data.tsv.bgz. We denoted exons with a phyloCSF max open reading frame (ORF) score > 1000 as highly conserved and those with phyloCSF max ORF score < -100 as lowly conserved (Supplementary Figure 5A) and evaluated their average usage in GTEx.

Using the base-level pext values that are used in the gnomAD browser, we filtered to intervals with high or low conservation, and calculated the average pext value in the interval. To evaluate regions with low conservation but high expression, we identified genes harboring unconserved regions with the pext value > 0.9 for pathway enrichment analysis and used the web browser for FUMA GENE2FUNC feature(40), which incorporates Reactome(41), KEGG(42), Gene Ontology(43) (GO) as well as other ontologies. Default parameters were used for FUMA, with all protein coding genes as the background list. Results from FUMA pathway analysis are available in Supplementary Figure 12, and full results are available in Supplementary Table 5.

Analysis of pext values for LOFTEE flags and the MAPS calculation were performed utilizing the gnomAD v2.1.1 exome dataset. Calculation of MAPS scores was previously described in Lek et al. 2016(21) and is implemented as a Hail module, as described in Karczewski et al. 2019(1). MAPS is a relative metric, and cannot be compared across datasets, but is a useful summary metric for the frequency spectrum, indicating deleteriousness as inferred from rarity of variation (high values of MAPS correspond to lower frequency, suggesting the action of negative selection at more deleterious sites). The MAPS scores were calculated on the gnomAD v.2.1.1 dataset partitioning upon the LOEUF score and expression bin. The script for generating MAPs scores is available in the tx-annotation Github repository under /analyses/maps/maps_submit_per_class.py

Manual evaluation of unexpressed regions in haploinsufficient developmental delay genes using the GENCODE workflow
As an orthogonal evaluation of regions flagged as unexpressed with the pext metric, we identified any region in 61 haploinsufficient disease genes with a mean pext value < 0.1 in all GTEx tissues and in GTEx brain samples, due to the relevance of brain tissues for these disorders, regardless of mutational burden in gnomAD. The resulting list of 128 regions was evaluated by the HAVANA manual annotation group of the GENCODE project(20).

The manual evaluation first established whether the transcript model corresponding to the region in question was correct in terms of structure, comparing exon / intron combinations, and the accuracy of splice sites against the RNA evidence supporting the model. Second, the functional biotype of each model was reassessed; in particular, whether the decision to annotate the model as protein-coding in GENCODE v19 was appropriate. Note that GENCODE models that incorporate alternative exons or exon combinations in comparison to the ‘canonical’ isoform are likely to be annotated as coding if they contain a prospective CDS that is considered biologically plausible, based on a mechanistic view of translation. These re-annotations are summarized in Supplementary Table 4.

We binned cases into three main categories, according to confidence in both the accuracy and potential functional relevance of the overlapping models: (1) ‘error’, where the model was seen to have an incorrect
transcript structure and/or a CDS that conflicted with updated GENCODE annotation criteria (these annotations had been or will be changed in future GENCODE releases based on this evaluation); (2) ‘putative’, where the model structure and CDS satisfied our current annotation criteria, although we judged the potential of the transcript represented to encode a protein with a functional role in cellular physiology to be nonetheless speculative (these have been maintained as putative protein-coding transcripts in GENCODE); (3) ‘validated’, where we believe it is highly probable that the model represents a true protein-coding isoform. High confidence in the validity of the CDS was based on comparative annotation, i.e. the observation of CDS conservation and also the existence of equivalent transcript models in other species. GENCODE also annotates transcript models as ‘nonsense-mediated decay (NMD) and ‘non-stop decay’ (NSD), where a translation is found that is predicted to direct the RNA molecule into cellular degradation programs. While it has been established that such ‘non-productive’ transcription events can play a role in gene regulation and thus disease, the interpretation of variants within NMD and NSD CDS remains challenging(44). These models were therefore classed in a separate category.

**Gene list comparisons**

To evaluate the filtering power of the pext metric for Mendelian variants, we evaluated the number of variants that would be filtered with an average GTEx pext cutoff of 0.1 (low expression) in the ClinVar and gnomAD datasets. We downloaded the ClinVar VCF from the ClinVar FTP (version dated 10/28/2018), imported it into Hail, annotated it with VEP v85 against Gencode v19, and added pext annotations with the tx_annotate function. All evaluated variants were annotated as HC by LOFTEE v1.0, and ClinVar variants were filtered to those marked as pathogenic, with no conflicts, and reviewed with at least one star status.

For variants in 61 haploinsufficient genes, we identified any variant identified in at least one individual with any zygosity in both datasets. For variants identified in autosomal recessive disease genes, we used a list of 1,183 OMIM disease genes deemed to follow a recessive inheritance pattern by Blekhman et al.(45) and Berg et al.(46) (available as https://github.com/macarthur-lab/gene_lists/blob/master/lists/all_ar.tsv). We compared the pext value for all pLoF variants identified in ClinVar versus any variant in a homozygous state in at least one individual in the gnomAD exome or genome datasets. Finally, we used a LOEUF cutoff of 0.35 to denote constrained genes, and compared any synonymous or pLoF variant in these genes in the gnomAD exome or genome datasets.

**De novo and rare variant analysis**

De novo variants were collated from previously published studies. We collected de novo variants identified in 5,305 probands from trio studies of intellectual disability/developmental disorders (Hamdam et al.(27): n = 41, de Ligt et al(28): N = 100, Rauch et al(29): N = 51, DDD(24): n = 4,293, Lelieveld et al(26): n = 820), 1,073 probands with congenital heart disease with co-morbid developmental delay (Sifrim et al(42): n = 512, Chih Jin et al(47) : 561), 6,430 ASD probands, and 5,179 unaffected controls from the Autism Sequencing Consortium(25). We also utilized a previously published dataset of variants in 8,437 cases with ASD and/or attention-deficit/hyperactivity disorder and 5,214 controls from the Danish Neonatal Screening Biobank (48). In this analysis, we analyzed pLoF variants identified in highly constrained genes (first LOEUF decile) with a combined total allele count of ≤10 in cases and controls.

We annotated both de novo and rare variants with VEP v85 against Gencode v19 and added pext annotations with the tx_annotate function. We then calculated the average pext metric across 11 GTEx brain samples and binned them as low (pext < 0.1), medium (0.1 ≤ pext ≤ 0.9) or high (pext > 0.9) expression. We then calculated the number of pLoF, missense, and synonymous variants per pext expression bin. To obtain case-control rate ratios and the 95% confidence intervals for de novo variant analyses, we used a two-sided Poisson exact test on counts(49). To obtain the odds ratio for the rare variant analysis in ASD/ADHD, we used the Fisher’s exact test for count data.

**Isoform quantifications via salmon**

To evaluate whether use of a different isoform quantification tool would affect results, we compared results of TCF4 base-level expression (shown in Figure 2B), MAPS (Figure 3C) and comparison of the number of variants filtered in haploinsufficient developmental disease genes in ClinVar vs gnomAD (Figure 4A) using RSEM quantifications used in this study with quantifications using salmon v.0.12(50). Due to the intractability of re-quantifying the entire GTEx dataset, we downloaded and requantified 151 GTEx brain – cortex CRAM files from the V7 dataset. We first converted CRAMs to fastq files using Picard 2.18.20 and ran salmon with the “salmon quant -i index -fastq1 – fastq2 –minAssignedFrag1 –validateMappings” command. The index was created with the “salmon index –t transcript.fa –type quasi –k 31” command using the GENCODE v19 protein-
Transcript expression aware annotation with fetal isoform expression dataset

While our analyses were based on transcript expression aware annotation from the GTEx v7 dataset, we provide necessary files for pext annotation from the Human Brain Development Resource (HBDR) fetal brain dataset\(^1\) in gs://gnomad-public/papers/2019-tx-annotation/data/HBDR_fetal_RNAseq. HBDR includes 558 samples from varying brain subregions across developmental time points. We downloaded HBDR sample fastq files from European Nucleotide Archive (study accession PRJEB14594) and obtained RSEM isoform quantification on HBDR fastqs using the GTEx v7 quantification pipeline, publicly available at https://github.com/broadinstitute/gtex-pipeline/\(^2\) which briefly involves 2-pass alignment with STAR v2.4.2a\(^3\) and isoform quantification with RSEM v1.2.22. Here, we also removed genes where the average pext across HBDR was below 0.2, resulting in the removal of 712 genes (3.5% of all analyzed genes). The dataset was also used for the analysis of baselevel expression values in SCN2A shown in Supplementary Figure 7D.

Supplementary Materials and Methods

**Fig S1:** Details of manual curation of 401 pLoF variants in 61 HI developmental disease genes

**Fig S2:** Technical artifacts in transcriptome sequencing experiments prevent the use of read pileup at exons as an unbiased proxy for expression

**Fig S3:** Details of calculating transcript-expression annotation

**Fig S4:** Baselevel TCF4 expression per GTEx tissue

**Fig S5:** Functional validation of pext

**Fig S6:** Results of GENCODE curation of 128 unexpressed regions in HI developmental disease genes

**Fig S7:** Evaluation of pext after accounting for LOFTEE flags in ClinVar and de novo analyses

**Fig S8:** Application of pext binning to de novo and rare variant analysis in additional datasets

**Fig S9:** Evaluation of using differing cutoffs for average, minimum or maximum pext across tissues

**Fig S10:** Using pext based on fetal RNA-seq isoform expression data for filtering variants in haploinsufficient genes

**Fig S11:** Comparison of key results using Salmon vs RSEM.

**Fig S12:** Results from FUMA GENE2FUNC analysis in unconserved regions with high expression values

**Table S1:** Summary of manual curation flags for 401 pLoFs in 61 HI developmental disease genes identified in gnomAD

**Table S2:** Summary of criteria for LoF verdicts of 401 pLoF in 61 HI developmental disease genes identified in gnomAD

**Table S3:** Manual curation results of 401 pLoFs in 61 HI developmental disease genes identified in gnomAD

**Table S4:** GENCODE curation results of 128 regions flagged as unexpressed by pext

**Table S5:** FUMA GENE2FUNC analysis results and run information

References

1. K. J. Karczewski et al., Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. 531210 (2019).
2. G. Consortium et al., Genetic effects on gene expression across human tissues. Nature 550, 204 (2017).
3. D. G. MacArthur et al., Guidelines for investigating causality of sequence variants in human disease. Nature 508, 469 (2014).
4. D. B. Goldstein et al., Sequencing studies in human genetics: design and interpretation. Nature Reviews Genetics 14, 460 (2013).
5. I. E. Dick, R. Joshi-Mukherjee, W. Yang, D. T. Yue, Arrhythmogenesis in Timothy Syndrome is associated with defects in Ca2+-dependent inactivation. Nature Communications 7, 10370 (2016).
6. I. Splawski et al., CaV1.2 Calcium Channel Dysfunction Causes a Multisystem Disorder Including Arrhythmia and Autism. *Cell* **119**, 19-31 (2004).

7. P. Liao, T. W. Soong, CaV1.2 channelopathies: from arrhythmias to autism, bipolar disorder, and immunodeficiency. *Pflügers Archiv - European Journal of Physiology* **460**, 353-359 (2010).

8. I. Splawski et al., Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations. **102**, 8089-8096 (2005).

9. J. Nousbeck et al., A Mutation in a Skin-Specific Isoform of <em>SMARCAD1</em> Causes Autosomal-Dominant Adermatoglyphia. *The American Journal of Human Genetics* **89**, 302-307 (2011).

10. A. Guven, A. Tolun, TBC1D24 truncating mutation resulting in severe neurodegeneration. **50**, 199-202 (2013).

11. A. M. Roberts et al., Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease. *Science translational medicine* **7**, 270ra276-270ra276 (2015).

12. D. G. MacArthur et al., A systematic survey of loss-of-function variants in human protein-coding genes. *Science (New York, N.Y.)* **335**, 823-828 (2012).

13. W. Zheng, L. M. Chung, H. J. B. B. Zhao, Bias detection and correction in RNA-SeQCing data. **12**, 290 (2011).

14. M. Teng et al., A benchmark for RNA-seq quantification pipelines. *Genome Biology* **17**, 74 (2016).

15. B. Li, C. N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

16. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525 (2016).

17. R. Patro, G. Duggal, M. I. Love, A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417 (2017).

18. J. D. Sweatt, Pitt–Hopkins Syndrome: intellectual disability due to loss of TCF4-regulated gene transcription. *Experimental & Molecular Medicine* **45**, e21 (2013).

19. M. F. Lin, I. Jungreis, M. Kellis, PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275-i282 (2011).

20. A. Frankish et al., GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47**, D766-D773 (2018).

21. M. Lek et al., Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285 (2016).

22. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics Chapter 7*, Unit7.20-Unit27.20 (2013).

23. M. J. Landrum et al., ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research* **46**, D1062-D1067 (2018).

24. S. Deciphering Developmental Disorders et al., Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433 (2017).

25. F. K. Satterstrom et al., Novel genes for autism implicate both excitatory and inhibitory cell lineages in risk. 484113 (2018).

26. S. H. Lelieveld et al., Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nature Neuroscience* **19**, 1194 (2016).

27. F. F. Hamdan et al., De novo mutations in moderate or severe intellectual disability. *PLoS genetics* **10**, e1004772-e1004772 (2014).

28. J. de Ligt et al., Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. **367**, 1921-1929 (2012).

29. A. Rauch et al., Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet* **380**, 1674-1682 (2012).

30. M. L. Tress, F. Abascal, A. Valencia, Alternative Splicing May Not Be the Key to Proteome Complexity. *Trends in biochemical sciences* **42**, 98-110 (2017).

31. B. J. Blencowe, The Relationship between Alternative Splicing and Proteomic Complexity. *Trends in biochemical sciences* **42**, 407-408 (2017).

32. E. V. Minikel et al., Evaluating potential drug targets through human loss-of-function genetic variation. 530881 (2019).

33. A. Dobin et al., STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15-21 (2013).

34. D. S. DeLuca et al., RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530-1532 (2012).

35. E. R. Riggs et al., Copy number variant discrepancy resolution using the ClinGen dosage sensitivity map results in updated clinical interpretations in ClinVar. **39**, 1650-1659 (2018).
36. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 (2013).
37. W. J. Kent et al., The Human Genome Browser at UCSC. 12, 996-1006 (2002).
38. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. 20, 110-121 (2010).
39. W. McLaren et al., The Ensembl Variant Effect Predictor. 17, 122 (2016).
40. K. Watanabe, E. Taskesen, A. van Bochoven, D. Posthuma, Functional mapping and annotation of genetic associations with FUMA. Nature Communications 8, 1826 (2017).
41. D. Croft et al., Reactome: a database of reactions, pathways and biological processes. Nucleic acids research 39, D691-D697 (2011).
42. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28, 27-30 (2000).
43. M. Ashburner et al., Gene Ontology: tool for the unification of biology. Nature Genetics 25, 25 (2000).
44. P. J. da Costa, J. Menezes, L. Romão, The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease. The International Journal of Biochemistry & Cell Biology 91, 168-175 (2017).
45. R. Blekhman et al., Natural selection on genes that underlie human disease susceptibility. Current biology : CB 18, 883-889 (2008).
46. J. S. Berg et al., An informatics approach to analyzing the incidentalome. Genetics in medicine : official journal of the American College of Medical Genetics 15, 36-44 (2013).
47. S. C. Jin et al., Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nature Genetics 49, 1593 (2017).
48. F. K. Satterstrom et al., ASD and ADHD have a similar burden of rare protein-truncating variants. 277707 (2018).
49. J. Przyborowski, H. Wilenski, Homogeneity of Results in Testing Samples from Poisson Series: With an Application to Testing Clover Seed for Dodder. Biometrika 31, 313-323 (1940).
50. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression. Nature methods 14, 417-419 (2017).
51. S. J. Lindsay et al., HDBR Expression: A Unique Resource for Global and Individual Gene Expression Studies during Early Human Brain Development. 10, (2016).

Acknowledgements
We thank all of the research participants for contributing their data. This work was supported by NIDDK U54 DK105566, NIGMS R01 GM104371, and the Broad Institute. KJK was supported by NIGMS F32 GM115208. A.O.L was supported by NICHD K12 HD052896. The GENCODE project is supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U41HG007234. The results published here are in part based upon data: 1) generated by The Cancer Genome Atlas managed by the NCI and NHGRI (accession: phs000178.v10.p8), Information about TCGA can be found at http://cancergenome.nih.gov, 2) generated by the Genotype-Tissue Expression Project (GTEx) managed by the NIH Common Fund and NHGRI (accession: phs000424.v7.p2), 3) generated by the Exome Sequencing Project, managed by NHLBI, 4) generated by the Alzheimer’s Disease Sequencing Project (ADSP), managed by the NIA and NHGRI (accession: phs000572.v7.p4). We thank Emma Pierce-Hoffman for prior analysis and thoughts on characterizing loss of function variants in haploinsufficient genes. We thank the iPSYCH/SSI/Broad Institute psychiatric genetics study for the use of exome count data.

We have complied with all relevant ethical regulations. This study was overseen by the Broad Institute’s Office of Research Subject Protection and the Partners Human Research Committee, and was given a determination of Not Human Subjects Research. Informed consent was obtained from all participants.

Group authors

Genome Aggregation Database Production Team: Jessica Alföldi1,2, Irina M. Armean3,1,2, Eric Banks4, Louis Bergelson5, Kristian Cibulskis4, Ryan L Collins1,5,6, Kristen M. Connolly1,7, Miguel Covarrubias4, Beryl Cummings1,2,8, Mark J. Daly1,2,9, Stacey Donnelly1, Yossi Farjoun1, Steven Ferriera10, Laurent Francioli1,2, Stacey Gabriel10, Laura D. Gauthier4, Jeff Gentry4, Namrata Gupta10,1, Thibault Jeandet4, Diane Kaplan4, Konrad J. Karczewski1,2, Kristen M. Laricchia1,2, Christopher Llanwarne4, Eric V. Minikel1, Ruchi Munshi4, Benjamin M Neale1,2, Sam Novod4, Anne H. O’Donnell-Luria1,11,12, Nikelle Petrillo4, Timothy Poterba9,2,1, David
