

Widespread naturally variable human exons aid genetic interpretation

Received: 29 August 2024

Accepted: 16 October 2025

Published online: 16 December 2025

 Check for updates

Hannah N. Jacobs^{1,2}, Bram L. Gorissen^{3,4}, Jeremy Guez^{2,3,4},
Masahiro Kanai^{2,4}, Kavi Gupta⁵, Hilary K. Finucane^{2,4},
Konrad J. Karczewski^{2,3,4} & Christopher B. Burge¹ ✉

Most mammalian genes undergo alternative splicing. The splicing of some exons has been acquired or lost in specific mammalian lineages, but differences in splicing within the human population are poorly understood. Using GTEx tissue transcriptomes from 838 individuals, we identified 57,271 “naturally variable exons” (NVEs) – exons which are included in mRNAs in some individuals but entirely excluded from others (or vice versa). NVEs impact three quarters of protein-coding genes, occur at all population frequencies, and are often absent from reference annotations. NVEs are more abundant in genes depleted of genetic loss-of-function mutations and aid in the interpretation of causal genetic variants. Genetic variants modulate the splicing of many NVEs, and 5′ untranslated region and coding-region NVEs are often associated with increased and decreased gene expression, respectively. Together, our findings characterize abundant splicing variation in the human population, with implications for a range of human genetic analyses.

Nearly all human genes undergo alternative splicing, in which distinct mRNAs derived from different combinations of exons and splice sites are produced from the same gene¹. Small changes in mRNA sequence can exert large effects, including the addition/deletion of protein domains^{2,3}, or production of functional versus nonfunctional isoforms, and dysregulation of splicing contributes to human disease^{4,5}. Understanding variation in alternative splicing is therefore important in understanding physiological and disease mechanisms.

Evolutionarily conserved alternative exons tend to preserve protein reading frame⁶, and often exhibit tissue-specific regulation⁷. In addition, comparison across mammalian species has identified thousands of cases of complete loss or gain of the splicing of exons⁸, with more recently evolved exons more likely to occur in 5′ UTRs, where they are often associated with increased gene expression⁸. Gain or loss of the splicing of an exon in the coding sequence (CDS) may also influence expression by changing the reading frame, often triggering nonsense-mediated mRNA decay (NMD)^{9,10}. However, the

evolution of alternative splicing within the human species remains poorly understood.

Most alternative exons are generally assumed to be spliced in all individuals, perhaps with different inclusion levels, a pattern which we call ‘canonical’ alternative splicing. Some such splicing differences are driven by genetic variation. Large-scale genetic association studies have associated changes in the usage of many alternatively spliced exons with single-nucleotide polymorphisms (SNPs), designated splicing quantitative trait loci (sQTLs)^{11,12}. However, these variants typically have modest effect sizes, and their molecular mechanisms are often unclear. Ultra-rare genetic variants can sometimes lead to inclusion of previously unseen, “cryptic” exons or splice sites¹³. Some recent studies suggest that many exons are missing from transcriptome reference annotations, occurring only in a subset of transcriptomic datasets^{11,14,15}. The availability of tissue transcriptomes from nearly 1000 individuals in the Genotype-Tissue Expression (GTEx) project presents an opportunity to characterize human variation at the exon

¹MIT Department of Biology, Cambridge, Massachusetts, USA. ²Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. ³Novo Nordisk Foundation Center for Genomic Mechanisms of Disease, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. ⁴Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. ⁵Computer Science and Artificial Intelligence Laboratory, MIT Cambridge, Massachusetts, USA. ✉ e-mail: cburge@mit.edu

level, asking how often the presence/absence of individual exons or splice sites varies between the transcriptomes of different individuals.

One key challenge in the study of unannotated transcripts is assessing which reflect technical artifacts or other types of noise^{16,17}, which are reproducible, and which have phenotypic effects¹⁸. A cross-species analysis has suggested that many unannotated splicing events in longer-lived organisms may represent “transcript drift”, and are often nonfunctional until they begin to splice in at higher levels¹⁹. However, another recent study found that lowly used exons that trigger NMD broadly impact gene expression in human lymphoblastoid cell lines, and can help interpret complex trait loci¹⁰.

Here, we define “naturally variable exons” (NVEs) as exons whose splicing is variable within the human population (i.e., not present in every individual), focusing on alternative “skipped” or “cassette” exons, and alternative splice sites, which extend or truncate exons. We introduce a statistical approach that estimates the population frequency with which NVEs are meaningfully included (at levels above 5%) in a human cohort, generating a quantitative atlas of splicing variation across tissues. Using data from GTEx, we find that NVEs impact most human genes, are sometimes associated with changes in gene expression, and can aid in the interpretation of non-coding genetic variants.

Results

Generating a catalog of human NVEs

In order to generate a catalog of high-quality NVEs, we obtained mapped RNA-seq reads from 14,000 GTEx samples representing 49 tissues from 838 individuals¹¹. To perform quality control and ensure accurate estimates of NVEs in the dataset, we devised a straightforward Bayesian approach. The method de-noises estimates by sharing signals across samples and enables estimation of the uncertainty in exon usage of each NVE.

To illustrate the method, consider an NVE of the alternative splice site variety, used in some individuals as an alternative to a canonical or “cognate” splice site, used in all people. The proportional use of the NVE within an individual – the “percent spliced in” (PSI or Ψ) – can be estimated from the proportion of RNA-seq reads that span the NVE and cognate exon junctions (EJs) (Supplementary Fig. 1A)²⁰. NVEs tend to have fewer total junction reads per sample, making some estimates of Ψ less certain (Supplementary Fig. 1B). We therefore de-noised these estimates using a beta-binomial model, employing a mixture of 3 beta distributions for reasons discussed below. The workflow is described in Supplementary Fig. 1C, Methods and Supplementary Note Sections A–D. The method fits a smooth function representing the NVE’s underlying Ψ distribution across the population from the observed splice junction reads for an NVE across individuals in a tissue. We chose a mixture of 3 betas model because of its flexibility to model many different distributions and because many NVEs are associated with sQTLs (see below), which have 3 possible genotypes that may each generate distinct levels of splicing. The mixture of betas model can be used to assess uncertainty in Ψ values, and to estimate the proportion of individuals that splice the NVE at any given level. Here, we defined NVEs as alternative exons or splice sites that have estimated Ψ values of at least 5% in between 1% and 99% of the population in a particular tissue (see “Methods”); exons with $\Psi \geq 5\%$ in over 99% of individuals are considered canonical alternative exons, and exons that don’t meet either of these criteria may be considered rare or cryptic exons. Using this definition, we identified 414,141 NVE-tissue pairs across all 49 tissues in the GTEx dataset (Supplementary Data 1), representing 57,271 unique NVEs. Like canonical alternative exons, NVEs are roughly balanced between the alternative splice site and skipped exon types (Supplementary Data 2).

Using these data, we define an interpretable summary statistic, exon frequency (EF), that describes the frequency of splicing of an NVE

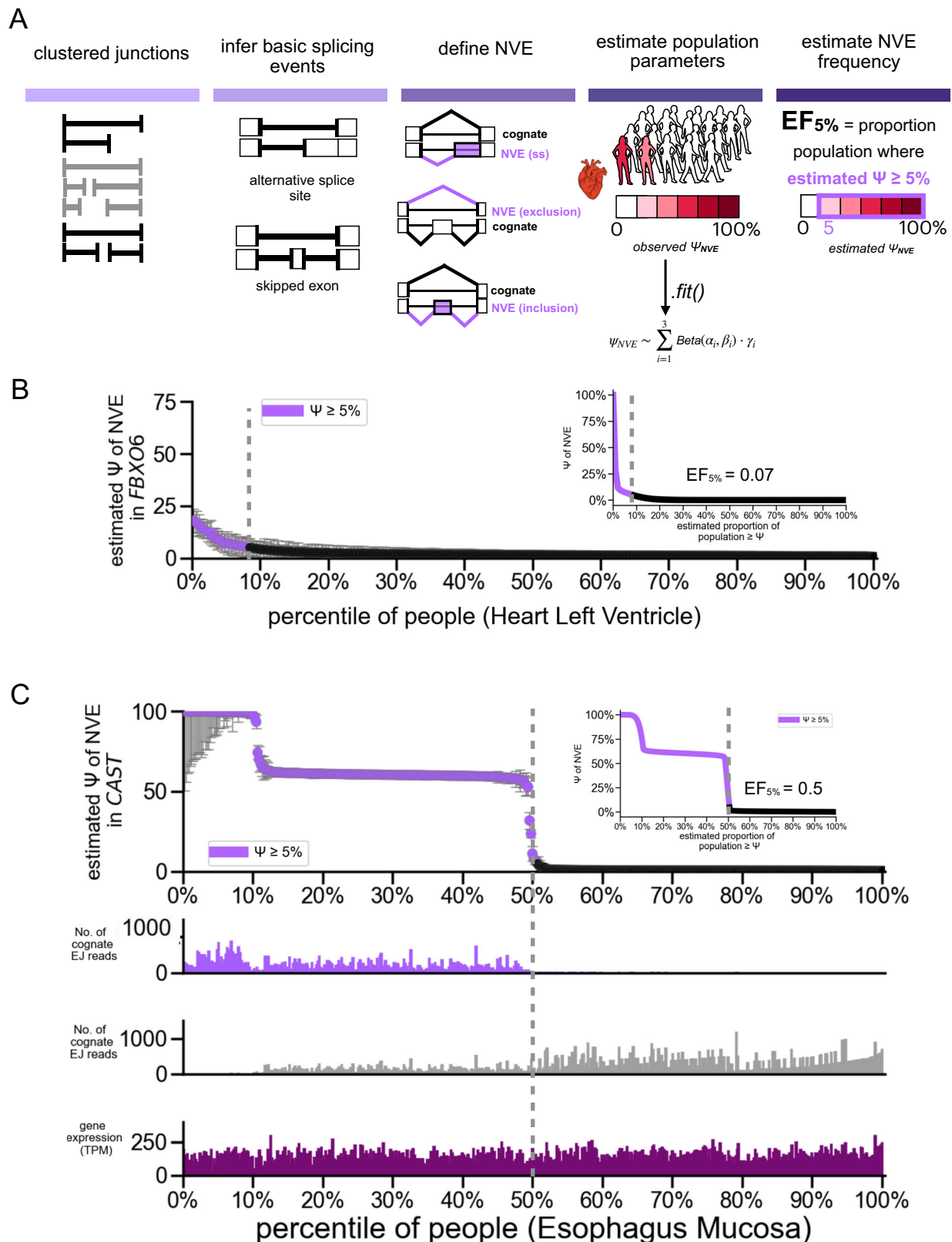
in a population, somewhat analogous to allele frequency (AF). The EF of an NVE in a tissue is the estimated proportion of individuals that splice the NVE at a threshold Ψ level or above. We typically employ a 5% Ψ threshold here because previous studies have suggested that this level of inclusion is near the lowest level where sequence conservation is commonly observed¹⁹. Distributions of EF values using different Ψ cutoffs are shown in Supplementary Fig. 1D, E; thresholds of 5, 10 and 20% yield generally similar results. The steps in data processing to estimate EF for an NVE are described in Fig. 1A. EF is in some ways analogous to AF, but because estimation of EF depends on the availability of RNA-seq datasets from relevant tissues, the precise value is probably less important than the utility in stratifying NVEs into rarer and more common subsets.

To illustrate, we provide an example of a fairly typical NVE, which occurs in the *FBXO6* mitotic regulator nucleolar GTPase gene and has an EF of 0.07 (Fig. 1B). This exon is very lowly or not included in most individuals, but has moderate inclusion of up to ~20% in a handful of individuals, with mean \pm standard deviation of Ψ well separated from zero in a small portion of the population. This splicing variation occurs against a background of fairly uniform gene expression (Supplementary Fig. 2A). The underlying factors that drive splicing differences between individuals might include genetic variation acting in *cis* or in *trans*, or environmental factors, and are not clear in this particular case. Examples of NVEs across the spectrum of EFs observed (which range from 1% to 99%) are shown in Supplementary Fig. 2A–D. The splice site motifs of NVEs match consensus motifs to a similar degree as the splice sites of cognate exons (Supplementary Fig. 2E, F), similar to observations for alternative exons overall²¹. We provide various summary statistics describing the splicing of NVEs, such as median and maximum Ψ across the spectrum (Supplementary Figs. 3 and 4), observing positive relationships between most measures of exon inclusion and EF (Supplementary Fig. 4C, D). NVEs can even be constitutively spliced in some people, and absent from others, as shown by a second example, occurring in the calpastatin gene *CAS7* (Fig. 1C). These examples highlight diversity in the splicing patterns of exons across individuals and flexibility provided by our three-component mixture model.

We next sought to understand the distribution of NVEs across genes and individuals. Since many NVEs are detected in multiple tissues, we focused on the tissue where the exon achieves its highest EF, which we call “top EF” and use as the default EF value for an NVE. (Similar results were obtained when using other criteria, such as the 90th percentile of EF across tissues.) Using this approach, we identified 28,694 alternative splice site NVEs (NVE_{alt ss}) and 28,577 skipped exon NVEs (NVE_{se}) in all of GTEx. We find that NVEs are widespread, occurring in 75% of protein-coding genes, with a median of 3 NVEs per gene (Fig. 2A). Across all 49 tissues, some variability was observed in the number of NVEs, with a median of ~3500 NVEs per tissue (Supplementary Fig. 5A). We estimated the average number of NVEs expressed per individual as ~13,100, using a conservative approach that sums top-EFs calculated from the ten best-sampled tissues. Although low-EF NVEs are numerous, those with higher EFs are more commonly observed, so the EF distribution of the set of NVEs that occur in any given individual will be skewed toward higher EF values (Supplementary Fig. 5B). As a result, the fraction of NVEs shared by any two unrelated individuals – calculated as the sum of the squares of EF values over all NVEs – is estimated at 28%.

Most low EF NVEs occur in coding regions and many high EF in 5'UTRs

We found that the distribution of EF values of NVEs is U-shaped, with many NVEs being either rare or frequent in the population (Fig. 2B). We estimate that the mean Ψ of an NVE in individuals that splice above our threshold level (5%) ranges from near 5% for low-EF NVEs (top EF 0.1 or below) to just over 25% for high-EF NVEs (top EF 0.9 or above) (Fig. 2C).



The RNA-seq read depths available in GTEx are more than adequate to distinguish these moderate ψ values from the absence of splicing, and NVEs possess splice site motifs very similar to those of other exons (Supplementary Fig. 2D), even for low-EF NVEs (Supplementary Fig. 2E).

We explored features associated with NVEs having different EF values. As noted above, more highly included exons (higher

median ψ) tended to have higher EF values (Supplementary Fig. 4C, D). For NVE_{alt ss}, their 5' and 3' splice site (SS) motif scores (using MaxEnt²²) tend to be slightly weaker than their associated cognate SS, with the difference shrinking at higher EF values (Supplementary Fig. 5C). These observations suggest that the SS strength of NVEs is a contributor to their inclusion and EF values.

Fig. 1 | Overview of the method to detect NVEs and examples. A Model figure describing NVE detection and our summary statistic, exon frequency. To define NVEs, we first filtered LeafCutter outputs for either alternative splice sites or skipped exons. For a given NVE observed in the RNA sequencing data, the Ψ of an individual can be directly estimated using the beta binomial model. The EF summary statistic is the percentage of individuals whose Ψ exceeds a given threshold, shown for a 5% Ψ threshold. **B** An example exon in *FBXO6* with $EF_{5\%} = 0.07$ in heart tissue. Individuals with heart samples ($n = 386$) are sorted by their estimated Ψ values for this exon, which is shown as percentiles on the x-axis. The mean posterior estimate of Ψ is plotted (highlighted in purple if the

estimate is at least 5% Ψ in the individual), along with error bars showing one standard deviation on Ψ . In the inset, an example of the population distribution used to calculate the exon frequency (EF) of this exon. Note that the y-axis scales differ between plots. **C** An example of an NVE in the *CAST* gene (with $EF_{5\%} = 0.5$) that has high variability amongst individuals. Individuals with Esophagus Mucosa samples ($n = 497$) are sorted by their estimated Ψ values for this exon, which is shown as percentiles on the x-axis. Inset: as in B. Below: bar plots of EJ reads for both NVE and cognate, and overall mRNA levels (TPM). Note that the y-axis scales differ between plots.

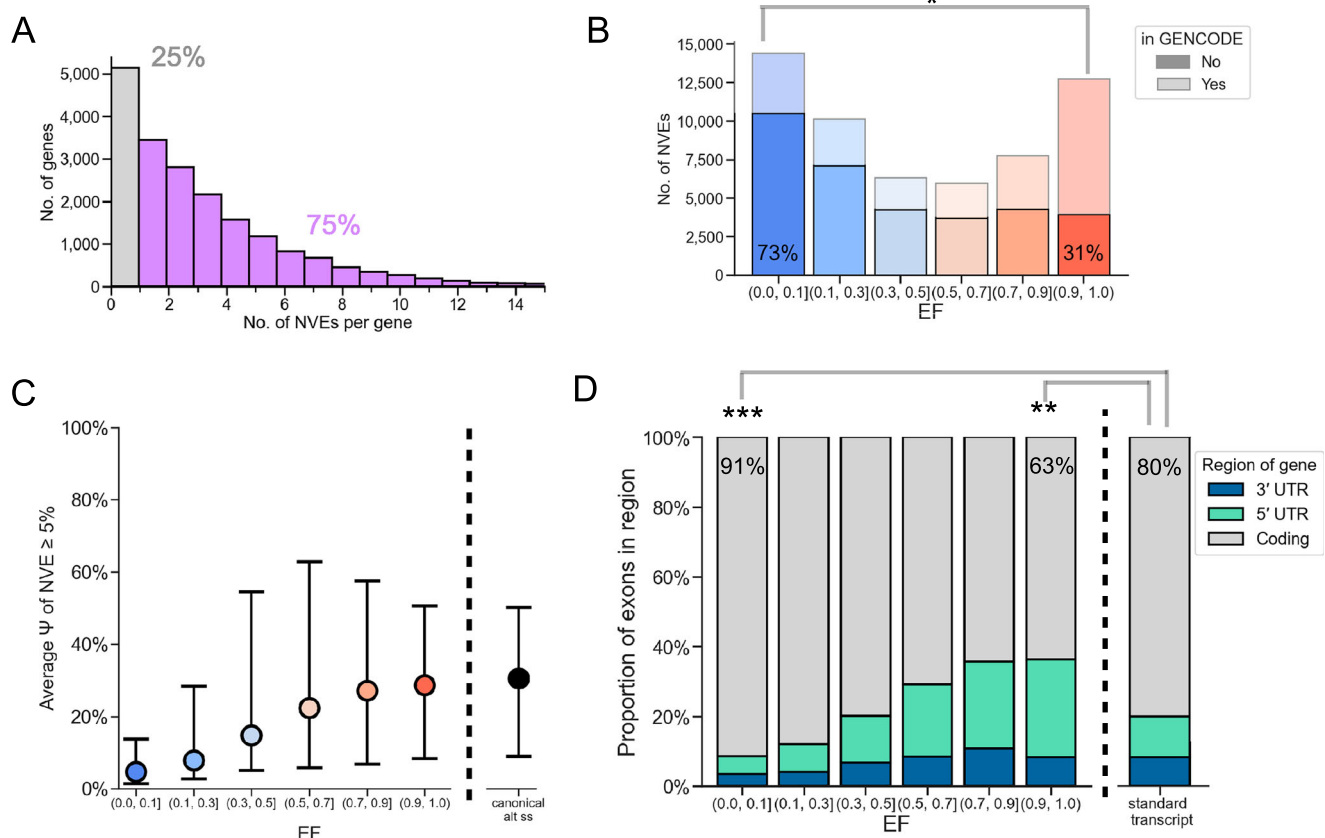


Fig. 2 | The distribution of NVEs across genes, gene regions, and their relative usage in individuals. A Histogram of the number of NVEs in a given gene, including those protein-coding genes with no NVEs (gray) at least one NVE (purple). **B** The top EF spectrum in GTEx. For each NVE, the tissue with the most usage of the NVE was taken, indicating the highest possible EF in GTEx. We use this color scheme throughout the paper, with lower EFs shown in blue and higher EFs shown in red. Lighter boxes indicate that the NVEs are not in GENCODE reference annotations. Star indicates significance ($p = 10^{-255}$ by performing a simple normal test for proportions (two-sided) of low and high EFs in reference annotations, without multiple

testing correction, and are connected by gray lines. **C** Average ψ of NVE in individuals with $\psi \geq 5\%$. 1000 NVEs were randomly sampled, with 10 bootstraps. We approximated the conditional expectation of $\psi \geq 5\%$ using the empirical mean of observed ψ values above that value, within each EF bin. Uncertainty was represented by the 2.5th and 97.5th percentiles. **D** Proportion of genes in MANE select/RefSeq transcripts in different gene regions (right) and Proportion of NVEs in given gene regions, split by exon frequency (left). Stars indicate to significance ($p < 10^{-24}$ and $p = 10^{-24}$) by performing a simple normal test for proportions (two-sided) of CDS proportions between groups connected by gray lines, without multiple testing correction.

Notably, 61% of NVEs present in GTEx are absent from comprehensive reference annotations (GENCODE 45). NVEs with low EF are predominantly (73%) unannotated, while around one third of high-EF NVEs are unannotated (Fig. 2B). Low-EF NVEs that are absent from reference annotations have more tissue-restricted splicing than those that are present in reference annotations (Supplementary Fig. 5D). Different NVEs were restricted to different subsets of tissues, with no prominent outlying tissues (Supplementary Fig. 5E). In total, NVEs expand the sequence space covered by the transcriptome by ~4 MB (Supplementary Fig. 5F). The number of NVEs detected in an individual was not strongly related with the depth of sequencing (Supplementary Fig. 5G).

In general, alternative exons occur most commonly in coding regions, but also occur often in 5' UTRs and rarely in 3' UTRs of genes (Fig. 2D). Similarly, most NVEs – termed cdsNVEs – occur within the region of the gene that contains coding exons, and this class of NVEs tends to have low EF values. A substantial minority of NVEs occur in 5' UTRs, and this subset tended to have higher EFs (Fig. 2D). This observation is consistent with previous work showing that evolutionarily more recent alternative exons arise most commonly in 5' UTRs⁸. NVEs rarely occur in 3' UTRs, likely because this gene region contains very few introns relative to others (presumably to avoid triggering NMD). For the most part, the size distributions of NVEs are consistent across EFs, particularly for the skipped exon type (Supp. Fig. 6A-E).

There are a substantial number of 3 bp alternative 3' splice sites (3SS), reflecting the abundance of NAGNAGs²¹ (Supplementary Fig. 6AC). Alternative 5' splice sites (5SS) located in 5' UTRs are on average twice as long as non-5'UTR alternative 5SS NVEs (Supplementary Fig. 6E). This observation may be related to the frequent presence of alternative promoters and 'hybrid exons' (overlapping first and internal exons) in the 5' UTRs of many human genes²². In summary, both splice site strength and location in the gene appear to contribute to the emergence and/or maintenance of NVEs in the human population.

NVEs impact mutationally-constrained genes

To explore the evolutionary properties of genes with and without NVEs, we considered gene tolerance to germline loss-of-function (LoF) mutations in the human population, which is a proxy for both gene function and natural selection²³. This property has been quantified across nearly all human genes by computing the LoF observed to expected upper bound fraction (LOEUF) across a population, where lower scores correspond to more constraint (greater intolerance to LoF) in the gene²³. Surprisingly, we find that the distribution of LOEUF scores is lower for NVE-containing genes than genes which lack NVEs (Fig. 3A, significant by KS test: p -value 10^{-212}), even after adjusting for gene expression (Fig. 3B, Supp. Fig. 6G). We observed a modest negative correlation (Spearman $\rho = -0.13$) between LOEUF and the "NVE rate" of a gene – the number of NVEs divided by the number of annotated exons (Supplementary Fig. 6F). Furthermore, the distribution of EFs was shifted toward lower values in more constrained genes (Fig. 3C). Such a pattern could occur if NVEs arise more commonly in more constrained genes. This could be due to a more favorable nuclear environment for splicing in subcompartments such as nuclear speckles²⁴, which have been shown to be involved in coupling splicing with actively transcribed genes, which is common for unconstrained genes²³. Though constrained genes may be more likely to be spliced for this reason, these genes often experience selection favoring lower ψ values, driving EFs lower. NVEs in less-constrained genes tend to have higher EFs, and might more often become canonical alternative exons over evolutionary time. This suggests the alternative possibility that NVEs can arise at certain frequencies in all genes, but their EFs may evolve more rapidly toward 1 (canonical alternative exon) or 0 (ψ below threshold in all people) – neither of which are considered NVEs – in less-constrained genes due to relaxed selection, making NVEs more evolutionarily transient in this subset of genes. Distinguishing among these possibilities would be difficult with current data, but might be enabled by comparing NVEs across human populations to help assess their evolutionary ages, once GTEx-style RNA-seq data are generated in other populations. Genes containing canonical alternative exons – which are spliced in all or virtually all people – showed little bias in LOEUF score relative to all genes (Supplementary Fig. 6H).

NVEs aid interpretation of GWAS variants

To explore whether NVEs could improve disease interpretation, we focused on variants that occur in the extended splice site motifs of NVEs not present in GENCODE comprehensive reference annotations. We analyzed the results of statistical fine-mapping of GWAS for around 1300 traits across three global biobanks: UK Biobank, FinnGen, and Biobank Japan (UKBB, FG, BBJ, respectively)²⁵. Statistical fine-mapping yields a posterior inclusion probability (PIP) for each variant, which reflects the probability that the variant causally drives the association at the locus and enables enrichment analyses of fine-grained annotations, such as NVEs, that are not powered for heritability-based analyses.

The observed enrichments of causal variants in NVEs are above the level of synonymous variants, particularly at high PIP values, but below those of missense variants (Fig. 3D). To ensure that the enrichments were not driven by existing annotations, the enrichments shown exclude GWAS variants already annotated as genetic LoF, splice region,

or missense variants. Performing enrichment analyses separately for each biobank reduced power but yielded similar trends (Supplementary Fig. 7C). While synonymous variants are typically null or nearly null in analyses of ultra-rare variants²⁶, some synonymous variants have been identified as high-confidence causal variants in cross-biobank fine-mapping²⁵, and may cause changes in splicing, mRNA stability or translation, for example, so the enrichment for synonymous variants in high-PIP bins is unsurprising.

We next explored the utility of unannotated NVEs for GWAS interpretation. One example is a pleiotropic synonymous variant in *ASGR1* (rs55714927, MAF 15%) that is both an sQTL and an expression quantitative trait locus (eQTL) and colocalizes with an NVE splice site with EF between 0.07 and 0.60 in different tissues. This common *ASGR1* variant is known to increase the inclusion of this unannotated NVE, and this exon contributes to phenotypes such as cholesterol and heart function^{25,27,28}. We found that, in many cases, the alternate allele was associated with increased splicing of the NVE to between 3 and 10%, while individuals with the reference allele had Ψ values near 0% (Supplementary Fig. 7B). We observed a negative correlation between the Ψ of this NVE and the expression of *ASGR1* (Supplementary Fig. 7B, inset), consistent with the potential of this exon to trigger NMD. This example illustrates a molecular mechanism for a synonymous variant and demonstrates that even NVEs with low Ψ and moderate EF values can be relevant to GWAS loci.

Low observed Ψ NVEs can have outsized impacts on gene expression

To explore the effects of NVEs on coding potential and expression more broadly, we built a custom pipeline to identify cdsNVEs that have NMD potential, including causing a frameshift or introducing a stop codon (Methods), which we refer to as nmdNVEs. We find that 55% of all NVEs and 68% of cdsNVEs are nmdNVEs. nmdNVEs tend to have Ψ values that are lower on average than cdsNVEs that lack NMD potential (Supplementary Fig. 7C).

We considered whether splicing of nmdNVEs tends to reduce gene expression. In order for an NVE to impact gene expression, it must be spliced in to a sufficient extent to meaningfully reduce the level of the cognate isoform. Degradation of the NVE-containing isoform by NMD is expected to reduce both the Ψ value that is observed in RNA-seq data (which reflects levels of mature mRNAs) for nmdNVEs, as well as the expression level (Fig. 4A). Because of the impact of NMD, even a relatively modest observed difference in cytoplasmic PSI value can be predictive of a fairly large change in gene expression. In the hypothetical example illustrated, destabilization of the NMD isoform by 5-fold relative to the canonical isoform²⁹ implies a nuclear PSI of 62% and a 2-fold reduction in gene expression associated with an NMD-triggering exon whose observed PSI is just 25%. Under the same assumptions, a frame-disrupting exon with an observed PSI of 10% would reduce expression by ~29%.

Sufficient read depth is therefore required to both observe any associated change in expression and to detect the splicing of the NVE. Because of these considerations, we chose to explore this question using a subset of NVEs that have large variability in splicing, occur in genes with variable expression, and have sufficient read depth for both types of variation to be readily observed. Specifically, we considered NVE_{alt ss} whose splicing is significantly associated with a genetic variant that impacts both splicing and gene expression, i.e., that is both an sQTL and an eQTL in GTEx. In this set, we observed a negative relationship between inferred NVE sQTL effect size and (directional) eQTL effect size for nmdNVEs, with a consistent relationship observed across nearly all EF thresholds (Fig. 4B). This observation supports the idea that, even at low observed Ψ values, nmdNVEs often reduce expression by shifting the mRNA output of a gene from productive, stable mRNAs to unproductive/unstable

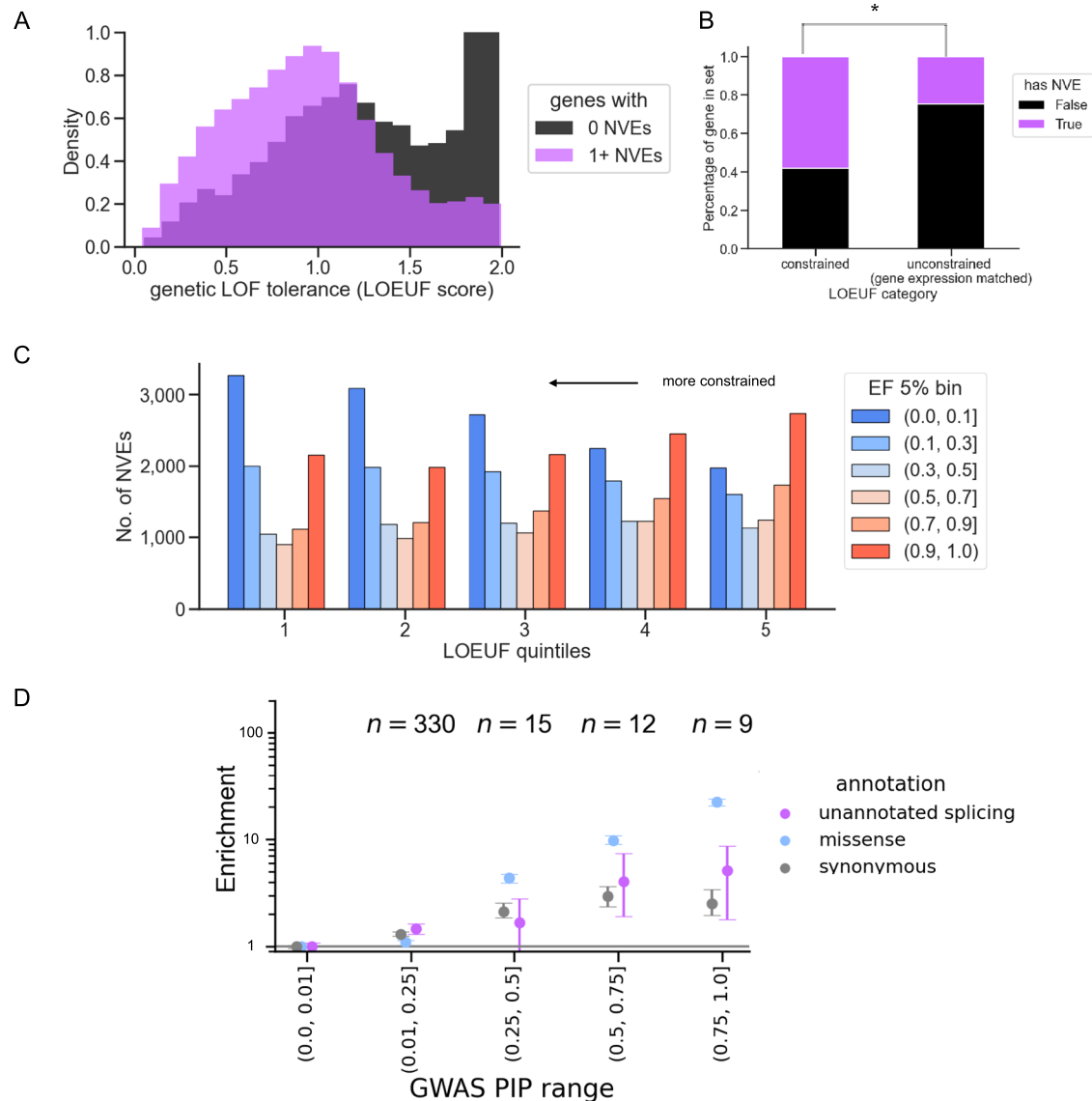


Fig. 3 | NVEs tend to impact more mutationally constrained genes and provide additional functional interpretations in GWAS. A Distribution of LOEUF scores of genes with NVEs (purple) and genes that do not have any NVEs (black), in bins of width 0.1. Significant by the KS test p -value 10^{-212} . **B** Percentage of genes with NVEs in constrained and unconstrained genes, matched on gene expression, in Whole Blood. For each gene, the median TPM across all individuals was taken to assess average expression in blood. Genes were separated as either being constrained (bottom quintile of LOEUF scores in blood) or unconstrained (top quintile of LOEUF scores in blood). Constrained genes were then filtered for those with median expression of over 1 TPM, and matched with unconstrained genes with similar expression, $n = 178$. Significant by proportion normal two-sided z-test in scipy, $p = 10^{-11}$. **C** The number

of genes stratified by NVE EF bin, represented in each LOEUF quintile, with low values indicating more constraint. Genes that contain low EF NVEs are enriched among the most constrained genes, whereas genes with high EF NVEs are largely unconstrained. Here, the EF represents the NVE that is most frequent across the gene. **D** GWAS from three global biobanks (UKBB, FG release 9, and BBJ) were pooled together. We included unannotated canonically alternative exons found in the NVE discovery process to boost power. For the unannotated splice site set, pLOF, annotated splice region, nonsense, and missense variants were filtered out of this set to ensure that enrichment was not driven by well-explained variants. The number of variants being categorized as within an unannotated splice site is shown as n values. Estimates shown as mean \pm 95% CIs of a binomial estimate given n .

mRNAs. By contrast, no significant relationship between sQTL and eQTL effect size/direction was observed for cdsNVEs that are not nmdNVEs. Such NVEs presumably yield unique protein isoforms, though we did not explore this direction further here.

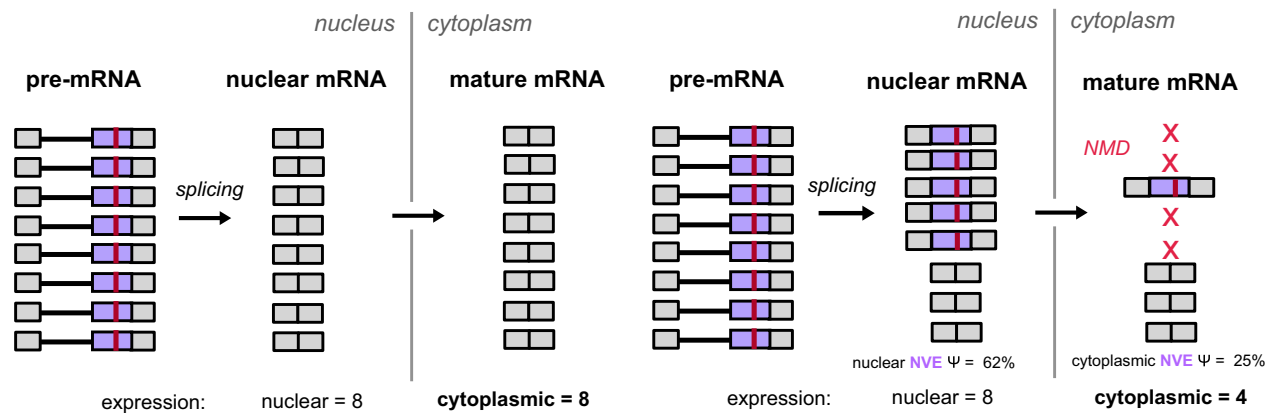
Because many NVEs occur in annotated introns, they may create non-functional mRNA and protein isoforms (whether they trigger NMD or not), which might often be mildly deleterious, particularly in constrained genes. We observed that nmdNVEs tend to have lower Ψ values in constrained genes than in unconstrained genes (Fig. 4C), likely reflecting selection against large perturbations of gene expression in low-LOEUF genes. Notably, for NVEs in 5' UTRs, higher Ψ values were associated with increased gene expression (Supp. Fig. 7D). NVEs

in this gene region may positively impact gene expression via exon-mediated activation of transcription starts (EMATS)³⁰. Together, the observations above indicate that the splicing of NVEs can be driven by genetic variants such as SNPs, that NVEs may impact gene expression negatively or positively, and that, in general, selection may favor reading frame-preserving NVEs and NVEs that exert smaller effects on gene expression over other NVEs.

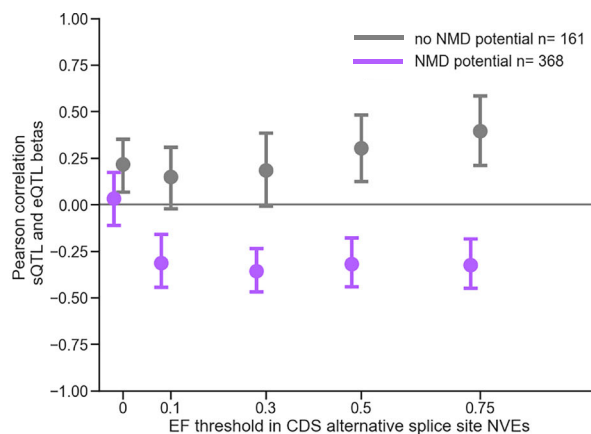
Common genetic variants impact NVEs when located near splice sites

Splicing can be impacted by both genetic and environmental factors^{11,31}. To understand the basis of inter-individual differences

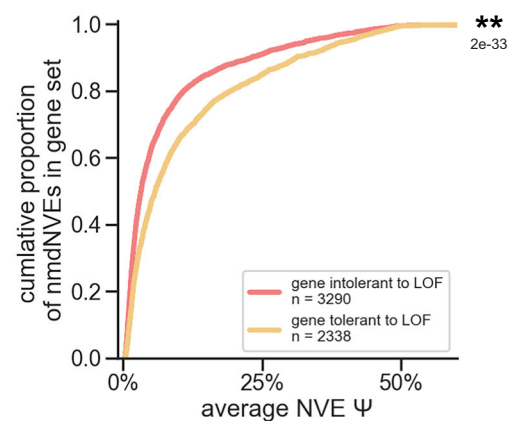
A



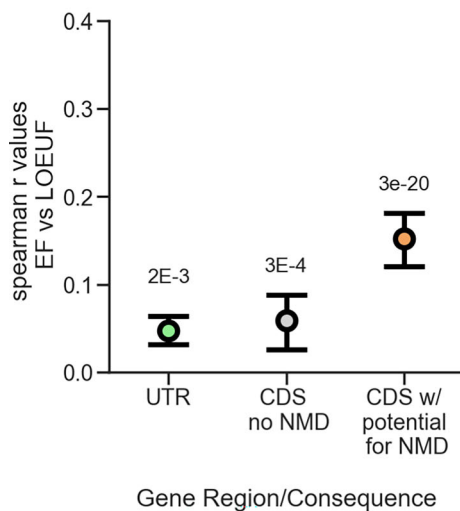
B



C



D



underlying differences in NVE splicing, we asked whether common genetic variants or other factors were associated with the presence of NVEs. We find that 60% of NVEs are associated with at least one sQTL (defined as a variant within a credible set in the fine-mapping dataset) in their top EF tissue, suggesting that many NVEs are modulated by *cis*-acting genetic variants. This proportion varied only modestly across EF bins (Fig. 5A).

This pattern is consistent with prior work showing that splicing variation in GTEx is mostly driven by *cis* genetic effects³². The

identification of causal variants that impact splicing remains challenging³⁸. sQTLs directly associate common genetic variants with splicing measurements, but because of pervasive linkage disequilibrium, they do not directly identify which variant causally impacts splicing, and so do not inform about the mechanism. Fine-mapped sQTLs are a gold standard, but because of limited statistical power, sets of fine-mapped sQTLs are far from complete. For rare variants (MAF < 1%), sQTLs and fine-mapped sQTLs are even less powered. We find that high-confidence (90% PIP or above) sQTL

Fig. 4 | Evidence of gene expression changes of NVEs with NMD-potential across the EF spectrum. **A** Left: NMD-causing NVE (purple, with red bar for stop codon) is not spliced into transcripts, will have no impact on expression. Right: Inclusion of the nmdNVE in 5%th of transcripts in the nucleus implied by the Ψ of 25% in the cytoplasm, assuming stability of nmdNVE isoforms is reduced 5-fold by NMD, resulting in a 2-fold reduction in gene expression. In general, if NMD reduces the stability of nmdNVE isoforms by k -fold of an nmdNVE with cytoplasmic Ψ value of x , gene expression will be reduced by $(k-1)x/(1+(k-1)x)$. **B** Effect size (slope of regression of QTL) shown for genetic variants nominally significant sQTLs for NVEs, and also significant eQTLs in the same gene in same tissue. Genetic variants filtered out if in complex loci (i.e., were eQTLs in other genes in the tissue). Association between variant and splicing event is with cognate splicing event, and a decrease in splicing would imply an increase in the overall NVE:cognate ratio (multiplied effect size by -1). cdsNVEs further separated into those with NMD potential (purple) and

without (black). Shown are mean and range (2%–98%) of 1000 bootstrapped Pearson correlations between sQTL and eQTL effect sizes, performed at increasing EF thresholds. **C** Average Ψ across individuals for a given NVE estimated in a tissue. Showing any EF NVE that causes NMD, further separated depending on if NVE was unconstrained (top 10% percentile of LOEUF) or constrained (bottom 10% of percentile LOEUF) gene. ** Indicates significant KS test (10^{-33}) between the two groups. **D** NVEs filtered into whether they were coding (CDS) and have potential to cause NMD (w/ potential for NMD, orange) or no potential (CDS no NMD, gray) or non-translated regions of the gene (UTR, green). Spearman correlations performed on 1000 sampled NVE per gene region (without replacement), to control for the fact that total coding region (CDS) NVEs are larger in number than total untranslated region (UTR) NVEs. The process was repeated for 100 bootstraps, and the range of estimates are plotted as box plots. Shown is the average p -value for the correlations performed in bootstrapping (and the range from min to max).

variants were typically quite proximal to the splice site, with 50% occurring within 50 bp of the nearest splice site (Supplementary Fig. 8A). Standard sQTL studies perform variant associations across 1 MB around the splice site³³. These observations suggest that the power to detect causal variants is decreased by assessing an unnecessarily large range. These considerations motivate the potential of predicting splice-affecting variants using different sources of information, such as distance to the splice site, and integrating it into a variant prediction model to assess the relative importance of each feature on predicting a variant that impacts splicing.

Using in-sample RNA sequencing improves common variant interpretation for splicing

Several published methods aim to integrate primary sequence information to predict a variant's impact on splicing, including SpliceAI¹⁶, Pangolin³⁴, MMSplice³⁵, and AbSplice¹⁷, among others. SpliceAI, a deep neural model trained on human GENCODE annotations, is one of the most widely-used tools for prediction of splice-impacting variants based on primary sequence¹⁶. However, this method, and methods built on it, such as AbSplice¹⁷, have been less successful in predicting effects of common variants, which often have low effect sizes relative to rare variants. We therefore attempted to improve the identification of more common splice-affecting variants.

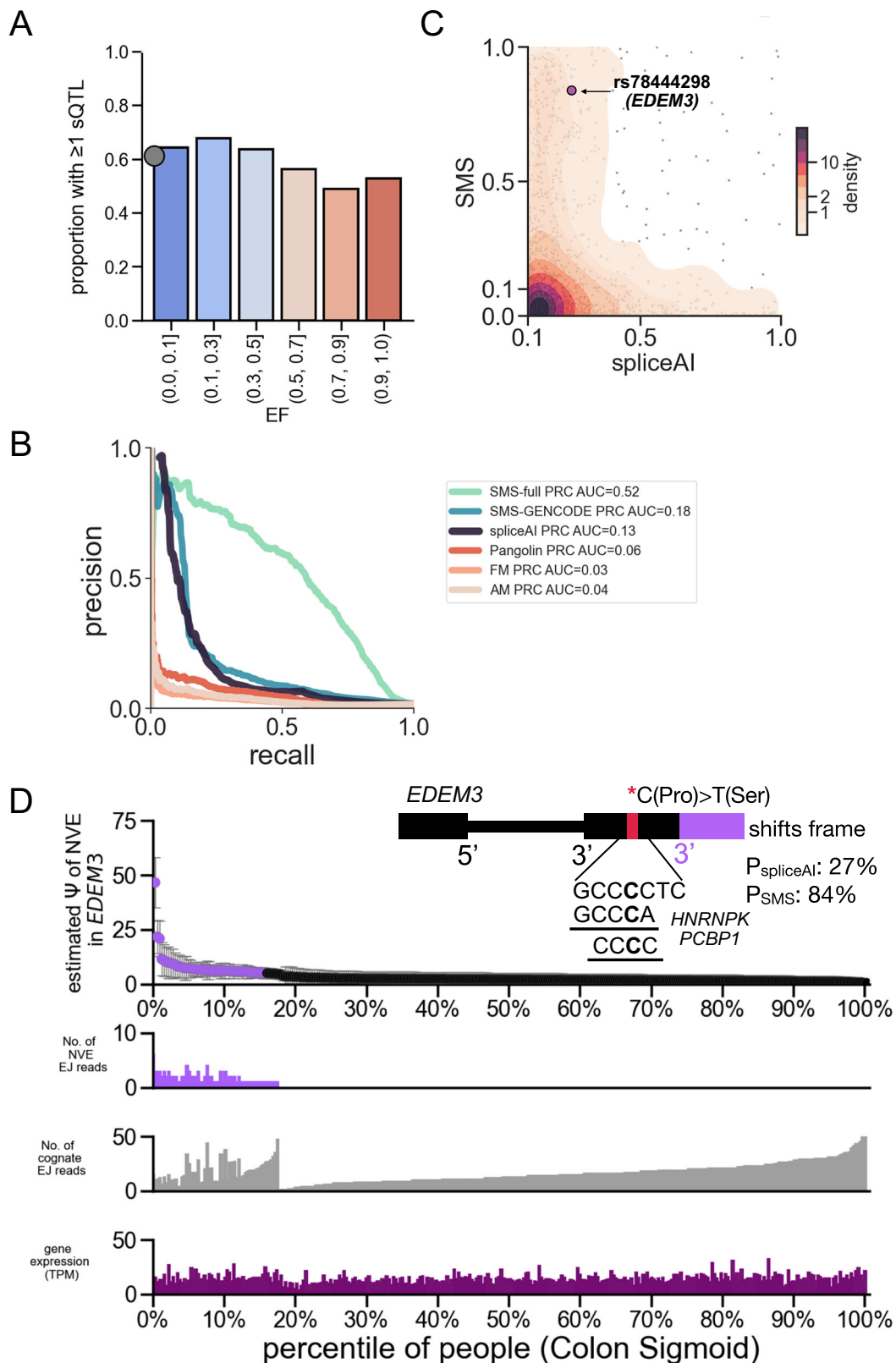
In the prediction of common splice-affecting variants, we considered two distinct situations: the case in which in-sample RNA-seq data is available, and the case where only reference transcriptome data are available. To this end, we developed the Splice Modifier Score (SMS), analogous to the Expression Modifier Score (EMS) for eQTLs³⁶, which uses a regression model to estimate the probability that a variant modifies splicing. We trained SMS using fine-mapped sQTL data³⁷, using very high probability sQTLs (PIP of 90% or above) as positives, and low-probability sQTLs (0.2% or below to improve class balance) as negatives. We consider the top association of a variant that is within 5 kb of a given splicing phenotype in a tissue (a tiny fraction of the original 1 MB window), and only consider skipped exons and alternative splice sites in protein-coding genes.

To train the model, we annotated genetic variants with a set of molecular features including distance to nearest splice site, known exonic splicing enhancer³⁸ and silencer³⁹ motifs, splicing-associated histone marks⁴⁰, and binding sites of RNA-binding proteins based on eCLIP data⁴¹. We trained SMS on 80% of sQTLs (randomly selected) and used the remaining data for testing, yielding stable performance across many trials, described in the Supplementary Fig. 8B, C, with model coefficients in Supplementary Fig. 9. When given access only to reference gene annotations, where the distance feature uses GENCODE splice sites (SMS-GENCODE), the method's performance improves moderately over SpliceAI¹⁶, Pangolin³⁴, and Adjusted Motif (AM) architecture models⁴² (Fig. 5B). When using a distance feature based on GTEx splice sites, including those belonging to NVEs, our

"SMS-full" model achieves an area under the precision recall curve (AUPRC) of 0.52, substantially better than SpliceAI or SMS-GENCODE (Fig. 5B). When training on the top splicing signal (top PIP), i.e., the strongest splicing association for each unique variant, we obtained qualitatively similar results and the same ranking of algorithms (AUC = 0.68 for full SMS model, 0.38 for SMS with GENCODE only, 0.26 for spliceAI).

Examining the performance of models using different combinations of features, we observed that distance to nearest GENCODE splice site contributes to precision, analogous to studies of eQTLs where distance to the reference transcription start site (TSS) is highly predictive⁴³, but that use of additional information related to splicing is needed to achieve a high AUPRC (Supplementary Fig. 8D). Together, these observations support that the improved accuracy of SMS-full over the SMS-GENCODE model is driven by inclusion of unannotated splice junctions, such as those of NVEs, and demonstrates the importance of in-sample RNA sequencing to better predict the genetic basis of splicing variation. Overall, we found that ~30% of splicing phenotypes associated with sQTLs are NVEs. We next compared predictions to the best-performing non-SMS model on our data, which was spliceAI. Comparing SMS-full to SpliceAI predictions for individual variants, we observed substantially different scores for the two methods in many cases (Fig. 5C), suggesting the potential complementarity of these models (Supplementary Fig. 8E). One explanation for the large divergence between predictions could be that SpliceAI performs best on rare variants that induce large changes in splicing¹⁶, while SMS-full is trained on data including more subtly spliced NVEs, potentially improving its ability to identify common variants which tend to exert smaller effects.

An example of a variant with low SpliceAI probability (0.25) but high probability with SMS (0.88) is rs78444298, located in the *EDEM3* gene with MAF 1.5%. This variant is associated with schizophrenia, as well as metabolic and blood phenotypes^{44–46}. Molecular studies of this variant have focused on *EDEM3* as a LoF phenotype⁴⁷. Though this is a predicted missense variant causing a proline to serine change, whether LOF occurs at the level of RNA abundance or protein function remains unclear. The variant is associated with an increase in splicing of a low EF nmdNVE in various tissues, such as Brain Cerebellar Hemisphere, and is associated with decreased gene expression in several of those tissues (Fig. 5D). SMS indicates that the variant is in a region of potential binding by several RBPs (HNRNPK, PCBP1, SRSF5, TAF15) near the 3SS of the nmdNVE. The variant alters predicted affinity for two of these RBPs (HNRNPK, PCBP1) by 5-fold and 4-fold, respectively, highlighting a potential molecular mechanism for this variant. The cognate exon and nmdNVE have moderate to high 3SS motif scores (6.7 and 10.4 bits, respectively). In summary, the SMS model and NVE information provide molecular explanations for this variant. SMS scores for all GTEx variants are provided (Supplementary Table 3).



Discussion

To better understand the impact of inter-individual variation in gene and isoform expression, we defined and identified over 50,000 NVEs and characterized their impacts. Many variants identified in GWAS studies lie within noncoding regions, some of which impact pre-mRNA splicing. Here, we found that NVEs tend to occur in more constrained genes and that NVEs enhance our ability to interpret GWAS variants,

especially in the UK BioBank. As RNA-seq data become more available from GWAS participants, in-sample splicing analysis and NVE identification may further improve variant interpretation.

NVEs have the potential to affect gene function or expression, commonly occurring in coding exons, their intervening introns or in 5' UTRs, and less commonly in 3' UTRs. NVEs that occurred at higher frequency in the population were particularly enriched in 5'

Fig. 5 | Splice modifier score (SMS) provides insight into variant induced splicing and helps interpret GWAS. **A** Fraction of NVEs in each EF bin explained via ≥ 1 cis sQTL in GTEx. Gray dot indicates the average across all NVEs. **B** Precision recall curves (PRC) of full logistic regression model trained on GTEx distances, PRC of full logistic regression trained on GENCODE distances only, and PRC of logistic regression trained only on SpliceAI score are shown, as well as Pangolin, and the maximum SAM-FM and SAM-AM scores. **C** Comparison of SpliceAI and SMS scoring. SMS scored using full model parameters and distance to GTEx splice sites across all GTEx variants. Variants with SpliceAI that scored > 0 were selected, and relative SMS and SpliceAI scores were compared in a scatter and kernel density plot. **D** Missense variant in *EDEM3* protein (rs78444298) associated with GWAS studies that may cause an NMD-inducing NVE in GTEx. Individuals with colon

sigmoid tissue samples ($n = 318$) sorted by their estimated Ψ values for this NVE. Mean posterior estimates of Ψ are plotted (in purple if the estimate is at least 5% Ψ), with error bars showing one standard deviation on Ψ . Below are bar plots of exon junction (EJ) reads for both NVE and cognate, and overall observed mRNA levels (TPM). Inset: Details on rs78444298, which induces a proline to serine mutation (shown in diagram) and occurs near a 3' splice site. NMD-inducing mechanism shown: low EF NVE boundaries from an alternative 3' SS with a maxent score of 10 (purple). Two of four RNA binding sites predicted to bind to the region (from oRNament database) are highlighted, showing two RNA binding proteins where the variant substantially reduces affinity. Probabilities of variants causing splicing changes using SpliceAI and SMS predictors shown at right.

UTRs, paralleling the high frequency of evolutionarily more recent exons observed in this gene region⁸. NVEs with lower EFs occurred largely in coding regions, particularly in constrained genes, and typically had lower Ψ values. Natural selection may more strongly limit which NVEs can rise to higher PSI and EF values in coding regions, because of constraints on the expression levels of canonical isoforms in constrained genes. One limitation of our current study is the modest number of available eQTL- and sQTL-associated NVEs with NMD potential (Fig. 4). These NVEs, whose splicing can impact gene expression by producing nonfunctional isoforms, could be targeted with splice-switching antisense oligonucleotides (ASOs) to therapeutically increase (or decrease) protein abundance^{48,49}. Inhibiting the splicing of an NVE would only be useful in people that splice the NVE, of course, but activating the splicing of an NVE via an ASO might be feasible even in individuals where the NVE is not normally included.

Recent efforts have begun to more comprehensively describe the impacts of inter-individual variation on splicing^{11,32,50}. In principle, NVEs might arise from cis-acting mutations that create or disrupt splice sites or splicing regulatory elements, from trans-acting mutations that impact the activity of splicing factors, or from various physiological or lifestyle factors such as sex, age, diet, environmental exposures, etc. In addition, the proportion of cell types present in a tissue sample might vary between individuals, for physiological or technical reasons, potentially contributing to NVE detection. Here, we find strong evidence that a majority of NVEs are impacted by cis-genetic variation (Fig. 5A). Trans-acting variation is also likely to be important but is more difficult to detect. Sex-specific differences in splicing are also known^{51,52}, and widespread splicing differences occur in individuals with certain disorders, including myotonic dystrophy and autism spectrum disorders^{3,53}.

All humans are thought to express similar sets of genes, in similar tissue-specific patterns⁷. However, we found here that individuals commonly differ in the use of specific exons, with each person expressing several hundred NVEs in each well-sampled tissue. Comparing two unrelated individuals, just over half of NVEs are expected to be shared, implying that unrelated people differ in the presence of hundreds of different mRNA isoforms in each of their tissues. Our findings using SMS emphasize the importance of using diverse and population-relevant RNA-seq data for the inference of genetic variation that modulates splicing. The abundance of lower-EF exons and their enrichment in constrained genes, which are more associated with disease, emphasizes the value of increasing population sample sizes rather than simply increasing read depth for increased detection of disease-relevant splicing.

Methods

Ethics statement

No primary data were generated for this study. Person-related data were obtained through authorized access from primary data controllers.

Statistics and reproducibility

No statistical method was used to predetermine sample size. We did not use any study design that requires randomization or blinding. In the GTEx data, no samples were excluded.

Datasets

GTEx release v8. From the GTEx download data portal, we downloaded splicing “phenotype” files, which were originally used as inputs of the original sQTL study in GTEx (v8p hg38). We accessed it using the GTEx data browser, and you can also use the requester pay google cloud bucket (gs://gtex-resources/GTEx_Analysis_v8_QTLs). Protected data, such as genotypes and metadata, were accessed via dbGaP (study accession: phs000424.v8.p2).

GTEx fine-mapped sQTLs and eQTLs. Baberia et al. fine-mapped sQTLs and eQTLs in GTEx using the DAP-G fine-mapping method. We accessed this dataset using their Zenodo link. Note that fine-mapping was performed only on European ancestry samples from GTEx.

gnomAD. We accessed LOEUF scores from gnomAD v4.0 using the gnomAD data browser (linked here), which computed the LOEUF across nearly all genes in the genome.

GENCODE exon annotations. We accessed comprehensive transcript annotations from GENCODE v44 using their data browser, for use as our reference set of exons.

SpliceAI scores. SpliceAI scores have been computed across the genome (all possible SNP mutations). To compare the differences in SMS and spliceAI scores, we used a dataset that released all variants with SpliceAI scores above a nominal threshold (> 0.1), which are available in the Illumina data browser (link: <https://basespace.illumina.com/s/otSPW8hnhazR>). All other variants were set to 0.

Fine-mapped BioBank (GWAS) data

UKBB (96 traits): <https://github.com/mkanai/finemapping-insights>, including *ASGR1* exon 4

Biobank Japan: <https://pheweb.jp/downloads>

FinnGen: https://www.finnngen.fi/en/researchers/data_available

Data preprocessing

GTEx LeafCutter phenotypes. LeafCutter intron clusters were required to include only 2 or 3 introns, so we could classify them as alternative 3' or 5' splice sites or skipped exons, respectively. We mapped to genes and strands by scoring every splice site in the cluster, and picking the strand that had MaxEnt scores > 0 for all splice sites, excluding any clusters that did not score above zero across all splice sites on either strand as potential artefacts. Only protein-coding genes were included. Lastly, exons were required to be ≤ 500 bp in length and alternative splice sites were required to be within 500 bp of each other.

Generating estimates of Ψ values of NVEs

Many NVEs have fairly low inclusion levels, so to estimate their usage accurately, we built a model to assess their abundance in an RNA-sequencing sample. We considered a potential NVE in relation to a more frequently observed “cognate” junction in transcripts from the same gene.

PSI (Ψ) estimates the fraction of a gene’s transcripts that contain the exon or splice site of interest and is a widely used statistic to quantify splicing. It can be estimated from any RNA-seq dataset with sufficient read coverage of the alternative region:

$$\Psi_{\text{NVE}} = \frac{\text{EJ reads}_{\text{NVE}}}{\text{EJ reads}_{\text{NVE}} + \text{EJ reads}_{\text{cognate}}} \quad (1)$$

We consider the exon or splice junction with fewer reads across all individuals a candidate NVE, and the exon or splice junction with more reads as the cognate.

Conditional on Ψ , the number of NVE reads in an RNA-seq dataset can be modeled as a binomial distribution, as defined by LeafCutter splicing³³ clusters:

$$\text{NVE reads} | \Psi_{\text{NVE}} \sim \text{Binomial}(\text{EJ reads}_{\text{NVE}} + \text{EJ reads}_{\text{cognate}}, \Psi_{\text{NVE}}) \quad (2)$$

Note that we have to frame the exon junction reads slightly differently in alternative splice sites as compared to skipped exons to model the NVE (Supp. Fig. 1A).

Mixture of Betas binomial model

Partial pooling involves using information across samples to estimate effect sizes of individual samples by fitting a population distribution to the observed data and letting that inform effect size estimates for individuals^{54,55}.

For proof of concept, the NVE population distribution can be modeled with a single beta distribution with parameters α and β :

$$\Psi_{\text{NVE}} \sim \text{Beta}(\alpha, \beta) \quad (3)$$

Where α can be described as the pseudocount on the splicing of the NVE splicing event, and β can be described as the pseudocount on the splicing of the cognate splicing event, respectively. In other words, in the absence of read count data, α and β provide an initial estimate of the splicing of the NVE.

A single beta distribution can capture several common situations (but not all – see Supplementary Note). One example of a distribution that a single beta cannot handle is a trimodal distribution, which would be expected for an NVE whose splicing differs between three subsets of the population. Because trimodal distributions are interesting biologically, potentially representing cases where a single variant (that may be absent or present in heterozygous or homozygous form) drives the inclusion of an NVE, we decided to fit a three-component mixture of betas model to all exons in the dataset. We compare this result to a single beta model in the Supplementary Note **Sections E-F**, and find that the summary statistic, EF, is fairly robust in both three- and one-component models.

For a three-component beta distribution, with parameters α and β , the NVE population distribution can be modeled with nine parameters:

$$\Psi_{\text{NVE}} \sim \sum_{i=1}^3 \gamma_i \text{Beta}(\alpha_i, \beta_i) \quad (4)$$

Where γ_i are weights that represent the fraction of the distribution coming from each individual beta distribution, requiring that all $\gamma_i \geq 0$,

and that they sum to 1. Note that the triple-beta distribution can model a two-beta or single beta distribution by setting one or two of the γ parameters to zero. We can then estimate the population distribution of a given tissue sample:

$$PDF_{\text{NVE}} = \sum_{i=1}^3 \gamma_i \cdot \text{PDF}(\text{Beta}(\alpha_i, \beta_i)) \quad (5)$$

The task is to predict the population distribution of a splicing event in a given tissue sample. We employ maximum likelihood estimation with the expectation maximization algorithm. The algorithm is initialized with uniform weights, $\alpha = (1, 2, 3)$, and $\beta = (3, 2, 1)$ as an unbiased starting point able to capture symmetric and asymmetric distributions. It then iteratively performs the E-step to analytically find optimal weights γ , and the M-step to determine the α, β using the L-BFGS-B optimization algorithm. The method is represented graphically in Fig. 1A.

Alternative splice site NVEs

We searched for the optimal α, β , and γ parameters that best fit the data, using the expectation maximization framework (see Supplementary Note **Sections B-C**).

Skipped exon NVEs

Skipped exons are represented by three intron clusters: two introns supporting the inclusion of the exon, and one longer intron excluding the exon. We include events where both inclusion introns have numbers of reads that are not drastically different (see Supplementary Fig. 11), keeping one of the two inclusion introns to simplify the analysis. We then followed the same approach as for alternative splice sites to find the optimal parameters.

Calculation of exon frequency (EF) at a given Ψ threshold

We estimated the Ψ of all NVEs of length ≤ 500 (for skipped exons) or inter-splice site distance ≤ 500 bp (see Supp. Fig. 16), as this size captures the vast majority ($>99\%$) of known internal exons and alternative splice sites in the human transcriptome. After obtaining the population parameters for all splicing events, we calculated the EF for a tissue at particular Ψ thresholds by using the population distribution of junction reads across individuals in the tissue. For each of four cutoff values $c = 1\%, 5\%, 10\%$, or 20% , we computed EF as the percentage of individuals with $\Psi \geq c$ from the CDF of the population distribution that was generated from the fitted model, and provide this value in a summary table (Supplementary Table 1). For a tissue with n individuals (almost always > 100), we included events where the EF 5% was between 1% and 99%, excluding poorly spliced exons with EF below 1% and “canonical alternative” exons with EF $> 99\%$.

Following ref. 56, to obtain an estimated Ψ value for an NVE in a single individual, we computed the mean posterior on Ψ and standard deviation of the estimate in each individual in the tissue by using the α and β parameters estimated from the 3-beta mixture binomial fit described above. To obtain a posterior mean on Ψ , a posterior on Ψ is computed using component priors is first estimated, and updated weights using the posterior estimates of α and β parameters are estimated. The weighted sum of these parameters for all components is the posterior mean on Ψ . That is, let:

$$x = \text{EJ}_{\text{NVE}}$$

$$n = \text{EJ}_{\text{NVE}} + \text{EJ}_{\text{cognate}}$$

$$\alpha_i = \text{EJ}_{\text{NVE}} + \alpha_i$$

$$\beta_i = EJ_{\text{cognate}} + \beta_i$$

And we use the following formula to find the updated weight, C_j for each component of the mixture:

$$C_i = \int_0^1 \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \binom{n}{x} \theta^{a_i+x-1} (1-\theta)^{b_i+n-x-1} d\theta \quad (7)$$

Then, we have

$$\left(\widehat{\Psi}_{\text{posterior}} \right)_i = \frac{a_i}{a_i + b_i}$$

$$\left(\widehat{\Psi}_{\text{posterior}} \right)_{\text{nve}} = \frac{1}{\sum_{i=1}^3 Y_i C_i} \sum_{i=1}^3 C_i Y_i \frac{a_i}{a_i + b_i} \quad (8)$$

The variance of each of the components is given by

$$\sigma_i^2 = \frac{a_i b_i}{(a_i + b_i)^2 (a_i + b_i + 1)} \quad (9)$$

Then, we use the law of total variance to get

$$\text{Var}\left(\left(\widehat{\Psi}_{\text{posterior}}\right)_{\text{nve}}\right) = \frac{1}{\sum_{i=1}^3 Y_i C_i} \sum_{i=1}^3 C_i Y_i \left[\sigma_i^2 + \left(\left(\widehat{\Psi}_{\text{posterior}} \right)_i - \left(\widehat{\Psi}_{\text{posterior}} \right)_{\text{nve}} \right)^2 \right] \quad (10)$$

Standard deviations were calculated as the square root of the variances estimated in this manner.

Estimation of total NVEs per individual and number of shared NVEs between individuals

Because EF values estimated from better-sampled tissues are likely more accurate than those estimated from lowly-sampled tissues, we calculated top EF values using only the top 10 most sampled tissues from GTEx in estimating the expected number of NVEs expressed in an individual. In this calculation, and the estimation of the number of NVEs shared between unrelated individuals, we assumed that NVEs occur independently of each other with probability estimated by their EF. Because other NVEs occur exclusively in tissues outside of these 10, these estimates are likely quite conservative. Overall, 23% of the data is included in the top 10 tissues. These tissues are: Whole Blood, Thyroid, Adipose Subcutaneous, Skin (Sun Exposed Lower leg), Nerve (Tibial), Artery (Tibial), Skin (Not Sun Exposed Suprapubic), Esophagus Mucosa, Lung, and Cells (Cultured fibroblasts).

Variant analyses

Fine-mapped sQTLs. As above, we filtered for intron clusters containing two introns (with one shared endpoint), which represent alternative splice sites, and clusters containing three introns in a pattern consistent with presence of a skipped exon. In these cases, any sQTL (i.e., any variant present in the fine-mapping dataset) associated with the intron cluster is considered to be associated with the NVE.

Top sQTL PIP file

We created pan-tissue clusters by merging the locations of each intron in a cluster, assigning pan-tissue cluster IDs. We then sorted by PIP in descending order, and dropped duplicates based on pan-tissue cluster IDs and variant pairs, keeping the highest PIP value. This ensured that

we were considering the top variant-phenotype pair for all splicing events. We performed the analogous analysis with eQTLs, filtering by top gene instead of cluster ID. To compute the percentage of NVEs with *cis* effects, we constructed the sQTL high-confidence set: sQTLs that had a fine-mapped PIP > 90%, and further filtered for variants within 5 kb of the splicing site to obtain, since most splicing regulation is thought to occur within a moderate distance from the splice site⁵⁷.

SMS algorithm

We use fine-mapped GTEx sQTLs and filters for LeafCutter splicing phenotypes that can be categorized as exon skipping or alternative 5' or 3' splice sites, as above (see partial pooling section). For training, we separated out high-PIP sQTLs (90% or above) as a positive class and low-PIP sQTLs (0.2% or below) as a negative class. The sample size is 1.5 million sQTLs across all PIPs, and 250,000 sQTLs after separating out high-PIP and low-PIP sQTLs for training/testing, of which about 4000 are high-PIP.

For features to train the model, we used PhyloP conservation, both 5SS and 3SS MaxEnt splice site scores (by scanning around the variant and looking for the maximum score of each splice site type), change in MaxEnt score due to variant, location within an exonic splicing enhancer³⁸ or exonic splicing silencer³⁹ motif, splicing-associated histone marks⁴⁰, location within the binding site of an RNA-binding protein, based either on eCLIP peak data⁴¹, or based on a mapping of *in vitro*-derived binding motifs in the transcriptome⁵⁸, and the log of the distance to the nearest splice site.

We trained a logistic regression to predict feature weights using an annotated feature matrix M . We found that feature weights changed depending on where the variant was located within the gene. For example, exonic splicing enhancers have stronger weights in annotated exons than in introns. The gene annotations considered are splice sites (including either GENCODE splice sites or GTEx SS in SMS-full), GENCODE annotated exon, or GENCODE annotated intron. So, we also include a gene annotation vector A . The vector contains the location of the exon within the gene. For a given variant i ,

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + (MA) \quad (11)$$

where p_i is the probability that the variant i modifies the splicing outcome. The gene annotation vector contains the following labels: not splice site, in any exon, both GTEx and reference splice region, just GTEx splice region, just reference splice region, reference exon, and reference intron.

Any features related to splice site strength, such as maxent score, were assigned separate weights based on being outside or inside the splice site ("not splice site" gene annotation). Any features using distance based on being outside or inside an exon ("in any exon" gene annotation). Feature weights across different gene annotations are shown for SMS-full in Supplementary Fig. 9C, D.

We trained this logistic regression model on 80% of randomly selected sQTLs to determine the relative importance of each feature in predicting causal sQTL variants. We trained on both the full model and on subsets of the features, and computed statistics such as AUPRC of these sets of models using held-out test sets, which is reported in the text. The GitHub repository includes all annotations for GTEx variants and a tutorial for rerunning many combinations of features for this model.

Comparison of other methods with SMS

To compare SMS with spliceAI, we also trained a logistic regression model using SpliceAI scores downloaded from [basespace](https://basespace.org). (Note that on basespace, all SpliceAI scores below 0.1 are listed as 0.) For Pangolin models, the model range included the entire context window, and for SAM models we include 5400 nt of context around each location. For

SAM-FM and -AM models we took 20% of all sQTLs in the dataset which had PIP values of $<0.2\%$ or $>90\%$, then located them in the hg38 genome. We then collected a context around them of model context width +201. We then ran the models on these sequences (both before and after the mutation), keeping the 201 middle predictions (those where the models had the full context). We then compute the largest absolute value of the difference in prediction in this range, which is the reported result.

Distribution of loss-of-function intolerant genes

The loss-of-function observed/expected upper bound fraction (LOEUF) scores, which was first described in ref. 23, were obtained from the [gnomAD browser](#). We used the most recent release (v4).

Enrichment of GWAS variants in splice sites across biobanks

First, associations of variants across all traits were concatenated into a file of trait-variant pairs. We filtered for single variant-trait pairs, taking the top PIP across all traits (so as to not repeat variants that had high PIP associations across many traits). Then for each variant, we computed the maximum PIP across traits in BBJ, FinnGen, and UKBB, and pooled these variants together. We estimated functional enrichment for each category as a relative risk (RR, i.e., a ratio of proportion of variants) between being in an annotation and fine-mapped. That is, $RR = (\text{proportion of variants in annotation with PIP within a specific bin}) / (\text{proportion of variants in annotation with PIP} \leq 0.01)$. We used PIP bins with the following boundaries: [0, 0.01, 0.01, 0.25, 0.5, 0.75, 1.0], where all of the intermediate values are open brackets.

To obtain enrichments of GWAS variants for unannotated GTEx splice sites, we filtered out any splice sites in GENCODE, and variants in pLoF, splice region, or missense variants. Enrichment is calculated using a fraction of variants observed in the lowest PIP bin (see range on plot) relative to the bin in question. Enrichment for missense and synonymous changes were also computed, keeping pLoF and missense variants in that set. We provide the locations of all unannotated GTEx splice site locations in Supplementary Data 4. Because unannotated GTEx splice sites comprise a very small fraction of the genome and tend to occur close to existing coding regions, we did not use linkage disequilibrium score regression (LDSC)⁵⁹ to assess whether unannotated GTEx contributed to SNP-heritability of traits.

Detection of unannotated splice sites impacting NMD

We devised a script that called unannotated splice sites with NMD potential in GTEx LeafCutter splicing phenotypes. The script: (1) intersects LeafCutter introns with annotated UTRs, excluding those that overlap; (2) maps remaining LeafCutter introns to corresponding exons in annotated coding regions; (3) Assesses putative NMD potential, if either (a) the NVE alters the frame of an annotated CDS, e.g., adding an exon whose length is not a multiple of three, or (b) the NVE contains an in-frame stop codon; (4) Excludes any NVE that impacts the last exon within the CDS; and (5) Excludes any NVE that represents the second-to-last exon of a CDS whose only stop codon(s) are within 50 bp from the NVE 5'SS. This script is intended to cover common NMD-triggering splicing changes, but does not capture some special cases. For example, if a pair of adjacent NVEs in the same gene each individually preserve frame they will be considered frame-preserving, even if when spliced together they generate a stop codon at the exon-exon junction. Conversely, if two NVEs each alter the reading frame, they will be considered as having NMD potential even if when they are spliced together the reading frame is restored.

sQTLs that were also eQTLs

First, GTEx variants were filtered for those that were both nominally significant sQTLs (v8.sqtl_signifpairs.txt) and nominally significant eQTLs (v8.egenes.txt) in the same tissue ([GTEx browser](#)). The sQTL

phenotypes considered only included splicing events that were alternative splice sites, where one of the sites was an NVE. The exact sQTL phenotype was the cognate intron, and not the NVE, because not many variants were nominally significant in the other direction (impacting the NVE levels), likely due to lower Ψ values of NVEs, so we transformed the effect size by -1 for all sQTLs, in order to more directly compare the impact of the NVE inclusion on expression of the gene.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data generated in this study have been deposited in the zenodo database under accession code <https://zenodo.org/records/15790343>. The individual level GTEx data is available under restricted access for privacy reasons, access can be obtained by obtained permission via dbGAP: accession number phs000424.v8.p2. The raw data used to fit generate EF estimates are available at GTEx data browser: <https://www.gtexportal.org/home/downloads/adult-gtex/ctl>. The pre-processed fine-mapped sQTLs of GTEx samples are available in the zenodo database <https://zenodo.org/records/3517189>. The gencode annotations are available at Gencode, <https://www.gencodegenes.org/releases/44.html>. Source data are provided in this paper.

Code availability

All code is available under MIT License. Code for running the mixture of betas is on Github: url: https://github.com/atgu/mixture_betas zenodo release: 10.5281/zenodo.16790816 including some test data from GTEx. The rest of the data to run all other files and generate the EFs for all tissues can be accessed in the [GTEx data browser](#). Code used to post-process these events is on GitHub. https://github.com/jacobs-hannah-mit/post_process_POVS (10.5281/zenodo.16790828)⁶⁰ and can be used to generate the EFs of NVEs reported in Supplementary Data 1 & 2. Code for logistic regression is also available on GitHub, called splice modifier score (called SMS) url: <https://github.com/jacobs-hannah-mit/SMS>, zenodo release <https://doi.org/10.5281/zenodo.16790834>, including a Jupyter notebook which can be used to retrain the model using different features, or regenerating SMS full model predictions (Supplementary Data 3).

References

- Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
- Martinez, N. M. et al. Alternative splicing networks regulated by signaling in human T cells. *RNA* **18**, 1029–1040 (2012).
- Irimia, M. et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511–1523 (2014).
- Hasimbegovic, E. et al. Alternative splicing in cardiovascular disease—A survey of recent findings. *Genes* **12**, <https://doi.org/10.3390/genes12091457> (2021).
- Ren, P. et al. Alternative splicing: a new cause and potential therapeutic target in autoimmune disease. *Front. Immunol.* **12**, 713540 (2021).
- Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T. & Burge, C. B. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc. Natl. Acad. Sci. USA* **102**, 2850–2855 (2005).
- Barbosa-Morais, N. L. et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).
- Merkin, J. J., Chen, P., Alexis, M. S., Hautaniemi, S. K. & Burge, C. B. Origins and impacts of new mammalian exons. *Cell Rep.* **10**, 1992–2005 (2015).

9. Mazin, P. V., Khaitovich, P., Cardoso-Moreira, M. & Kaessmann, H. Alternative splicing during mammalian organ development. *Nat. Genet.* **53**, 925–934 (2021).
10. Fair, B. et al. Global impact of unproductive splicing on human gene expression. *Nat. Genet.* **56**, 1851–1861 (2024).
11. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
12. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.* **12**, 727 (2021).
13. Kim, J. inkuk. et al. Patient-customized oligonucleotide therapy for a rare genetic disease. *N. Engl. J. Med.* **381**, 1644–1652 (2019).
14. Hoffman, G. E. et al. CommonMind consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Sci. Data* **6**, 180 (2019).
15. Wilks, C. et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* **22**, 323 (2021).
16. Jaganathan, K. et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548 (2019).
17. Wagner, N. et al. Aberrant splicing prediction across human tissues. *Nat. Genet.* **55**, 861–870 (2023).
18. Dawes, R. et al. SpliceVault predicts the precise nature of variant-associated mis-splicing. *Nat. Genet.* **55**, 324–332 (2023).
19. Bénitière, F., Necsulea, A. & Duret, L. Random genetic drift sets an upper limit on mRNA splicing accuracy in metazoans. *Elife* **13**, <https://doi.org/10.7554/elife.93629> (2024).
20. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
21. Hinzpeter, A. et al. Alternative splicing at a NAGNAG acceptor site as a novel phenotype modifier. *PLoS Genet.* **6**, <https://doi.org/10.1371/journal.pgen.1001153> (2010).
22. Fiszbein, A. et al. Widespread occurrence of hybrid internal-terminal exons in human transcriptomes. *Sci. Adv.* **8**, eabk1752 (2022).
23. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
24. Bhat, P. et al. Genome organization around nuclear speckles drives mRNA splicing efficiency. *Nature* **629**, 1165–1173 (2024).
25. Kanai, M. et al. Insights from complex trait fine-mapping across diverse populations. Preprint at <https://doi.org/10.1101/2021.09.03.21262975> (2021).
26. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
27. Nioi, P. et al. Variant ASGR1 associated with a reduced risk of coronary artery disease. *N. Engl. J. Med.* **374**, 2131–2141 (2016).
28. Ali, L. et al. Common gene variants in ASGR1 gene locus associate with reduced cardiovascular risk in absence of pleiotropic effects. *Atherosclerosis* **306**, 15–21 (2020).
29. Kolakada, D. et al. A system of reporters for comparative investigation of EJC-independent and EJC-enhanced nonsense-mediated mRNA decay. *Nucleic Acids Res.* **52**, e34 (2024).
30. Fiszbein, A., Krick, K. S., Begg, B. E. & Burge, C. B. Exon-mediated activation of transcription starts. *Cell* **179**, 1551–1565 (2019).
31. Botto, A. E. C. et al. Reciprocal regulation between alternative splicing and the DNA damage response. *Genet. Mol. Biol.* **43**, e20190111 (2020).
32. García-Pérez, R. et al. The landscape of expression and alternative splicing variation across human traits. *Cell Genom.* **3**, 100244 (2023).
33. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
34. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biol.* **23**, 103 (2022).
35. Cheng, J. et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* **20**, 48 (2019).
36. Wang, Q. S. et al. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat. Commun.* **12**, 3394 (2021).
37. Barbeira, A. N. et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 1–24 (2021).
38. Fairbrother, W. G. et al. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.* **32**, W187–W190 (2004).
39. Wang, Z. et al. Systematic identification and analysis of exonic splicing silencers. *Cell* **119**, 831–845 (2004).
40. ENCODE Project Consortium et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
41. Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
42. Gupta, K. et al. Improved modeling of RNA-binding protein motifs in an interpretable neural model of RNA splicing. *Genome Biol.* **25**, 23 (2024).
43. Westra, H.-J. & Franke, L. From genome to function by studying eQTLs. *Biochim. Biophys. Acta* **1842**, 1896–1902 (2014).
44. Wuttke, M. et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).
45. Trubetskoy, V. et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
46. Teumer, A. et al. Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria. *Nat. Commun.* **10**, 4130 (2019).
47. Xu, Y.-X. et al. EDEM3 Modulates Plasma Triglyceride Level through Its Regulation of LRP1 Expression. *iScience* **23**, 100973 (2020).
48. Neil, E. E. & Bisaccia, E. K. Nusinersen: A Novel Antisense Oligonucleotide for the Treatment of Spinal Muscular Atrophy. *J. Pediatr. Pharmacol. Ther.* **24**, 194–203 (2019).
49. Lim, K. H. et al. Antisense oligonucleotide modulation of non-productive alternative splicing upregulates gene expression. *Nat. Commun.* **11**, 3501 (2020).
50. Walker, R. L. et al. Genetic control of expression and splicing in developing human brain informs disease mechanisms. *Cell* **179**, 750–771 (2019).
51. Trabzuni, D. et al. Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.* **4**, 2771 (2013).
52. Blekhman, R., Marioni, J. C., Zumbo, P., Stephens, M. & Gilad, Y. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* **20**, 180–189 (2010).
53. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2016).
54. Carpenter, B. *Hierarchical Partial Pooling for Repeated Binary Trials* <https://mc-stan.org/users/documentation/case-studies/pool-binary-trials.html> (2016).
55. Young-Xu, Y. & Chan, K. A. Pooling overdispersed binomial data to estimate event rate. *BMC Med. Res. Methodol.* **8**, 58 (2008).
56. Farrow, M. *MAS3301 Bayesian Statistics, School Mathematics Statistics*. (2008).
57. Wainberg, M., Alipanahi, B. & Frey, B. Does conservation account for splicing patterns? *BMC Genomics* **17**, 787 (2016).
58. Benoit Bouvrette, L. P., Bovaird, S., Blanchette, M. & Lécuyer, E. oRNAment: a database of putative RNA binding protein target sites in the transcriptomes of model species. *Nucleic Acids Res.* **48**, D166–D173 (2020).
59. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

60. Jacobs, H. Widespread naturally variable human exons aid genetic interpretation. https://github.com/jacobs-hannah-mit/post_process_POVS 10.5281/zenodo.16790828 (2025).

Acknowledgements

We thank Michael McGurk for guidance on analyses, feedback on the early manuscript, as well as beta-binomial model framework. We thank Jacob Ulirsch and Masahiro Kanai for discussions of the *ASGR1* exon. We also thank Ran Cui for providing guidance on fine-mapping based analyses. We thank Chi-Hsien Chang for help testing the code. We thank Athma Pai, Kaitlin Samocha, and David McWatters for direct feedback on the manuscript. This work is supported by the Novo Nordisk Foundation (NNF21SA0072102) and by NIH grants HG002439 and GM085319 (to C.B.B.).

Author contributions

H.N.J. performed the analyses and authored the manuscript. B.L.G. helped implement the mixture of beta-binomial analyses and provided feedback on the manuscript. J.G. helped with construction of NMD analyses with guidance from H.N.J. K.G. helped with model comparisons with SMS. M.K. provided guidance on GWAS results. H.K.F., K.J.K., and C.B. provided guidance on analyses, feedback on manuscript, and helped guide project directions.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-65476-7>.

Correspondence and requests for materials should be addressed to Christopher B. Burge.

Peer review information *Nature Communications* thanks Marta Melé and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025