nature genetics

Technical Report

https://doi.org/10.1038/s41588-023-01373-3

Aberrant splicing prediction across human tissues

Received: 5 April 2022

Accepted: 14 March 2023

Published online: 4 May 2023

Check for updates

Nils Wagner [1,2,7], Muhammed H. Çelik[1,3,7], Florian R. Hölzlwimmer[1], Christian Mertes [1,4], Holger Prokisch [5,6], Vicente A. Yépez [1] & Julien Gagneur [1,2,5,6]

Aberrant splicing is a major cause of genetic disorders but its direct detection in transcriptomes is limited to clinically accessible tissues such as skin or body fluids. While DNA-based machine learning models can prioritize rare variants for affecting splicing, their performance in predicting tissue-specific aberrant splicing remains unassessed. Here we generated an aberrant splicing benchmark dataset, spanning over 8.8 million rare variants in 49 human tissues from the Genotype-Tissue Expression (GTEx) dataset. At 20% recall, state-of-the-art DNA-based models achieve maximum 12% precision. By mapping and quantifying tissue-specific splice site usage transcriptome-wide and modeling isoform competition, we increased precision by threefold at the same recall. Integrating RNA-sequencing data of clinically accessible tissues into our model, AbSplice, brought precision to 60%. These results, replicated in two independent cohorts, substantially contribute to noncoding loss-of-function variant identification and to genetic diagnostics design and analytics.

Identifying noncoding loss-of-function DNA variants is a major bottleneck of whole genome interpretation, as predicting function outside coding regions is difficult[1]. Variants altering splicing represent an important class of noncoding loss-of-function variants because they can lead to drastically altered RNA isoforms, for instance, by inducing frameshifts or ablations of functionally important protein domains. If the variant strongly alters splicing isoform choice, the remaining abundance of functional RNA isoforms can be so reduced that the function of the gene is lost. Due to the relevance of splicing for variant interpretation, notably in rare disease diagnostics and in oncology, algorithms have been developed to predict whether variants affect splicing[2–9]. However, only recently, aberrant splicing events, that is, rare large alterations of splice isoform usage, have been called in human tissues[10–12]. While a method to a posteriori prioritize candidate causal rare variants for observed aberrant splicing events has been proposed[12], the forward problem, that is, predicting among rare variants which ones will result in aberrant splicing, has not been addressed.

Here, we set out to establish models predicting whether a rare variant associates with aberrant splicing in any given human tissue. First, we assumed only DNA to be available and later on further considered complementary RNA-sequencing (RNA-seq) data of clinically accessible tissues (CATs) (Fig. 1).

Results

A benchmark dataset for aberrant splicing predictions

We created a benchmark using the aberrant splicing caller FRASER (Find RAre Splicing Events in RNA-seq)[10] on 16,213 RNA-seq samples of the Genotype-Tissue Expression (GTEx) dataset, spanning 49 tissues and 946 individuals. Compared with other splicing outlier detection methods[11,12], FRASER consistently showed the highest agreement with sequence-based predictors and was therefore subsequently used for

[1]School of Computation, Information and Technology, Technical University of Munich, Garching, Germany. [2]Helmholtz Association – Munich School for Data Science (MUDS), Munich, Germany. [3]Center for Complex Biological Systems, University of California, Irvine, Irvine, CA, USA. [4]Munich Data Science Institute, Technical University of Munich, Garching, Germany. [5]Institute of Human Genetics, School of Medicine, Technical University of Munich, Munich, Germany. [6]Computational Health Center, Helmholtz Center Munich, Neuherberg, Germany. [7]These authors contributed equally: Nils Wagner, Muhammed H. Çelik. e-mail: gagneur@in.tum.de

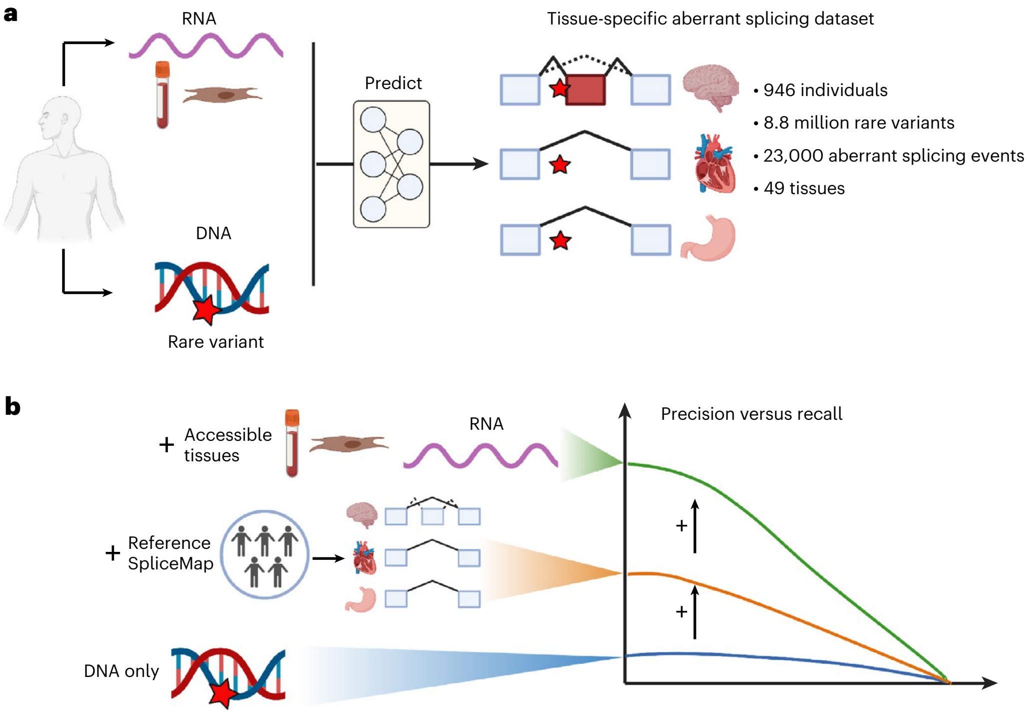


Fig. 1 | Study design and main findings. We set out to predict whether rare variants associate with aberrant splicing across 49 human tissues. **a**, We established a comprehensive benchmark for aberrant splicing by processing GTEx samples with a recently published aberrant splicing caller[10] based on which we could assess and develop predictors that could take as input DNA sequence and, optionally, RNA-seq data of CATs. **b**, Benchmarking revealed modest performance of currently used algorithms based on DNA only, a substantial performance improvement when integrating these models with SpliceMap, a quantitative map of tissue-specific splicing we developed in this study, and further improvements when also including direct measures of aberrant splicing in accessible tissues.

our evaluations (Extended Data Fig. 1). For every individual, we considered every protein coding gene carrying at least one rare variant (minor allele frequency (MAF) less than 0.1% based on the Genome Aggregation Database (gnomAD)[13] and found in no more than two individuals across GTEx) and set out to predict in which tissue, if any, is this gene aberrantly spliced. We defined a gene to be aberrantly spliced in a sample if it was called as a transcriptome-wide significant splicing outlier and with a sufficient amplitude (differential percent spliced-in (Ψ) larger than 0.3; Methods, and see Extended Data Fig. 1 for results with alternative cutoffs). Previous studies had reported that as many as 75% of aberrant splicing events in GTEx RNA-seq samples are not replicated across tissues[10,12] and thus may reflect technical artifacts or aberrant splicing that is not genetically driven. We quantified the enrichment of replicated splicing outliers across tissues of the same individual with respect to the distance to the closest rare variant and found them to be enriched up to a distance of 250 base pairs (bp) (Extended Data Fig. 2). Therefore, we also required a rare variant to be less than 250 bp away from the boundaries of any intron associated with the aberrantly spliced splice site (Methods and Extended Data Fig. 3). This filter yielded similar results as filtering for replicated aberrant events with the extra advantage of being applicable to independent cohorts that have a single sample per individual (Extended Data Fig. 4).

State-of-the art sequence-based models poorly predict tissue-specific aberrant splicing

We then assessed the performance of two complementary state-of-the-art sequence-based deep learning models: modular modeling of splicing (MMSplice)[3], which predicts quantitative usage changes of predefined splice sites within a 100-bp window of a variant, and SpliceAI[2], which is independent of gene annotations and predicts creation or loss of splice sites within a 50-bp window of a variant (Extended Data Fig. 5). Using larger prediction window sizes for SpliceAI did not improve the results (Supplementary Fig. 1). For individuals with multiple rare variants on a gene, we retained the highest score of each model. Out-of-the-box application of MMSplice and SpliceAI showed a modest performance, with an overall precision of 8% for MMSplice and of 12% for SpliceAI at 20% recall, and an area under the precision–recall curve (auPRC) of 4% ± 1 percentage point across tissues for MMSplice and 5% ± 2 percentage points for SpliceAI.

Tissue-specific splicing annotations improve aberrant splicing predictions

We observed that many false predictions originated from inaccurate genome annotations. On the one hand, standard genome annotations are not tissue-specific, leading to false positive predictions. This includes predictions for genes that are not expressed in the tissue of interest, as for the gene *TRPC6* in the brain (Fig. 2a), and, among expressed genes, predictions for exons that are not canonically used in the tissue, as for exon 2 of *C2orf74* in the tibial nerve (Fig. 2b). On the other hand, many splice sites are missing from standard genome annotations[14,15]. These nonannotated splice sites are often spliced at a low level, yet can be strongly enhanced by variants (see Fig. 2c for an example) and are suspected to be a major cause of aberrant splicing[16,17]. To address all these issues, we created a tissue-specific splice site map, which we named SpliceMap, using GTEx RNA-seq data. SpliceMap excludes untranscribed splice sites and introns for each tissue and includes nonannotated splice sites and introns reproducibly observed among samples of the same tissue (Methods). The standard genome annotation GENCODE[18] (release 38 of hg38) contains 244,189 donor sites and 235,654 acceptor sites, of which 93% were detected at least in one GTEx tissue (Fig. 2d). SpliceMap contains 168,004 ± 9,288 donor sites and 164,702 ± 8,950 acceptor sites per tissue (Extended Data Fig. 6).

From this total, 7,060 ± 3,706 donor sites and 8,222 ± 3,740 acceptor sites were unannotated, with testis containing the maximum number of nonannotated donor and acceptor sites (29,673 and 29,911 respectively), in line with the unique transcriptional and splicing patterns of testis[19,20]. SpliceMap is robust to variations in sample size and to different split-read counting tools[21,22] (Supplementary Fig. 2). Moreover, we found that currently available long-read RNA-seq data in GTEx[23] were not yet sensitive enough[24] to reliably identify nonannotated splice sites (Supplementary Fig. 2). Applying MMSplice on the tissue-specific splice sites defined by SpliceMap increased the precision of MMSplice to 13% at 20% recall (Fig. 2e), with a significantly higher auPRC consistently across tissues (Fig. 2f). Similarly, applying SpliceMap on SpliceAI increased precision to 22% at 20% recall.

Quantified reference isoform proportions improve aberrant splicing predictions

Variants affecting splicing typically associate with abundance ratio fold-changes of competing splicing isoforms, which result in nonlinear effects on isoform proportions according to the so-called scaling law of splicing[25,26]. For instance, starting from a 1:1 ratio between one splicing isoform and its alternative in a major allele background, a tenfold decrease leads to a 1:10 ratio, which amounts to around 40 percentage points decrease (from 50% to approximately 10%). However, the same ratio fold-change starting from a 1:10 ratio amounts to less than 1 percentage point decrease (Extended Data Fig. 7). Hence, the scaling law of splicing implies that the variation of isoform abundance between tissues in major allele background alone can explain some of the tissue-specific effects of variants on isoform proportion[25], as exemplified with exon 7 of the gene *TRPC6* (Fig. 3a). We estimated major allele background levels of alternative donor and acceptor splice site usage proportions for all introns and all tissues of SpliceMap (Extended Data Fig. 7). Integrating these reference levels further improved the MMSplice predictions by 1.6-fold consistently across tissues, and to a lesser extent the SpliceAI predictions (Fig. 3b,c and Methods). We suspect that MMSplice showed stronger relative improvement compared with SpliceAI because it models percent spliced-in of predefined splice sites and can integrate in a principled fashion reference levels using the scaling law. In contrast, SpliceAI models creation or loss of splice sites. We integrated reference levels with SpliceAI by applying filters (Methods). However, predicted activations of annotated splice sites and predicted deactivations of unannotated splice sites are already masked in SpliceAI, thereby qualitatively capturing the effect of using reference level filters for a large number of splice sites.

AbSplice-DNA predicts the probability that a variant causes aberrant splicing in a given tissue

Next, to leverage the complementarity of MMSplice and SpliceAI predictions[7], we trained a generalized additive model using the scores from both deep learning models as well as annotation features from tissue-specific SpliceMaps (Methods). This model, which we call AbSplice-DNA, achieved an additional 1.5-fold improvement (Fig. 3b,c). The AbSplice-DNA scores are probability estimates which we found to be well calibrated on GTEx (Extended Data Fig. 8). AbSplice predicts for each variant how likely aberrant splicing of some sort takes place in a given tissue and reports the splice site with the strongest effect (see Supplementary Table 1 for an example). To ease downstream applications we suggest three cutoffs (high: 0.2, medium: 0.05, low: 0.01), which have approximately the same recalls as the high, medium and low cutoffs of SpliceAI (Fig. 3b).

We also tested integration of other predictors into AbSplice-DNA by including scores from Combined Annotation Dependent Depletion-Splice (CADD-Splice)[7], Multi-tissue Splicing (MTSplice)[9] and Super Quick Information-content Random-forest Learning of Splice variants (SQUIRLS)[8] (Methods). However, those models only led to minor improvements (Extended Data Fig. 9). We decided to incorporate only MMSplice and SpliceAI into the final model so as not to have a model confounded by conservation information (used by CADD-Splice and SQUIRLS), and to keep the possibility to easily integrate new tissues which would not be the case with MTSplice. Nevertheless, the code of AbSplice can easily be modified to incorporate new features. We also tried random forest and logistic regression as alternative machine learning models, which gave similar performances to the generalized additive model (Methods and Extended Data Fig. 9).

We evaluated the model performances in more detail by stratifying the results on two different scenarios. First, we stratified by variant categories. As expected, the precision was the best on variants affecting the donor and acceptor dinucleotides on all models, followed by variants in the splice region (within 1–3 bases of the exon or 3–8 bases of the intron), then in the exonic, and lastly in the intronic regions (Methods and Fig. 3d). AbSplice-DNA outperformed all models throughout all variant categories, including intronic variants, whose effects are notoriously more difficult to predict. Second, we analyzed the model performance for five nonexclusive aberrant splicing outcomes: exon elongation, exon truncation, exon skipping, any alternative donor or acceptor choice outlier, and any splicing efficiency outlier. AbSplice-DNA performed better for exon skipping than for exon elongation and truncation, as well as better for alternative donor or acceptor choice than for splicing efficiency outliers. Moreover, AbSplice-DNA outperformed all other models throughout all investigated outlier outcome categories (Fig. 3e).

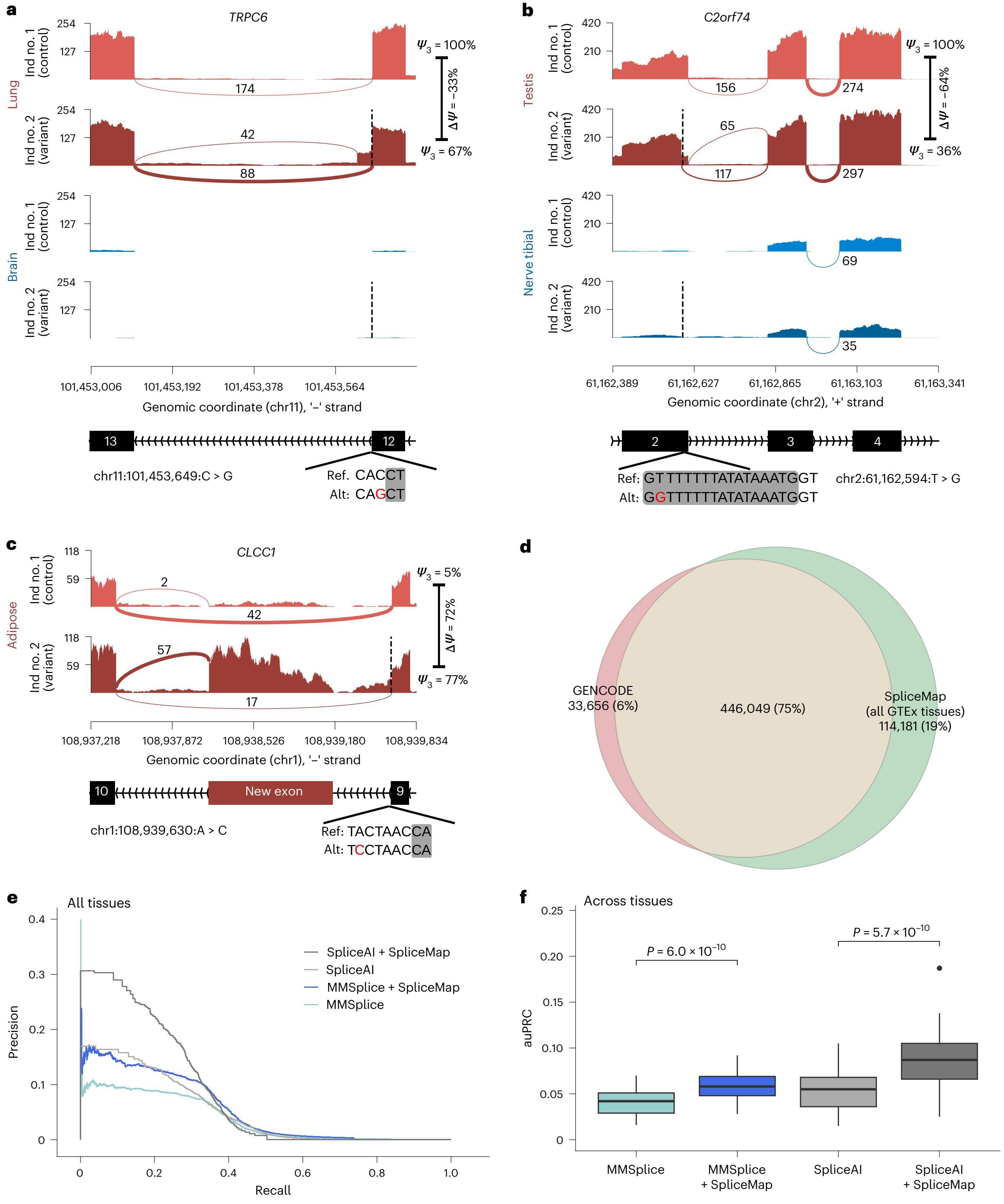
AbSplice-DNA performance is confirmed on independent data

Having established our model on GTEx, we next assessed how well the performance replicated in independent cohorts. We first evaluated a dataset consisting of RNA-seq samples from skin fibroblasts of 303 individuals with a suspected rare mitochondriopathy[27]. We found that there was a large overlap (86%) of splice sites in SpliceMaps generated from GTEx fibroblasts and from this cohort (Fig. 4a and Supplementary Fig. 3). Moreover, we observed consistent reference levels of splicing between the two datasets (Fig. 4b, Pearson correlation 0.87). We applied AbSplice-DNA trained on GTEx using the SpliceMap from GTEx fibroblasts on the subset of this data for which whole genome sequencing (WGS) was available ($n = 20$) and used aberrant splicing calls performed on the RNA-seq samples to assess the predictions. The relative improvements between the baseline models and AbSplice-DNA replicated. AbSplice-DNA achieved 13.2 ± 1.5% auPRC, 2.5-fold higher

Fig. 2 | Tissue-specific splice site map improves prediction performance. **a**–**c**, Sashimi plots showing RNA-seq read coverage (*y* axis) and the numbers of split reads spanning an intron indicated on the exon-connecting line (using pysashimi[50]) for instances illustrating the benefits of the SpliceMap annotation. For each instance, two individuals are displayed. The individual with the rare genetic variant (located at the dashed black line) is shown in the lower track (darker color). SpliceMap catalogs expressed genes and splice sites in each tissue and can thus help in identifying cases for which there is no variant effect in tissues not expressing the whole gene (**a**) or the exon (**b**) in proximity of the variant. Moreover, SpliceMap includes weak splice sites, which are spliced at a low level, but can be activated and create novel exons in the presence of a variant (**c**). **d**, Venn diagram comparing annotated splice sites in standard genome annotation (GENCODE release 38) and SpliceMap aggregating all GTEx tissues. **e**, Precision–recall curves comparing the overall prediction performance across all GTEx tissues ($n = 49$) of MMSplice applied to GENCODE splice sites, MMSplice applied to tissue-specific splice sites according to SpliceMap, SpliceAI and SpliceAI using tissue-specific SpliceMaps. **f**, Distribution of the auPRC across all GTEx tissues of the models in **e**. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles. *P* values were computed using the paired one-sided Wilcoxon test. Alt, alternative; Ind, individual; Ref, reference.

than SpliceAI or MMSplice alone (Fig. 4c). From a rare variant prioritization standpoint, AbSplice-DNA typically gave about twofold fewer candidate predictions at the same level of recall than SpliceAI, itself comparing favorably over MMSplice (Supplementary Fig. 4). Hence, AbSplice-DNA can help rare disease diagnostics by providing substantially shorter lists of predicted candidate variants to investigate compared with state-of-the-art sequence-based models.

We next considered a cohort of WGS samples paired with RNA-seq and proteomics data of induced pluripotent stem cell (iPSC)-derived spinal motor neurons from 245 amyotrophic lateral sclerosis



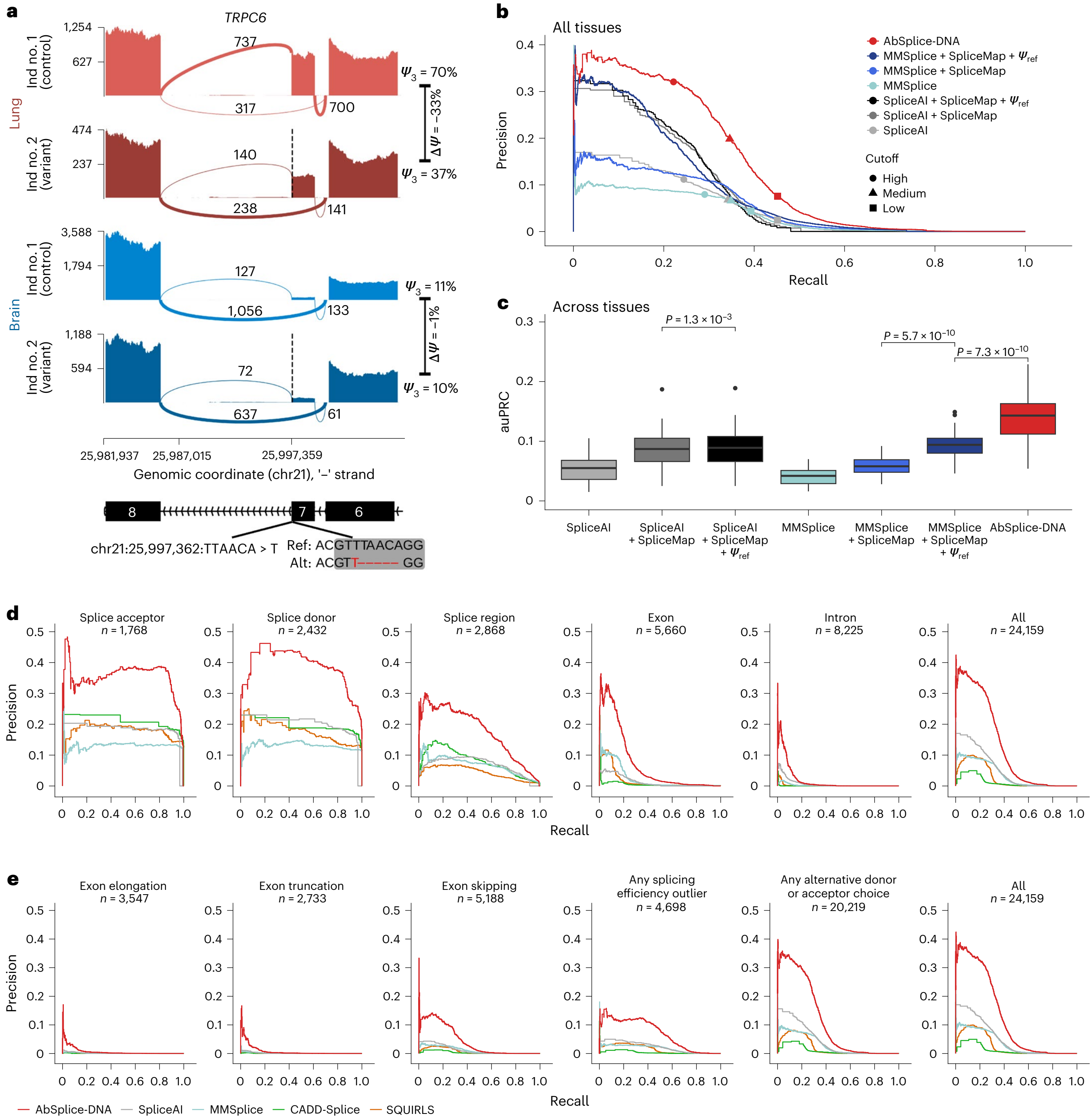


Fig. 3 | Quantitative splicing levels further improve prediction performance. **a**, Sashimi plot of *TRPC6* around exon 7 in lung and brain for two individuals, one carrying no rare variant in this region (control, upper tracks), and one carrying an exonic rare deletion (dashed line and lower tracks) associated with reduced splicing of exon 7. The donor sites of exon 6 and exon 7 compete against each other for splicing with the acceptor site of exon 8. For the control individual, the donor site of exon 7 is used 70% of the time in the lung, and only 11% of the time in the brain. The variant associates with a stronger difference (33 percentage points) in the lung than in the brain (1 percentage point). **b**, Precision–recall curve comparing the overall prediction performance on all GTEx tissues of SpliceAI, SpliceAI using SpliceMap, SpliceAI using SpliceMap along with quantitative reference levels of splicing, MMSplice using GENCODE annotation, MMSplice using SpliceMap annotation, MMSplice using SpliceMap annotation along with quantitative reference levels of splicing and the integrative model AbSplice-DNA. Different cutoffs are shown (SpliceAI, high: 0.8, medium: 0.5, low: 0.2; MMSplice (score absolute value), high: 2, medium: 1.5, low: 1; AbSplice-DNA, high: 0.2, medium: 0.05, low: 0.01). **c**, Distribution of the auPRC of the models in **b** across tissues ($n = 49$). Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles. *P* values were computed using the paired one-sided Wilcoxon test. **d**, Model performance across different VEP[51] variant categories. Categories are ordered from left to right by decreasing severity. Each annotated variant is labeled by its most severe category. The 'Exon' category consists of the VEP categories stop gained, stop lost, missense and synonymous. **e**, Model performance across nonexclusive outlier outcome categories (Methods). For panels **d** and **e**, the 'All' category contains all unique variants (independent of the VEP annotation and outlier outcome categories) and *n* is the number of variants associated with outliers.

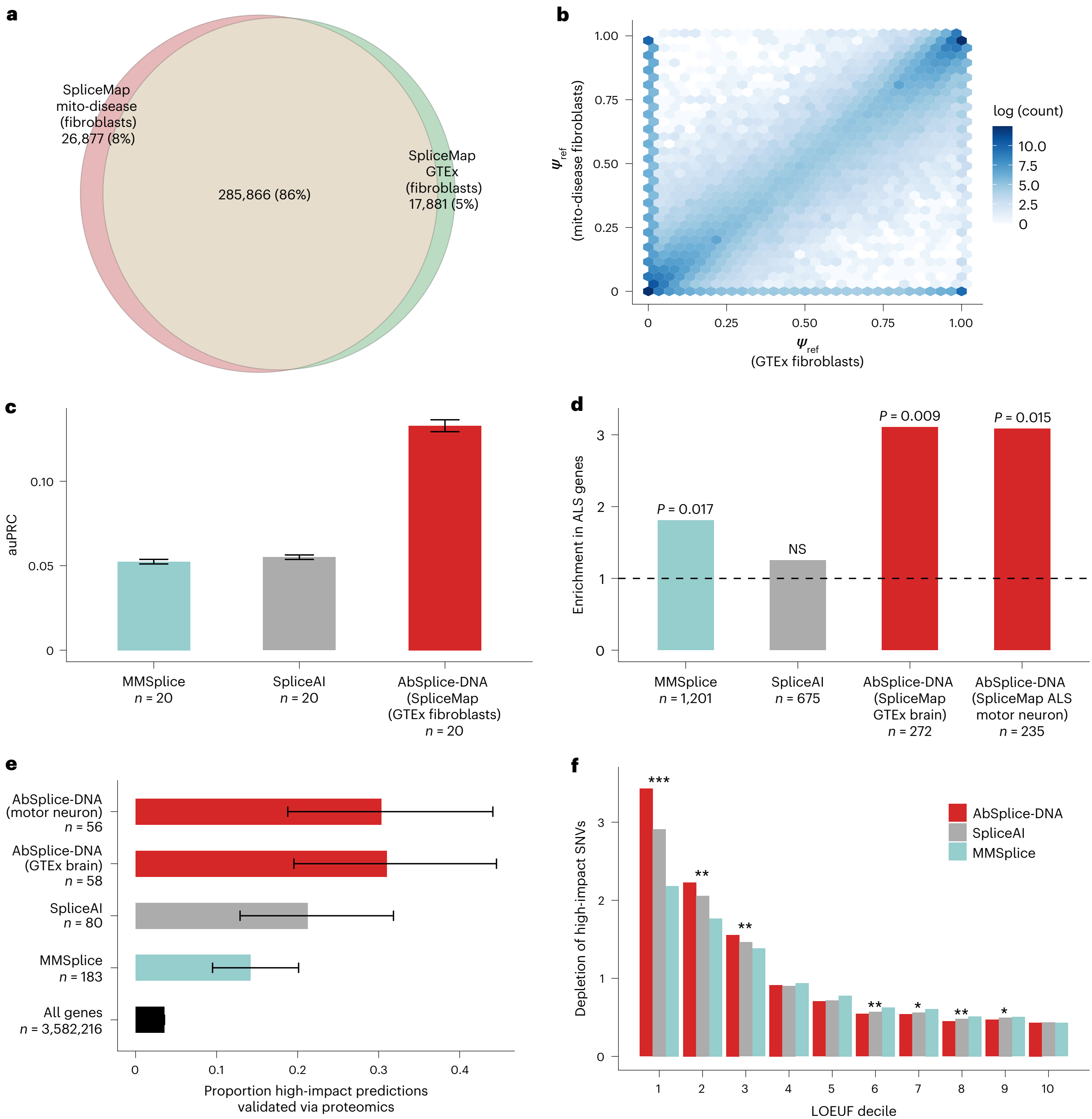


Fig. 4 | Application of AbSplice-DNA on independent data. a, Venn diagram comparing the splice sites in the SpliceMap generated from fibroblasts from a mitochondrial disease dataset ($n = 303$) and GTEx ($n = 492$). **b**, Correlation of the reference Ψ values from the union of introns of the SpliceMaps from **a**. For the union of introns: $n = 736{,}503$, Pearson correlation = 0.87, $R^2 = 0.74$, where reference Ψ of nonintersecting introns was set to zero. For the intersection of introns: $n = 522{,}876$, Pearson correlation = 1.0, $R^2 = 0.99$. **c**, AuPRC for classification of aberrant splicing events from rare variants in the mitochondrial disease dataset for SpliceAI, MMSplice and AbSplice-DNA trained on GTEx using the GTEx fibroblasts SpliceMap from **a**. Error bars represent s.e.m. (Jackknife over samples, $n = 20$). **d**, Enrichment of high-score predictions in ALS genes ($n = 165$). Cutoffs are for SpliceAI (high: 0.8), MMSplice (high: 2) and AbSplice-DNA (high: 0.2). The sample size n in the x-axis labels corresponds to the total number of predictions above the cutoff. P values were computed using one-sided Fisher tests considering all protein coding genes as the universe. **e**, Proportion of rare variants that pass the high cutoffs described in **d** for MMSplice with GENCODE annotation, SpliceAI and AbSplice-DNA trained on GTEx and using GTEx brain SpliceMaps as well as a SpliceMap from ALS motor neurons, validated using proteomics (Z-score < −2; Methods) in the ALS dataset. The sample size n in the y-axis labels corresponds to the total number of predictions above the cutoff. Error bars represent 95% CIs from the binomial test. **f**, Genome-wide depletion of high-impact variants among rare SNVs (gnomAD MAF < 0.1%) within a gene ($n = 19{,}534$) as a function of LOEUF score deciles. High-impact variants are defined by a SpliceAI score > 0.8, MMSplice score > 2 (absolute score) and an AbSplice-DNA score > 0.2 in at least one tissue. Asterisks mark significance levels of two-sided Fisher tests of AbSplice-DNA compared with SpliceAI (*<0.05, **<10^{-4}, ***<10^{-8}). NS, not significant.

(ALS)-affected and 45 healthy individuals from the Answer ALS project (Methods). As iPSC-derived spinal motor neurons were not profiled in GTEx, we considered two approaches. On the one hand, we used the Answer ALS healthy controls to generate a SpliceMap for iPSC-derived spinal motor neurons. On the other hand, we used the SpliceMap of GTEx brain tissues as a proxy which showed the highest overlap from all GTEx tissues (Supplementary Fig. 5). We found that the GTEx SpliceMap from brain tissues agreed reasonably well with the one derived from this cohort both qualitatively (76% shared splice sites) and quantitatively (Pearson correlation 0.86; Supplementary Fig. 5). Here, too, AbSplice-DNA outperformed SpliceAI and MMSplice. Interestingly, AbSplice-DNA achieved similar performances using the SpliceMap from GTEx brain tissues or using the SpliceMap from iPSC-derived spinal motor neurons, suggesting that AbSplice-DNA can be applied robustly in absence of control samples using SpliceMaps from proxy tissues (Supplementary Fig. 6). Moreover, AbSplice-DNA predictions were enriched for genes associated with ALS[28–32] (threefold enrichment; Fig. 4d), which was less so for MMSplice predictions and not the case for SpliceAI predictions. We further validated AbSplice-DNA using proteomics data available for this cohort. At our recommended cutoff, AbSplice-DNA predicted 58 genes to be aberrantly spliced, of which 31% (18 of 58; 95% confidence interval (95% CI), 20–45%) of the corresponding proteins showed significantly low abundance (Z-score < −2; Methods), consistent with RNA degradation via nonsense-mediated decay or protein isoforms resulting from aberrant splicing events that are more poorly translated or less stable. Similarly, independent confirmation by proteomics led to validation rates of MMSplice (26 of 183; 95% CI, 9–20%) and SpliceAI (17 of 80; 95% CI, 13–32%) consistent with the validation rates we originally observed at those cutoffs using the GTEx RNA-seq benchmark (Fig. 3b). Altogether, the proteomics analyses confirm the relative improvements of the different models and are overall consistent with our precision estimates.

Furthermore, we applied AbSplice-DNA to 203,306,868 rare variants (MAF < 0.1%) from the gnomAD dataset using SpliceMaps from all GTEx tissues. In highly constrained genes, defined as the 10% of genes most strongly depleted for loss-of-function variants in gnomAD[13], rare variants were more strongly depleted for high AbSplice-DNA scores in at least one tissue (3.4-fold depletion), than for high SpliceAI scores (2.9-fold depletion, $P < 10^{-21}$; Fig. 4f) or high MMSplice scores (2.2-fold depletion). A stronger depletion than with SpliceAI or MMSplice also held when relaxing the AbSplice-DNA cutoff to match the total number of predictions of SpliceAI (Supplementary Fig. 7).

Collectively, these results on independent data demonstrate the robustness and the applicability of AbSplice-DNA and suggest its utility for rare disease diagnostics and rare variant interpretation.

AbSplice-RNA incorporates RNA-seq from CATs

Sequencing transcriptomes of CATs such as skin or body fluids is of increasing interest in rare disease research as it allows direct detection of aberrant splicing for those splice sites used both in the CAT and in tissues of suspected disease relevance[16,33–35]. The GTEx dataset consists of post-mortem-collected RNA-seq samples across a vast variety of tissues and thereby offers the unique opportunity to evaluate to what extent aberrant splicing in an accessible tissue reflects aberrant splicing of another tissue of interest. One positive example in GTEx is aberrant splicing of *DDX27* in the heart which can also be observed in skin fibroblasts (Fig. 5a). Consistent with a previous study[35] based on the Ensembl gene annotation[36], we found that among the CATs, fibroblasts have the highest overlap of transcribed splice sites according to SpliceMap with nonaccessible tissues, followed by lymphocytes and whole blood (Fig. 5b). To predict aberrant splicing in nonaccessible tissues, we considered ranking genes of an individual first for showing significant and large aberrant splicing in a CAT (false discovery rate (FDR) < 0.1 and $|\Delta\Psi| > 0.3$) and then by significance level. This simple method yielded a markedly increased precision compared with the DNA-based models, up to nearly 40% recall (Fig. 5c and Extended Data Fig. 10a). However, RNA-based predictions remain limited to those splice sites expressed and spliced in the CAT. Therefore, we next trained models integrating AbSplice-DNA features together with RNA-seq-based features from CATs, including differential splicing amplitude estimates to leverage the splicing scaling law and the SpliceMaps (Methods). These models, which we call AbSplice-RNA, outperformed all other models (Fig. 5c and Extended Data Fig. 10a). We found that using fibroblasts only led to the same performance as using all CATs, reaching around 60% precision at 20% recall and amounting to a twofold improvement over AbSplice-DNA (Fig. 5c and Extended Data Fig. 10b). Those improvements were consistent across target tissues (Fig. 5d). As expected, AbSplice-RNA outperformed AbSplice-DNA for genes expressed in CATs and remained on par with it otherwise (Extended Data Fig. 10c). Altogether, these results establish a formal way to integrate direct measurements of aberrant splicing along with sequence-based models to predict aberrant splicing in a tissue of interest.

Discussion

We established a comprehensive benchmark for predicting variants leading to aberrant splicing in human tissues, revealing limited performance of state-of-the-art sequence-based models. We created a tissue-specific splicing annotation (SpliceMap) based on GTEx which maps acceptor and donor splice sites and quantifies their usage in 49 human tissues. We showed that integrating SpliceMap with DNA-based prediction models leads to a threefold increase of precision at the same recall. Additionally, we found that RNA-seq from CATs complements DNA-based splicing predictions when incorporated into an integrative model.

The prediction of splicing-perturbing variants has a long history of over 20 years' work[2–9,26,37–44]. This includes tissue-specific models for mouse[43,44] and more recently human[9,41]. Those models showed successes in various splicing prediction tasks, such as quantitative change of percent spliced-in, splice site usage or splicing efficiency. This study mainly focuses on the prediction of extreme splicing effects (outliers), which has not yet been assessed. This modeling task could be investigated only now, after the development of aberrant splicing callers[10–12] which enabled the establishment of a ground truth for splicing outlier prediction. We foresee that the paradigm of predicting extreme effects in splicing from DNA could be an inspiration for future research and further be extended to aberrant expression or protein abundance prediction. Furthermore, the large multi-tissue cohorts provided by GTEx allowed us to assess and develop tissue-specific predictors. Using aberrant splicing predictions for tissues that are mechanistically related to the disease of interest may prove to be helpful to identify the effector gene, just as tissue-specific predictions are important for transcriptome-wide association studies[45].

Some splicing variant effect prediction models leverage conservation as further evidence of the functional relevance of a variant[7,8]. Even though conservation is a strong indicator of function, we decided not to include conservation in our final model, as variants causing aberrant splicing do not necessarily have to reside on conserved regions. Moreover, conservation depends on the functional importance of the gene. A nucleotide strongly affecting splicing of a nonconserved gene may be less conserved than a nucleotide with a milder effect on splicing located in a highly conserved gene. Also, a nucleotide can be conserved due to its other potential roles besides splicing. For example, exonic regions near splice sites might be conserved due to their role in protein function. Altogether, even though conservation could still marginally yet significantly improve our model (Extended Data Fig. 9), we opted to provide to the community a model predicting aberrant splicing per se by integrating models solely trained on DNA sequence and splicing metrics measured from RNA-seq or massively parallel reporter assays (MPRAs) (SpliceAI and MMSplice). AbSplice users can still benefit from conservation evidence in post-processing steps to further prioritize variants.

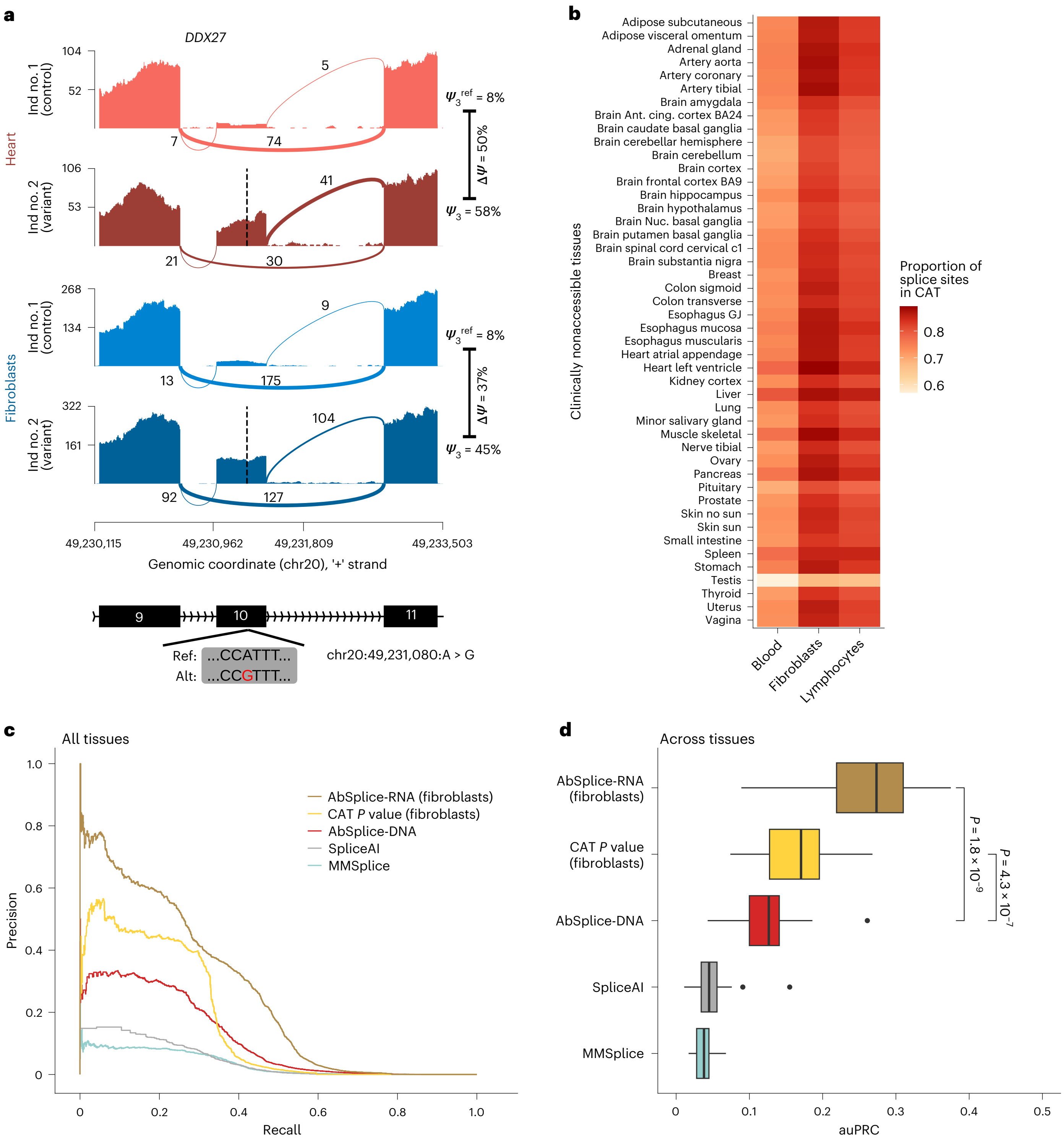


Fig. 5 | Integrating RNA-seq data of CATs to predict aberrant splicing in nonaccessible tissues. a, Sashimi plot of *DDX27* around exon 10 for two individuals in heart and fibroblasts. One individual carries no rare variant in this region (control, upper tracks), and one carries an exonic rare variant (dashed line, lower tracks) associated with increased splicing of exon 10. This exon shows a similar usage in fibroblasts and in the heart (reference donor site percent spliced-in, $\Psi_3 = 8\%$, according to SpliceMap in both tissues, in line with the measured values for the displayed control individual: $\Psi_3 = 6\%$ in heart and $\Psi_3 = 5\%$ in fibroblasts). The effect associated with the variant in fibroblasts approximates well the one in heart (difference of donor site usage, $\Delta\Psi_3 = 50\%$ in heart and 37% in fibroblasts). In this case, aberrant splicing can be directly detected from the accessible tissue. **b**, Proportion of splice sites used in GTEx clinically nonaccessible target tissues (rows) also used in GTEx CATs (columns). **c**, Precision–recall curve comparing the overall prediction performance on all GTEx tissues of SpliceAI, MMSplice using GENCODE annotation, AbSplice-DNA, gene-level FRASER *P* values in fibroblasts and AbSplice-RNA, which integrates AbSplice-DNA features with features from RNA-seq from fibroblasts. **d**, Distribution of the auPRC of the models in **c** across tissues ($n = 49$). Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles; *P* values were computed using the paired one-sided Wilcoxon test.

We constructed SpliceMaps and detected aberrant splicing from short-read RNA-seq. We found that current long-read RNA-seq data available for GTEx[23] did not provide sufficient coverage to detect unannotated splice sites (Supplementary Fig. 2). Since split short reads reveal splice sites, we foresee that the major added value of long-read over short-read sequencing is not about calling splice sites but identifying the complete RNA isoforms. This could be used in the future to develop models predicting the exact splicing outcome (for example, exact elongated or truncated exon boundaries, exon combinations and so on) caused by the variant, which is beyond the scope of current models trained primarily on short-read data.

We showed how RNA-seq of CATs effectively complements DNA-based predictions. An alternative to this approach is to reprogram or transdifferentiate cells into the suspected mechanistically involved cell type and perform RNA-seq on them[46]. This approach has, however, important caveats. First, it is not ensured that the suspected mechanistically involved cell type is the correct one, as symptoms may manifest more strongly in downstream affected tissues. Second, this approach is cost, time and labor intensive. Third, cell reprogramming can induce and select mutations which may lead to false identifications. Therefore, predictive models that can leverage RNA-seq of CATs will probably remain relevant in practice[47]. Furthermore, RNA-seq reveals the consequence of the splicing defect on the resulting transcript isoform (for example, frameshift or exon truncation), which is crucial for diagnostics.

By increasing the precision at 20% recall from about 10% to 60%, the cumulative improvements of our models are substantial. Still, a majority of the aberrant splicing events are not recalled and there remains a majority of false positives. An unknown and potentially large fraction of events that are not recalled might be aberrant splicing calling artifacts, as suggested by the high number of singleton calls. In this study, we implemented strategies aiming at improving the proportion of genuine genetically driven aberrant splicing events in the ground truth while not introducing biases favoring particular models (Extended Data Figs. 2–4). However, every classification task is founded on a reliable ground truth. As splicing is a complex process and not all aberrant events can be reliably called by state-of-the art aberrant splicing callers, the ground truth in the prediction task remains a proxy. Progress in aberrant splicing calling or better understanding of the technical reasons could reduce the number of incorrectly called aberrant splicing events and improve the recall. Moreover, some of the apparent false positive predictions may be actually correct. This is the case when the aberrant splicing isoform contains a premature termination codon and, often, though not systematically[48], gets rapidly degraded by nonsense-mediated decay. Rapidly degraded isoforms barely have any reads in RNA-seq data and hence are typically not detected by aberrant splicing callers. In diagnostic applications, those variants remain relevant. Moreover, dedicated experiments can be done to test whether aberrant splicing is taking place, for instance, using the translation inhibitor cycloheximide.

As WGS becomes more readily available in research and healthcare, there is a growing need for accurate annotation of noncoding variants with strong deleterious effects for establishing genetic diagnostics of rare disorders, identifying effector genes of common diseases and more precisely stratifying patients with cancer based on their tumor genetic alterations. Variants causing aberrant splicing are not only a major class of such noncoding loss-of-function variants, but their mechanisms of action also now become targetable for an increasingly rich therapeutic arsenal[49]. Hence, because of its high precision and its focus on extreme events, we foresee AbSplice to be instrumental for genome-based diagnostics and therapy design.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01373-3.

References

1. Zappala, Z. & Montgomery, S. B. Non-coding loss-of-function variation in human genomes. *Hum. Hered.* **81**, 78–87 (2016).
2. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548.e24 (2019).
3. Cheng, J. et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* **20**, 48 (2019).
4. 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).
5. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).
6. Xiong, H. Y. et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
7. Rentzsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 31 (2021).
8. Danis, D. et al. Interpretable prioritization of splice variants in diagnostic next-generation sequencing. *Am. J. Hum. Genet.* **108**, 2205 (2021).
9. Cheng, J., Çelik, M. H., Kundaje, A. & Gagneur, J. MTSplice predicts effects of genetic variants on tissue-specific splicing. *Genome Biol.* **22**, 94 (2021).
10. Mertes, C. et al. Detection of aberrant splicing events in RNA-seq data using FRASER. *Nat. Commun.* **12**, 529 (2021).
11. Jenkinson, G. et al. LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics* **36**, 4609–4615 (2020).
12. Ferraro, N. M. et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369**, eaaz5900 (2020).
13. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
14. Wilks, C. et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* **22**, 323 (2021).
15. Ling, J. P. et al. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nat. Commun.* **11**, 137 (2020).
16. Kremer, L. S. et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8**, 15824 (2017).
17. Dawes, R., Joshi, H. & Cooper, S. T. Empirical prediction of variant-activated cryptic splice donors using population-based RNA-Seq data. *Nat. Commun.* **13**, 1655 (2022).
18. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
19. Elliott, D. J. & Grellscheid, S. N. Alternative RNA splicing regulation in the testis. *Reprodroduction* **132**, 811–819 (2006).
20. de la Grange, P., Gratadou, L., Delord, M., Dutertre, M. & Auboeuf, D. Splicing factor and exon profiling across human tissues. *Nucleic Acids Res.* **38**, 2825–2838 (2010).
21. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
22. Cotto, K. C., Feng, Y. Y., Ramu, A. et al. Integrated analysis of genomic and transcriptomic data for the discovery of splice-associated variants in cancer. Nat Commun **14**, 1589 (2023).
23. Glinos, D. A. et al. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**, 353–359 (2022).

24. Amarasinghe, S. L. et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).
25. Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J. & Lehner, B. Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. *Cell* **176**, 549–563.e23 (2019).
26. Cheng, J., Çelik, M. H., Nguyen, T. Y. D., Avsec, Ž. & Gagneur, J. CAGI 5 splicing challenge: improved exon skipping and intron retention predictions with MMSplice. *Hum. Mutat.* **40**, 1243–1251 (2019).
27. Yépez, V. A. et al. Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *Genome Med.* **14**, 38 (2022).
28. Abel, O., Powell, J. F., Andersen, P. M. & Al-Chalabi, A. ALSoD: a user-friendly online bioinformatics tool for amyotrophic lateral sclerosis genetics. *Hum. Mutat.* **33**, 1345–1351 (2012).
29. Gregory, J. M., Fagegaltier, D., Phatnani, H. & Harms, M. B. Genetics of amyotrophic lateral sclerosis. *Curr. Genet. Med. Rep.* **8**, 121–131 (2020).
30. Pecoraro, V. et al. The NGS technology for the identification of genes associated with the ALS. A systematic review. *Eur. J. Clin. Invest.* **50**, e13228 (2020).
31. Hardiman, O. et al. Amyotrophic lateral sclerosis. *Nat. Rev. Dis. Primers* **3**, 17071 (2017).
32. McCann, E. P. et al. Evidence for polygenic and oligogenic basis of Australian sporadic amyotrophic lateral sclerosis. *J. Med. Genet*. https://doi.org/10.1136/jmedgenet-2020-106866 (2020).
33. Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).
34. Frésard, L. et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat. Med.* **25**, 911–919 (2019).
35. Aicher, J. K., Jewell, P., Vaquero-Garcia, J., Barash, Y. & Bhoj, E. J. Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. *Genet. Med.* **22**, 1181–1190 (2020).
36. Yates, A. D. et al. Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).
37. Pertea, M., Lin, X. & Salzberg, S. L. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* **29**, 1185–1190 (2001).
38. Desmet, F.-O. et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37**, e67 (2009).
39. Ke, S. et al. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* **21**, 1360–1374 (2011).
40. Jian, X., Boerwinkle, E. & Liu, X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* **42**, 13534–13544 (2014).
41. Xiong, H. Y. et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
42. Sonnenburg, S., Schweikert, G., Philips, P., Behr, J. & Rätsch, G. Accurate splice site prediction using support vector machines. *BMC Bioinf.* **8**, S7 (2007).
43. Barash, Y. et al. Deciphering the splicing code. *Nature* **465**, 53–59 (2010).
44. Xiong, H. Y., Barash, Y. & Frey, B. J. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* **27**, 2554–2562 (2011).
45. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
46. Gonorazky, H. D. et al. Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease. *Am. J. Hum. Genet.* **104**, 466–483 (2019).
47. Martorella M. Noninvasive, low-cost RNA-sequencing enhances discovery potential of transcriptome studies. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/2022.09.06.506813v1 (2022).
48. Teran, N. A. et al. Nonsense-mediated decay is highly stable across individuals and tissues. *Am. J. Hum. Genet.* **108**, 1401–1408 (2021).
49. Rogalska, M. E., Vivori, C. & Valcárcel, J. Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-022-00556-8 (2022).
50. Zhang, Y., Zhou, R. & Wang, Y. Sashimi.py: a flexible toolkit for combinatorial analysis of genomic data. Preprint at *bioRxiv* https://doi.org/10.1101/2022.11.02.514803 (2022).
51. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

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

Methods

Ethics statement

No primary data were generated for this study. Person-related data were obtained through authorized access from primary data controllers. The study adheres to the ethical and research agreements between the Technical University of Munich and the primary data controllers. All participant informed consents were collected by and remain with the primary data controllers.

Statistics and reproducibility

No statistical method was used to predetermine sample size. We did not use any study design that required randomization or blinding. In the GTEx data we excluded tissues with fewer than 50 samples. In the ALS and mitochondrial disease datasets we did not exclude any samples.

Datasets

GTEx. We downloaded the RNA-seq read alignment files (BAM files) and the variant calling files (VCF files) from WGS from GTEx v8p (hg38) from the database of Genotypes and Phenotypes (dbGaP) (study accession: phs000424.v8.p2). We used data from 946 individuals with paired WGS and RNA-seq measurements ($n = 16{,}213$) in at least one tissue. For the long-read RNA-seq data, we downloaded the transcript annotation (GTF) generated by FLAIR[52] based on 88 Nanopore samples from the GTEx portal.

Mitochondrial disease dataset. The dataset consists of 303 patients with mitochondriopathy described by Yépez et al.[27], all of which have RNA-seq from skin-derived fibroblasts. For 20 individuals, WGS is also available.

ALS dataset. The dataset consists of WGS, RNA-seq and proteomics data from 245 individuals diagnosed with ALS and 45 control samples. RNA-seq data were obtained from iPSC-derived spinal motor neurons. We downloaded the data from the Answer ALS portal (dataportal.answerals.org). Genes known to be involved in ALS disease development were manually curated from literature[28–32].

Data preprocessing

Rare variants. Variants had to be supported by at least ten reads and had to pass the conservative genotype-quality filter of $GQ \geq 99$. These criteria were used for single nucleotide variants (SNVs) and indels in the same way. We considered a variant to be rare if it had an MAF in the general population ≤0.001 based on gnomAD (v.3.1.2) and was found in at most two individuals within each cohort.

Splicing outlier detection. Splicing outliers were called using FRASER[10] (v.1.6.0) as implemented in the Detection of RNA-seq Outliers Pipeline[53] (v.1.1.2). FRASER was used to detect introns (including de novo introns) and to count split reads for each intron. Based on the split-read counts, three intron-centric metrics were calculated: alternative acceptor usage with the ψ_5 metric, alternative donor usage with the ψ_3 metric, and splicing efficiencies as defined with the θ_5 and θ_3 metrics[54]:

$$\Psi_5(D,A) = \frac{n(D,A)}{\sum_{A'} n(D,A')} = \frac{k}{N_5}$$

$$\Psi_3(D,A) = \frac{n(D,A)}{\sum_{D'} n(D',A)} = \frac{k}{N_3}$$

$$\theta_5 = \frac{\sum_{A'} n(D,A')}{n(D) + \sum_{A'} n(D,A')}$$

$$\theta_3 = \frac{\sum_{D'} n(D',A)}{n(A) + \sum_{D'} n(D',A)}$$

where k is the number of split reads supporting the intron from donor D to acceptor A. The sum in the denominator of $\psi_5(D,A)$ goes over all possible acceptors A' for donor D, and the sum in the denominator of $\psi_3(D,A)$ goes over all possible donors D' for acceptor A. In the splicing efficiencies, the denominator contains $n(D)$ or $n(A)$ which are the numbers of nonsplit reads spanning the exon–intron boundary of donor D or acceptor A, respectively. The advantage of these intron-centric metrics over the exon-centric metric percent spliced-in (ψ) is that they do not require exons to be mapped, which is an ill-defined task when starting from short-read RNA-seq data.

FRASER models these metrics while controlling for latent confounders and reports both splice-site-level and gene-level FDRs. We called aberrant spliced genes using the gene-level FDR < 0.1 as in Mertes et al.[10] Furthermore, we requested the gene to contain at least one significant splice site (FDR < 0.05, FRASER default) supported by 20 reads and with an absolute deviation of $\psi_{5,3}$ from the FRASER-modeled expected value larger than 0.3 (denoted $|\Delta\psi_{5,3}| > 0.3$). The same filters were applied to the splicing efficiency metrics.

To discard aberrant splicing calls that probably have no genetic basis[10], we additionally applied and compared different filtering methods (Extended Data Fig. 4). In the GTEx dataset, where multiple RNA-seq samples from the same individual are available, we investigated including splicing outliers from at least two tissues from the same individual (Filter 2; Extended Data Fig. 4b). Here, a gene-level outlier was considered to be replicated if the same splice-site-level outlier was detected in multiple tissues. As this strategy cannot be applied to other single-tissue datasets, we alternatively filtered for splicing outliers containing a rare variant in the vicinity of ±250 bp of every splice site based on RNA-seq from the sample (Filter 3; Extended Data Fig. 4c). Importantly, this filter was applied to all splice sites identified by FRASER, which includes both annotated splice sites as well as cryptic splice sites (Extended Data Fig. 3). For consistency, all reported results are based on Filter 3.

Aberrant splicing prediction benchmark

Aberrant splicing prediction task. The task is to predict whether a protein coding gene with one or more rare variants within the gene body is aberrantly spliced in a given tissue of an individual.

Performance evaluation metric. Due to the large class imbalance in the splicing outlier prediction benchmarking dataset, we chose to evaluate models using precision–recall curves. As evaluation metric we used the auPRC, computed using the average precision (AP) score[55] (which represents the mean of precisions for each threshold weighted by the recall difference):

$$\mathrm{AP} = \sum_{n} (R_n - R_{n-1}) P_n$$

where P_n and R_n are the precision and recall at the nth threshold.

Tissue-specific SpliceMap

For each tissue separately, we created a SpliceMap that lists all active introns along with aggregate statistics about acceptor and donor site usage useful for aberrant splicing prediction purposes.

Active introns. We started from all introns reported by FRASER. We filtered out untranscribed splice sites and background noise by filtering out introns not supported by any split-read in more than 95% of the samples. For this and other operations involving genomic ranges we used PyRanges[56] (v.0.0.115).

Aggregate statistics. Aggregate statistics were calculated on donor and acceptor sites independently. For donor site usage, the SpliceMap aggregate statistics are (1) the total number of split reads across samples (s) supporting the intron ($\Sigma_s k$), (2) the total number of split reads across samples sharing the same acceptor site ($\Sigma_s N_3$), (3) the median number of split reads per sample sharing the same acceptor site and (4) the reference isoform proportion (ψ_3^{ref}), defined as $\psi_3^{\text{ref}} = \frac{\Sigma_s k}{\Sigma_s N_3}$. Aggregate statistics were computed analogously for acceptor site usage.

Exclusion of rare variant data in aggregate statistics. To prevent information leakage, the aggregate statistics were computed so that they do not contain information about splicing events associated with rare variants (specifically, we excluded from the computations of the aggregate statistics data from samples with a rare variant within ±250 bp of any donor or acceptor site).

SpliceMap generation using alternative counting tools. SpliceMaps were also created from split-read (introns) counts using Regtools[22] (v.0.5.2) and STAR[21] (v.2.5.3) for the 'Skin – Not Sun Exposed (Suprapubic)' tissue. We ran Regtools using BAM files. Regtools performs annotation-free counting; thus, it also calls unannotated introns and splice sites. We downloaded STAR split-read counts from the GTEx portal. The GTEx pipeline filters unannotated splice sites, although the STAR two-pass approach could call unannotated splice sites and introns. During SpliceMap generation, active introns and aggregate statistics were computed as described above.

Aberrant splicing prediction models

SpliceAI. SpliceAI[2] (v.1.3.1) is a deep learning model that predicts splice site alteration for acceptor and donor sites from sequence. SpliceAI is annotation free and can therefore score all variants including cryptic splice sites created by deep intronic variants. SpliceAI provides precomputed scores for all SNVs and indels up to the length of 4 nucleotides. These variant scores were computed with 50 bp as the maximum distance between the variant and gained/lost splice sites. We downloaded precomputed variant scores from Illumina BaseSpace and stored them in a RocksDB[57] (v.6.10.2) key-value database for fast lookup. We ran SpliceAI to obtain variant scores for long indels not available in the database. Also, we used masked scores of SpliceAI as recommended by the authors for variant interpretation. This masking sets Delta scores to zero if SpliceAI predicts activation for annotated splice sites and deactivation for unannotated splice sites.

SpliceAI + SpliceMap. We used tissue-specific splice site annotations from SpliceMap together with SpliceAI predictions. For each tissue, we retained those variant scores that contained an annotated splice site within a 100-bp window.

SpliceAI + SpliceMap + $\boldsymbol{\psi}_{\text{ref}}$. As SpliceAI was trained to predict creation or loss of splice sites and not ψ, there is no principled way to apply the splicing scaling law to include reference levels. Therefore, we used reference levels only to filter predictions. Analogously to the masking of scores representing annotated acceptor/donor gain and unannotated acceptor/donor loss performed by the authors of SpliceAI, we used tissue-specific ψ_{ref} values for filtering. Specifically, variant scores associated with acceptor/donor gain and a splice site with $\psi_{\text{ref}} \geq 0.95$ as well as with acceptor/donor loss and a splice site with $\psi_{\text{ref}} \leq 0.05$ were filtered out.

MMSplice. MMsplice[3] (v.2.3.0) is a deep learning model that predicts the impact of a variant (in a 100-bp window of annotated splice sites) on alternative usage of a nearby donor or acceptor site. MMSplice predicts the effect of a variant in log-odds ratios (denoted $\Delta\text{logit}\psi_5$ or $\Delta\text{logit}\psi_3$). MMSplice requires a splice site annotation. We used the GENCODE (release 38 of hg38) annotation.

MMSplice + SpliceMap. We ran MMSplice on tissue-specific splice site annotations from SpliceMap.

MMSplice + SpliceMap + $\boldsymbol{\psi}_{\text{ref}}$. MMSplice is a quantitative model predicting percent spliced-in for which the splicing scaling law can be leveraged to integrate reference levels. For conversion of the variant effect into natural scale, reference levels of donor site and acceptor site usages are required. For the sake of shorter notations, we write in the following ψ instead of ψ_5 and ψ_3. We used MMSplice to predict $\Delta\text{logit}(\psi)$ values. $\Delta\text{logit}(\psi)$ values were then combined with the corresponding reference ψ value (ψ_{ref}) in SpliceMap: first in logit scale to adjust the predicted variant effect by MMSplice to the correct reference level; then in natural scale by using the sigmoid function (Extended Data Fig. 7a):

$$\Delta\text{logit}(\Psi) = \text{logit}(\Psi_{\text{alt}}) - \text{logit}(\Psi_{\text{ref}})$$

$$\hat{\Psi}_{\text{alt}} = \sigma\left(\Delta\text{logit}\left(\Psi\right) + \text{logit}\left(\Psi_{\text{ref}}\right)\right)$$

$$\Delta\hat{\Psi} = \hat{\Psi}_{\text{alt}} - \Psi_{\text{ref}}$$

$$\sigma^{-1} = \text{logit}$$

Variants further away than 100 bp from any SpliceMap splice site were scored 0 (no effect).

MTSplice. MTSplice[9] (v.2.3.0) is a tissue-specific version of MMSplice. The model scores each exon–variant pair for 56 tissues. With respect to each annotated exon boundary, the model takes as input a sequence of 100 bp in the exon and 300 bp in the intron. MTSplice predicts the tissue-specific effect of a variant in log-odds ratios (denoted $\Delta\text{logit}(\psi)$). MTSplice requires a splice site annotation. We used the GENCODE (release 38 of hg38) annotation.

CADD-Splice. CADD-Splice[7] is an ensemble model that combines CADD scores (contains conservation scores) together with splicing predictions from SpliceAI and MMSplice. We ran CADD-Splice v.1.6. CADD-Splice provides raw and PHRED-scaled scores. We used the PHRED score.

SQUIRLS. SQUIRLS[8] is based on engineered splicing features for donor and acceptor sites that are extracted from a genome annotation. SQUIRLS predicts the probability of a variant to alter the splicing pattern. We downloaded the SQUIRLS database v.2203 and ran SQUIRLS v.2.0.0.

AbSplice-DNA. AbSplice-DNA is a generalized additive model, namely the ExplainableBoostingClassifier from the python package interpretml[58]. Similar performance was achieved using a random forest or logistic regression model from scikit-learn[55]. The features of AbSplice-DNA were the prediction score from MMSplice + SpliceMap, MMSplice + SpliceMap + ψ_{ref}, the SpliceAI Delta score and a binary feature from SpliceMap indicating if the splice site is expressed in the target tissue (using a cutoff of 10 reads for the median number of split reads sharing the splice site). The model includes interaction terms, thereby de facto capturing the effect of combining SpliceMap with SpliceAI scores. The model was trained on a variant level using outliers within 250-bp distance of rare variants as ground truth (Extended Data Fig. 4c before aggregation to gene level). The model was trained with fivefold-stratified cross-validation, grouped by individuals to avoid information leakage, and such that the proportions of the negative (variant is associated with no outlier on the gene) and positive (variant is associated with an outlier on the gene) classes were preserved in each fold.

Predictors using RNA-seq from CATs. We used different features from RNA-seq of three CATs from GTEx (Whole blood, Cells transformed fibroblasts, and Cells Epstein-Barr virus (EBV)-transformed lymphocytes) to predict aberrant splicing in nonaccessible target tissues.

As one predictive feature we used the $-\log_{10}$ nominal gene-level P values obtained using FRASER. In the benchmark, we ranked all splicing outlier genes (FDR < 0.1 and $|\Delta\psi| > 0.3$) lower than the remaining genes, and further ranked genes within each of these two groups by increasing P value.

Additionally, we used SpliceMaps from the accessible and the nonaccessible tissues together with ψ measurements from RNA-seq and applied the splicing scaling law to infer $\Delta\psi$ values in the nonaccessible target tissue:

$$\Delta\text{logit}\left(\Psi\right) = \text{logit}\left(\Psi^{\text{CAT}}\right) - \text{logit}\left(\Psi^{\text{CAT}}_{\text{ref}}\right)$$

$$\Psi^{\text{target}} = \sigma\left(\Delta\text{logit}\left(\Psi\right) + \text{logit}\left(\Psi^{\text{target}}_{\text{ref}}\right)\right)$$

$$\Delta\Psi^{\text{target}} = \Psi^{\text{target}} - \Psi^{\text{target}}_{\text{ref}}$$

where Ψ^{CAT} is the splicing level in the CAT and $\Psi^{\text{CAT}}_{\text{ref}}$ is the reference level of splicing obtained from SpliceMap, and the difference of these two values provides the tissue unspecific variant effect, $\Delta\text{logit}(\Psi)$. Then, adding $\Delta\text{logit}(\Psi)$ with the reference level of splicing of the target tissue $\text{logit}\left(\Psi^{\text{target}}_{\text{ref}}\right)$ in logit scale and converting back to natural scale provides Ψ^{target} in the target tissue. Subtracting the reference level of splicing of the target tissue, $\Psi^{\text{target}}_{\text{ref}}$, provides the predicted splicing change in the target tissue, $\Delta\Psi^{\text{target}}$, using RNA-seq measurements in CAT.

All precision–recall curves involving CATs have been computed on a subset of the data, excluding CATs from the target tissues and only containing individuals that have RNA-seq measurements from multiple tissues (including the CAT).

AbSplice-RNA. We trained integrative models using the two predictors from RNA-seq data from CATs described above in addition to DNA-based features used in AbSplice-DNA.

We trained AbSplice-RNA models using a single CAT and all CATs together. For the model using all CATs together we trained AbSplice-RNA in a CAT-agnostic manner such that the model predicts outliers regardless of the CAT source. This might be helpful in a diagnostic setting as it might be that the available CAT differs from the CATs that AbSplice-RNA was trained on.

Gene-level aggregation. For genes with multiple variants, we retained the largest score per model.

Model performance per variant and outlier category

Variant categories were annotated with the Ensembl Variant Effect Predictor (VEP)[51]. For each variant, the most severe VEP annotation was considered. For the 'Exon' category, the following VEP categories were grouped together: synonymous_variant, missense_variant, stop_lost, stop_gained. For the nonexclusive splicing outlier categories, we defined 'exon elongation', 'exon truncation', 'exon skipping' using FRASER's branch: https://github.com/c-mertes/FRASER/tree/junction_annotation ref. 59. We defined the category 'Any alternative donor or acceptor choice' as any ψ_5 or ψ_3 outlier, and the category 'Any splicing efficiency outlier' as any θ outlier.

Enrichment in known ALS genes

The enrichment of 165 manually curated genes involved in ALS[28–31] was computed as the proportion of high-splicing-impact variants within those genes, divided by all the high-score predictions of the respective models. Depletion was computed as 1/enrichment.

Proteomics in ALS

We downloaded the protein intensities matrix from the ALS cohort consisting of 4,442 proteins and 204 samples from the Answer ALS portal. We considered the 178 affected individuals. Proteins with missing values in more than 30% of the samples were filtered out, with 3,329 remaining. We then ran PROTRIDER[60], a denoising autoencoder-based method to detect outliers on proteomics data. The encoding dimension was optimized by injecting outliers. No covariates were provided. Z-scores were extracted from the results table.

Depletion in loss-of-function intolerant genes

For all possible rare SNVs (gnomAD MAF < 0.1%) in 19,534 protein coding genes, we computed AbSplice-DNA scores and obtained the SpliceAI precomputed scores from Illumina BaseSpace. The loss-of-function observed/expected upper bound fraction (LOEUF) scores were downloaded from https://gnomad.broadinstitute.org/downloads. For each LOEUF decile we computed the proportion of high-splicing-impact variants to the total sum of high-impact variants and divided it by the proportion of rare variants in each decile.

Reporting summary

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

Data availability

No primary data were generated for this study. Rare variants from gnomAD v.3.1.2 are publicly available at https://gnomad.broadinstitute.org. The GTEx v8 dataset is available at (under dbGaP protection) https://gtexportal.org/home. The ALS dataset is available at http://dataportal.answerals.org after a registration and approval process. The mitochondrial dataset is described by Yépez et al.[27]. Precomputed SpliceAI scores are publicly available at Illumina Basespace, https://basespace.illumina.com/s/otSPW8hnhaZR, after registration. SpliceMaps for all 49 GTEx tissues and iPSC-derived spinal motor neurons from ALS (hg38) are available at Zenodo, https://doi.org/10.5281/zenodo.6387937. Precomputed AbSplice-DNA scores (hg38) in all 49 GTEx tissues are available at Zenodo, https://doi.org/10.5281/zenodo.6408331. Due to potential donor re-identification when revealing rare variants, the benchmark dataset cannot be shared without restrictions. Users with access to the GTEx data can reproduce the benchmark using the code repository below.

Code availability

SpliceMaps can be generated using the custom-written python package 'splicemap' (publicly available at: https://github.com/gagneurlab/splicemap ref. 61). AbSplice predictions using the enhanced SpliceMap annotation can be performed with the custom-written python package 'absplice' (publicly available at: https://github.com/gagneurlab/absplice ref. 62). We also provide a fast implementation of computing SpliceAI predictions using a wrapper based on fast lookup from a database of precomputed scores for existing variants and running SpliceAI for not precomputed variants at https://github.com/gagneurlab/spliceai_rocksdb ref. 63. Fast lookup of all gnomAD variants can be performed with https://github.com/gagneurlab/gnomad_rocksdb ref. 64. The analyses are available under https://github.com/gagneurlab/AbSplice_analysis ref. 65.

References

52. Tang, A. D. et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.* **11**, 1438 (2020).
53. Yépez, V. A. et al. Detection of aberrant gene expression events in RNA sequencing data. *Nat. Protoc.* **16**, 1276–1296 (2021).
54. Pervouchine, D. D., Knowles, D. G. & Guigo, R. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29**, 273–274 (2013).
55. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

56. Stovner, E. B. & Sætrom, P. PyRanges: efficient comparison of genomic intervals in Python. *Bioinformatics* **36**, 918–919 (2020).
57. Dong, S., Kryczka, A., Jin, Y. & Stumm, M. RocksDB: evolution of development priorities in a key-value store serving large-scale applications. *ACM Trans. Storage* **17**, 26:1–26:32 (2021).
58. Nori, H., Jenkins, S., Koch, P. & Caruana, R. InterpretML: a unified framework for machine learning interpretability. Preprint at arXiv190909223 Cs Stat https://doi.org/10.48550/arXiv.1909.09223 (2019).
59. Mertes, C., Scheller, I. & Gagneur, J. FRASER code used in AbSplice publication. *Zenodo* https://doi.org/10.5281/zenodo.7447804 (2022).
60. Kopajtich, R. et al. Integration of proteomics with genomics and transcriptomics increases the diagnostic rate of Mendelian disorders. Preprint at https://www.medrxiv.org/content/10.1101/2021.03.09.21253187v1 (2021).
61. Wagner, N. et al. SpliceMap code used in AbSplice publication. *Zenodo* https://doi.org/10.5281/zenodo.7626022 (2022).
62. Wagner, N. et al. AbSplice code used in AbSplice publication. *Zenodo* https://doi.org/10.5281/zenodo.7626035 (2022).
63. Wagner, N. et al. Code to generate SpliceAI rocksdb used in AbSplice publication. *Zenodo* https://doi.org/10.5281/zenodo.7626078 (2022).
64. Wagner, N. et al. Code to generate gnomAD rocksdb used in AbSplice publication. *Zenodo* https://doi.org/10.5281/zenodo.7625641 (2022).
65. Wagner, N. et al. Analysis code used in AbSplice publication. *Zenodo* https://doi.org/10.5281/zenodo.7628868 (2022).

Acknowledgements

We thank I. Scheller for feedback and advice. M.H.C. thanks X. Xie and A. Mortazavi for institutional support. The German Bundesministerium für Bildung und Forschung (BMBF) supported the study through the Model Exchange for Regulatory Genomics project (MERGE; grant no. 031L0174A to F.R.H. and J.G.), the German Network for Mitochondrial Disorders (mitoNET; grant no. 01GM1906B to V.A.Y.), the ERA PerMed project PerMiM (grant no. 01KU2016A to H.P. and grant no. 01KU2016B to J.G.) and the EJP RD project GENOMIT (grant no. 01GM1920A to H.P.). N.W. is supported by the Helmholtz Association under the joint research school 'Munich School for Data Science – MUDS'. This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), via the projects 'Identification of host genetic variation predisposing to severe COVID-19 by genetics, transcriptomics and functional analyses' (grant no. 466168909 to V.A.Y. and J.G.) and NFDI 1/1 'GHGA - German Human Genome-Phenome Archive' (grant no. 441914366 to C.M. and J.G.). Figure 1 was created with BioRender.com. The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by the NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS. This study was supported by data provided by the Answer ALS Consortium, administered by the Robert Packard Center for ALS at Johns Hopkins. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

J.G. conceptualized the project. N.W., M.H.C. and J.G. designed the methodology. N.W. and M.H.C. provided the software. N.W., M.H.C., F.R.H., H.P. and V.A.Y. performed validations. N.W., M.H.C., F.R.H., V.A.Y. and C.M. performed the formal analysis. N.W., M.H.C., F.R.H. and V.A.Y. curated the data. N.W., M.H.C., V.A.Y. and J.G. wrote the original draft of the manuscript. All authors reviewed and edited the manuscript. N.W., M.H.C., F.R.H., V.A.Y. and J.G. performed visualizations. J.G. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

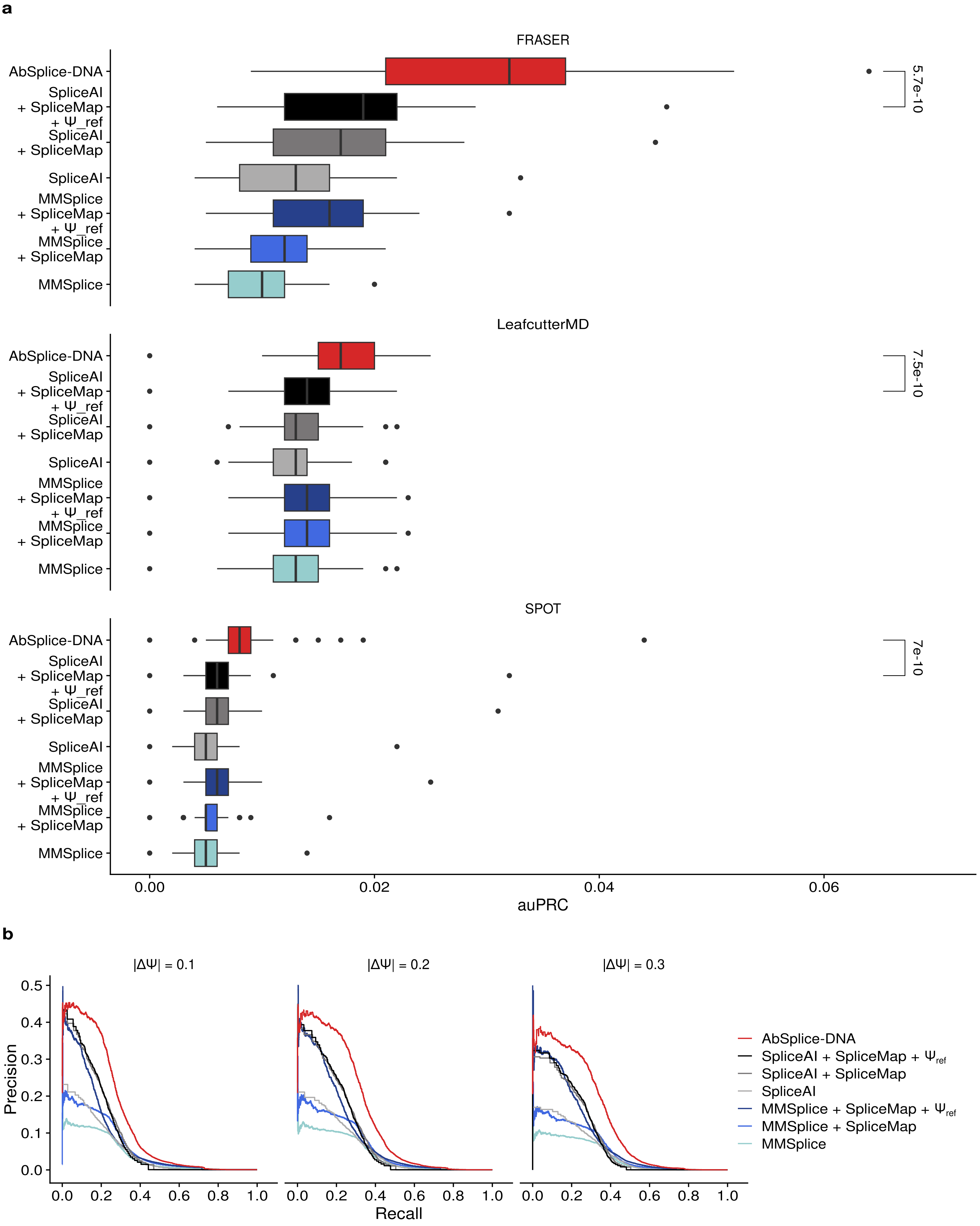
Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01373-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01373-3.

Correspondence and requests for materials should be addressed to Julien Gagneur.

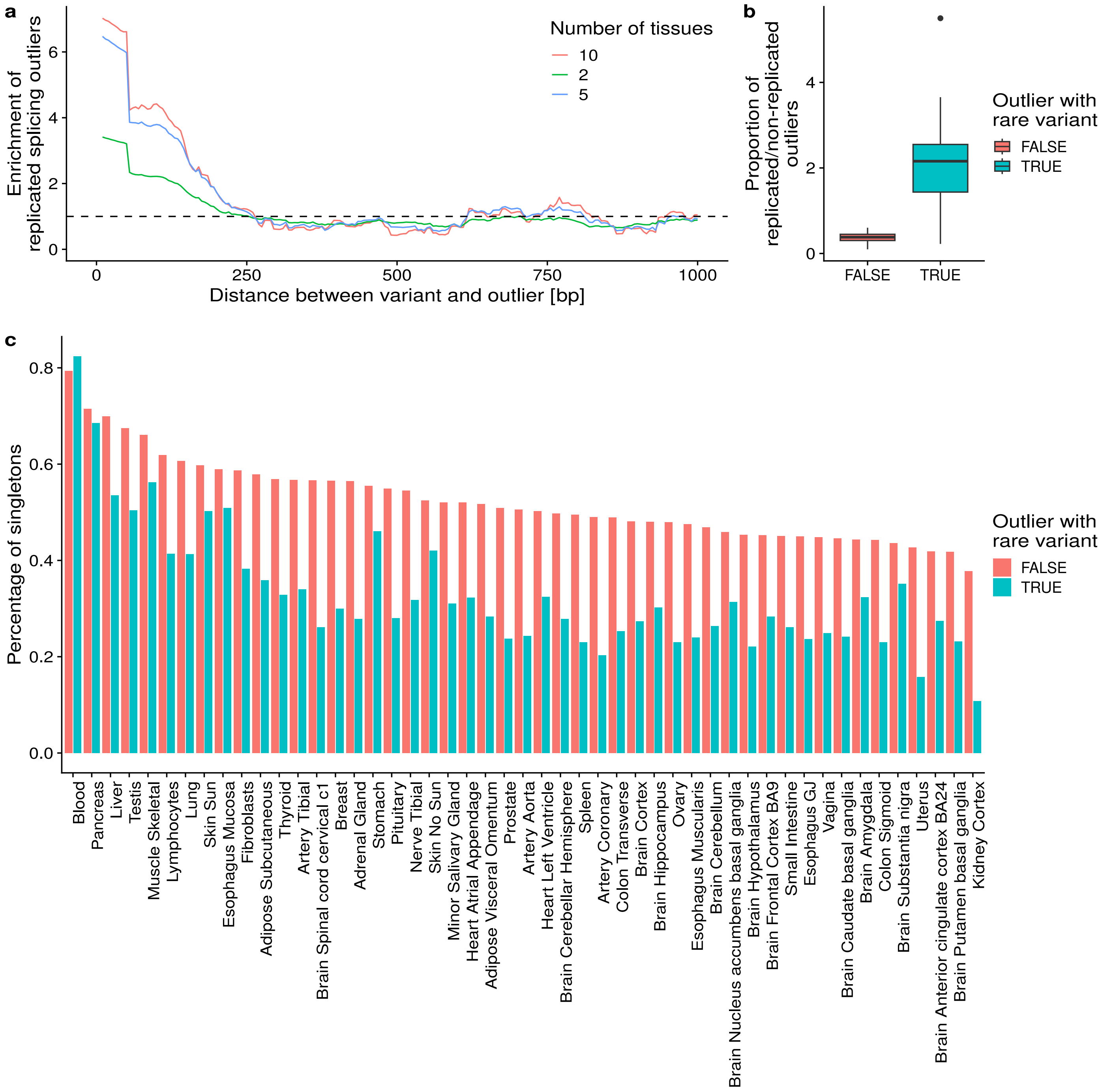
Peer review information *Nature Genetics* thanks Jamie Ellingford, Xin Gao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

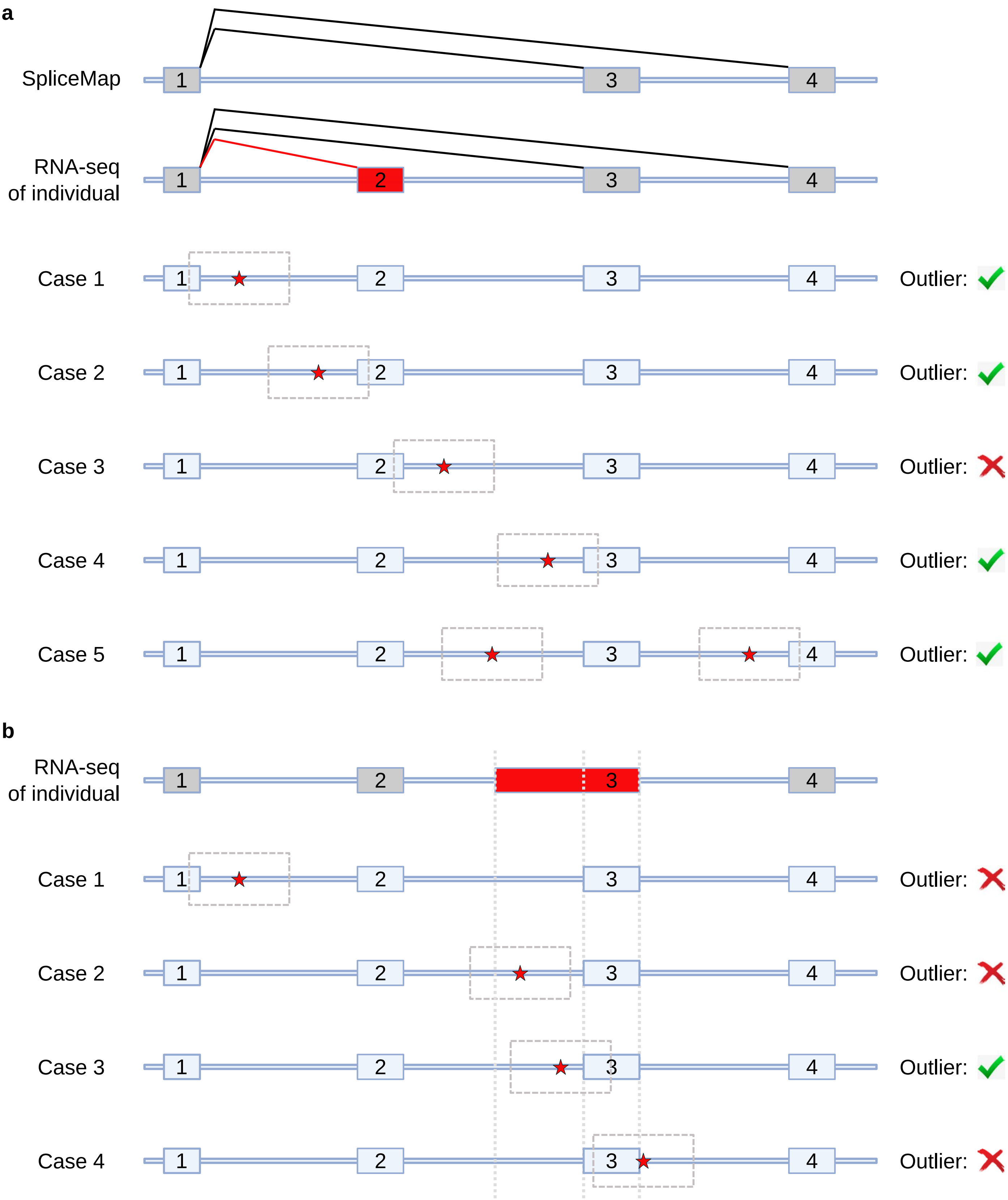


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Performance comparison with different outlier detection methods and different differential splicing cutoffs. a, Distribution of the area under the precision-recall curve across GTEx tissues (n = 49) of different prediction methods (SpliceAI, SpliceAI using SpliceMap annotation, SpliceAI using SpliceMap annotation along with quantitative reference levels of splicing, MMSplice using GENCODE annotation, MMSplice using SpliceMap annotation, MMSplice using SpliceMap annotation along with quantitative reference levels of splicing, and the integrative model AbSplice-DNA) taking as ground truth 3 different aberrant splicing callers: FRASER, LeafcutterMD and SPOT. A gene was considered aberrantly spliced if it contained at least one significant splicing outlier reported by the aberrant splicing caller without applying any additional replication or rare variant filter (Extended Data Fig. 4a for FRASER). Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles. P values were computed using the paired one-sided Wilcoxon test. **b**, Precision-recall curves comparing the overall prediction performance on all GTEx tissues of the same models as in **a**, using FRASER as the outlier caller and the rare variant filter in Extended Data Fig. 4c with 250 bp together with different differential splicing cutoffs, namely |ΔΨ| = 0.1, 0.2, 0.3.

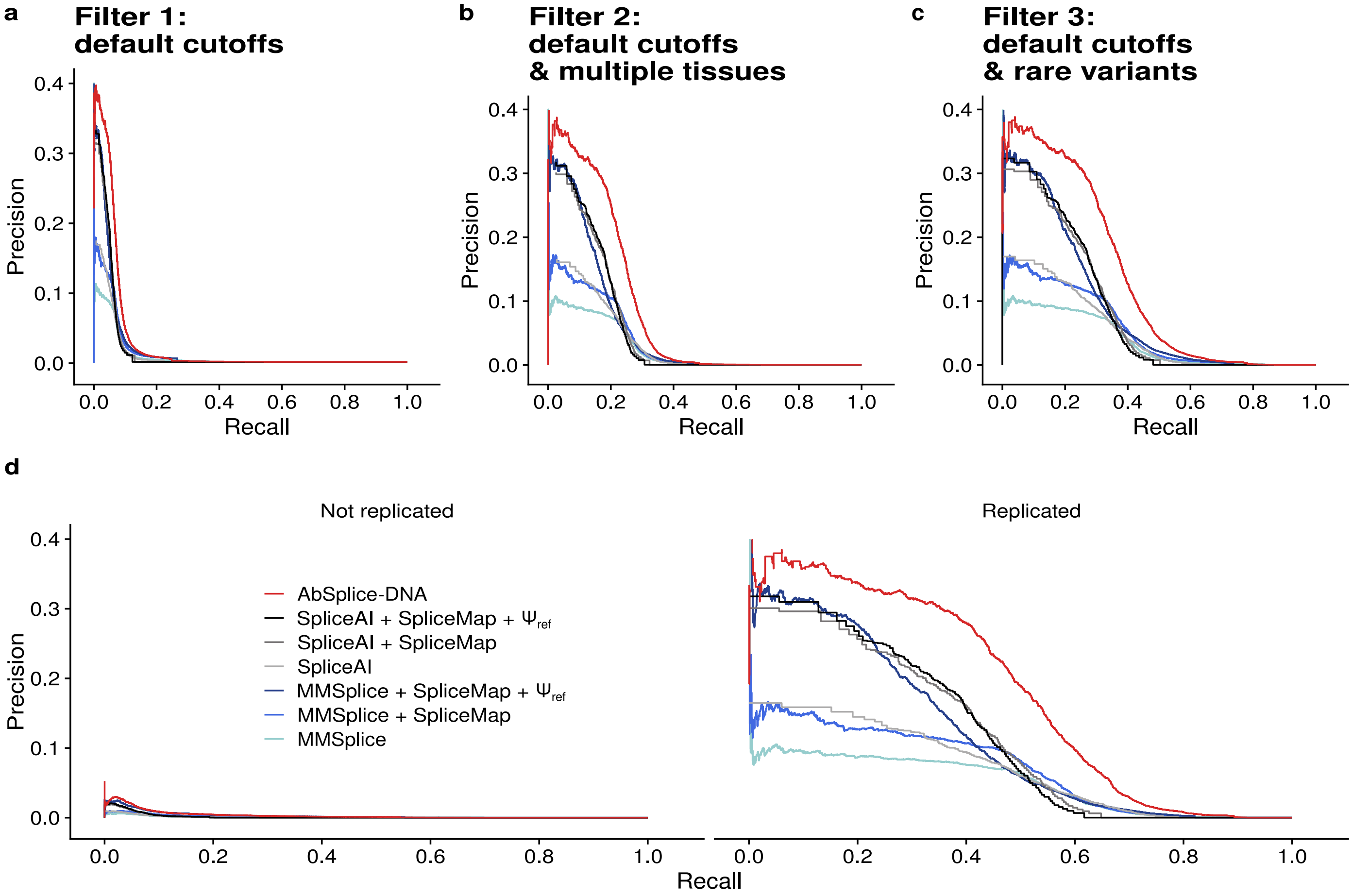


Extended Data Fig. 2 | Splicing outliers with a rare variant in the vicinity are enriched for replicated events. a, Enrichment of replicated splicing outliers across tissues with respect to the distance to the nearest rare variant. Note that there is an enrichment up to a distance of 250 bp. 'Number of tissues' denotes the minimum number of tissues from an individual with a shared splicing outlier such that the outlier is considered to be replicated. **b**, Replication rate of aberrant splicing events between tissues (n = 49) of a sample for all aberrant splicing events (red) compared with aberrant splicing events that contain a rare variant within a 250 bp window (blue). Filtering for aberrant splicing events with a rare variant reduces the amount of singletons probably by filtering out technical artifacts. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles. **c**, Percentage of singletons (aberrant splicing events that are observed only in one tissue) among all outliers (in red) and among outliers with a rare variant (in blue) for each tissue. There are nearly no replicated RNA-seq samples in the GTEx dataset. Therefore, among all singleton events, genuinely tissue-specific aberrant splicing events are hard to distinguish from non-reproducible technical artifacts.

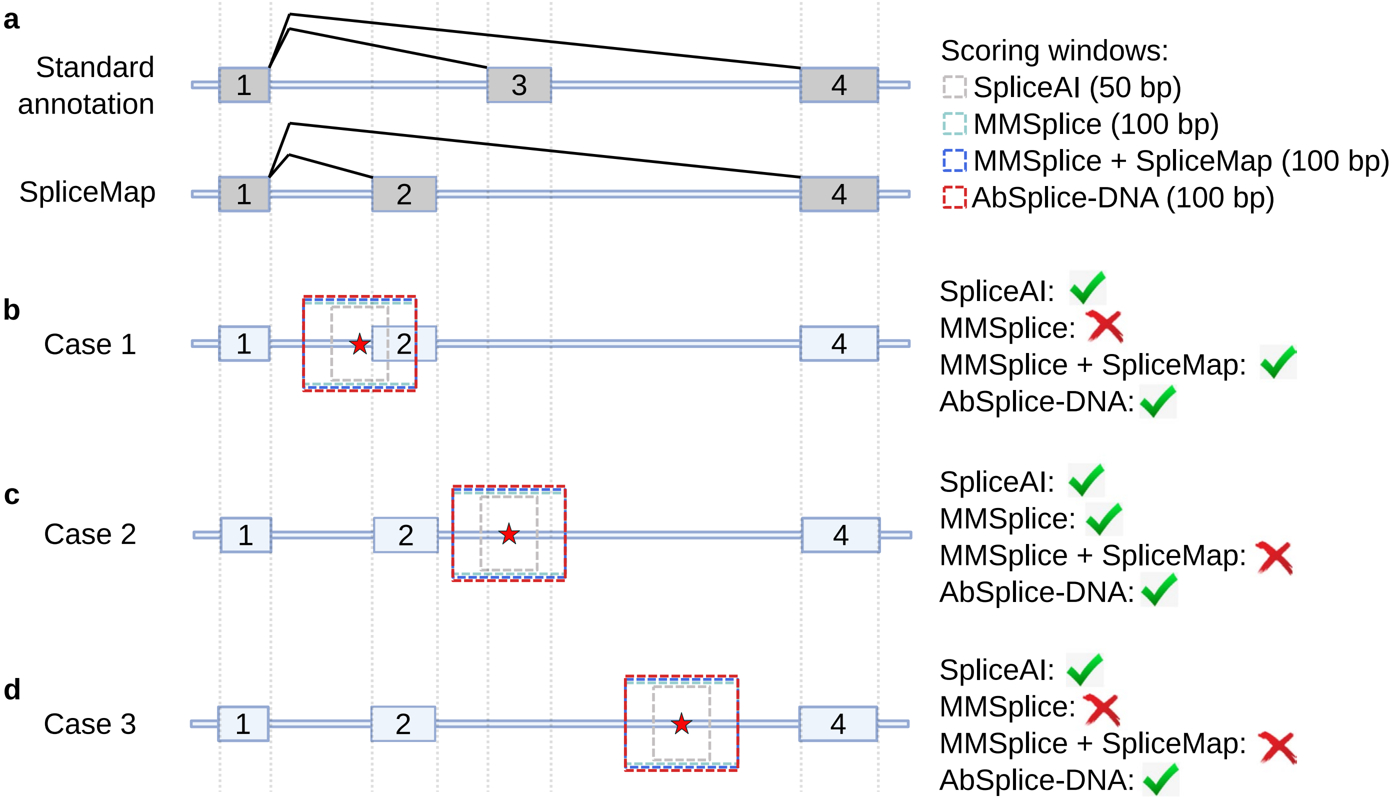


Extended Data Fig. 3 | See next page for caption.

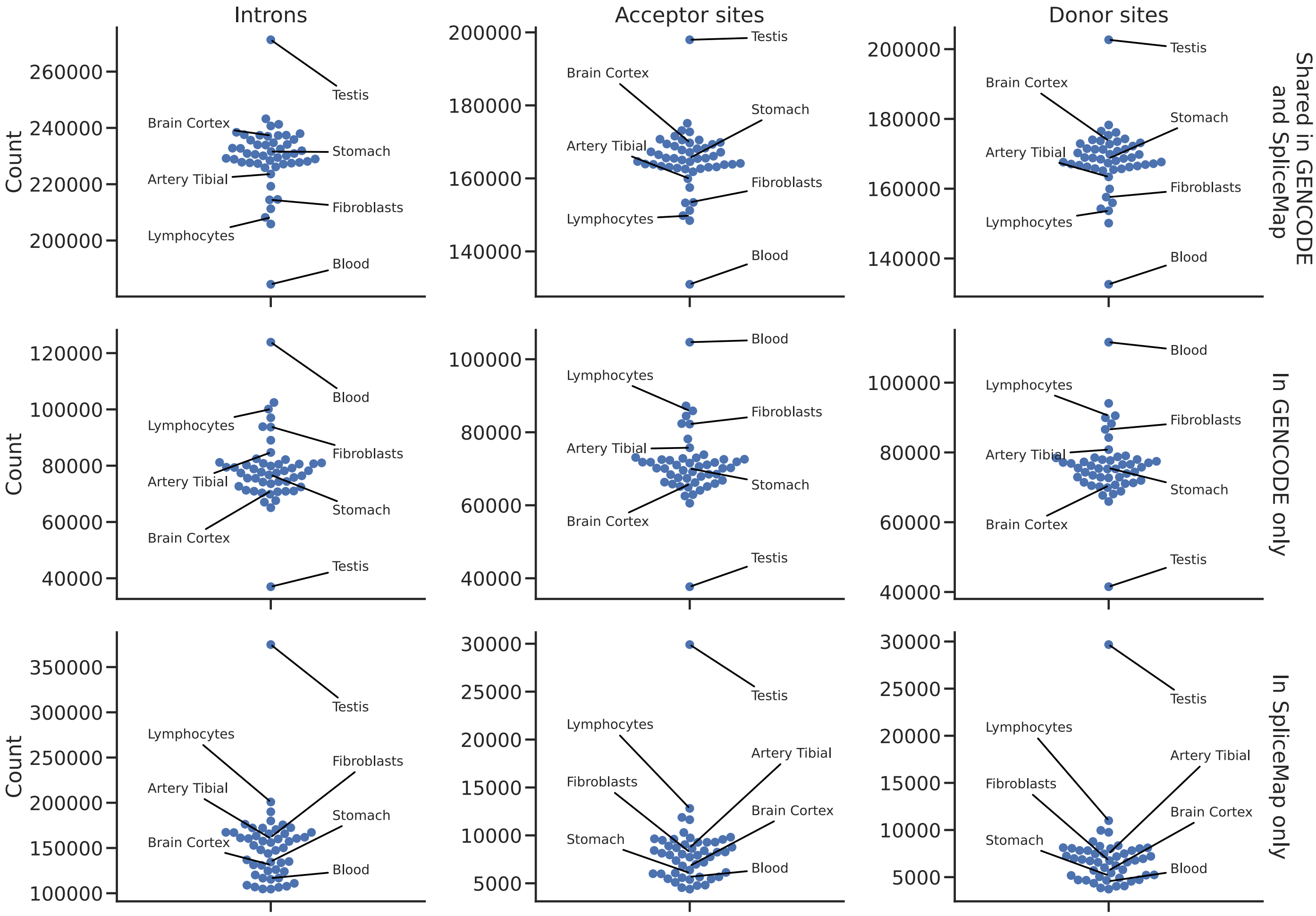
Extended Data Fig. 3 | Outlier filtering. Visualization of different cases for the rare variant outlier filter (corresponds to Filter 3 in Extended Data Fig. 4). **a**, Exons 1, 3 and 4 were annotated in SpliceMap. Exon 2 is a novel exon detected on an individual whose splice sites are not in SpliceMap. If there exists a rare variant within 250 bp of any splice site (in SpliceMap or not) that shares a junction with either the donor or acceptor site of the outlier event, the outlier passes the 'rare variant filter'. Cases 1 and 2: The individual has a rare variant within 250 bp of either the donor site of exon 1 or the acceptor site of exon 2, which are the splice sites of the outlier junction. Importantly, exon 2 was not quantified by SpliceMap, but the outlier filter solely depends on split reads. Case 3: The individual has a rare variant within 250 bp of the donor site of exon 2. However, this donor site is not part of the outlier event. Case 4: The individual has a rare variant within 250 bp of the acceptor site of exon 3, which forms a splicing junction with the donor site of exon 1. Case 5: The individual has two rare variants, one further than 250 bp of any splice site, the other within 250 bp of the acceptor site of exon 4. Notably, a variant can be far from the outlier junction and still be involved in the outlier event. **b**, Exon elongation detected as a splicing efficiency outlier. For splicing efficiency outliers, only the affected splice-site with altered splicing efficiency is considered for the variant filter. Case 1: The individual has a rare variant within 250 bp of the donor site of exon 1. Case 2: The individual has a rare variant that overlaps the acceptor site of the elongated exon 3, but is further than 250 bp from the acceptor site of exon 3. Case 3: The individual has a rare variant within 250 bp of the acceptor site of exon 3. Case 4: The individual has a rare variant within 250 bp of the donor site of exon 3, but the donor is not related to the exon elongation.



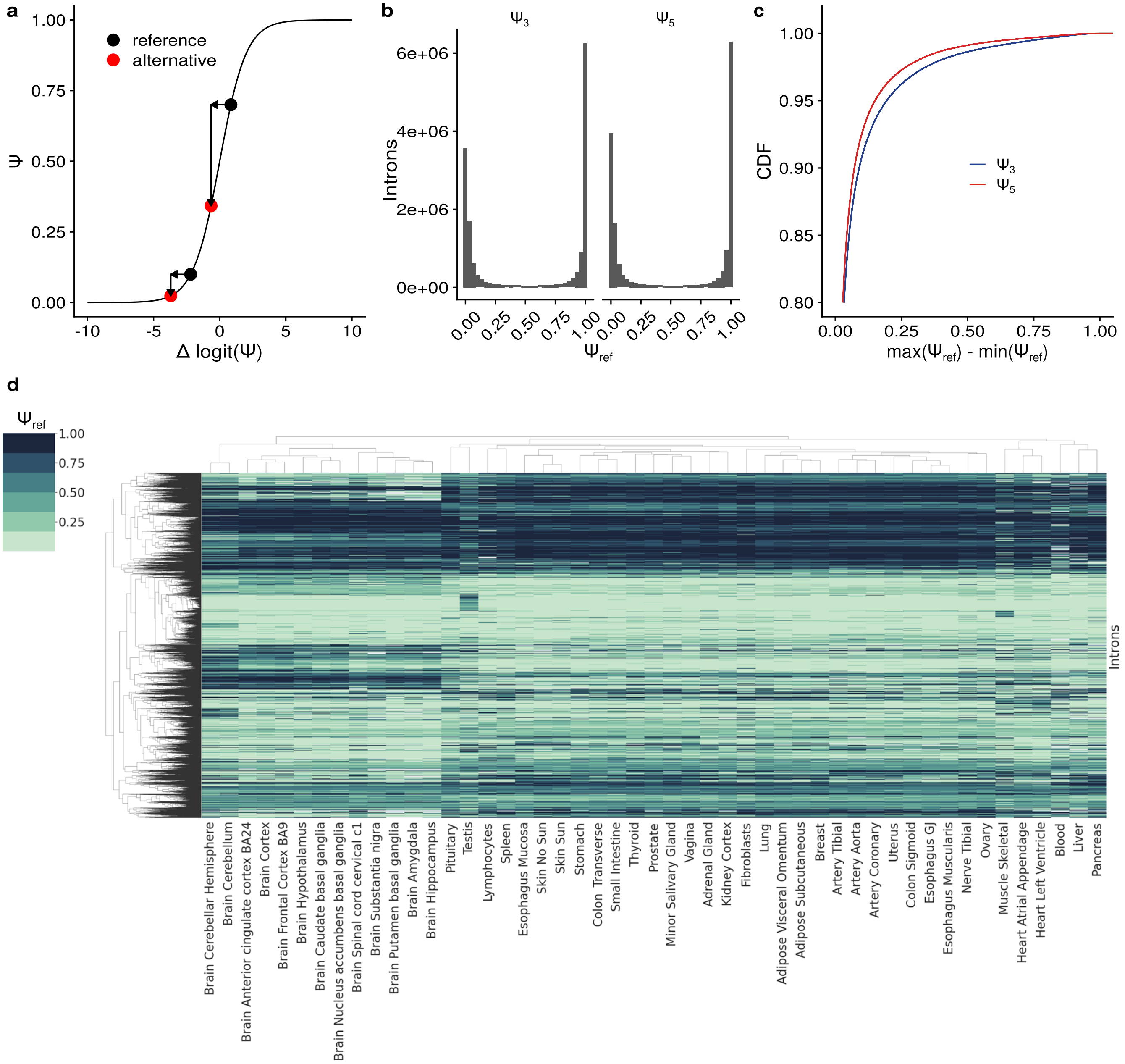
Extended Data Fig. 4 | Performance with different filters. Precision-recall curve comparing the overall prediction performance on all GTEx tissues of SpliceAI, SpliceAI using SpliceMap annotation, SpliceAI using SpliceMap annotation along with quantitative reference levels of splicing, MMSplice using GENCODE annotation, MMSplice using SpliceMap annotation, MMSplice using SpliceMap annotation along with quantitative reference levels of splicing, and the integrative model AbSplice-DNA, using different filters for aberrantly spliced genes. **a**, Filter 1: FRASER default cutoffs ($|\Delta\Psi| > 0.3$, FDR < 0.05, 126,308 aberrant events) **b**, Filter 2: same as **a**, but restricting to genes that are aberrantly spliced in at least two different tissues from the same individual (32,886 aberrant events). **c**, Filter 3: same as **a**, but restricting to genes that have a rare variant within 250 bp of the splice sites (22,766 aberrant events). While the results are best with Filter 3, the relative improvements in terms of precision at the same recall between the methods is the same as with Filter 2. In particular, having restricted to variants 250 bp away from any detected split read boundary (Filter 3) did not bias our analysis for the splice-site centric method MMSplice over SpliceAI. **d**, After applying Filter 3, outliers were stratified into 'replicated' (14,030 aberrant events), that is appearing in at least two different tissues of the same individual, and 'not replicated' (8,736 aberrant events). All models showed a significantly higher performance for aberrant splicing events replicated in two or more samples compared to those reported in a single sample only.



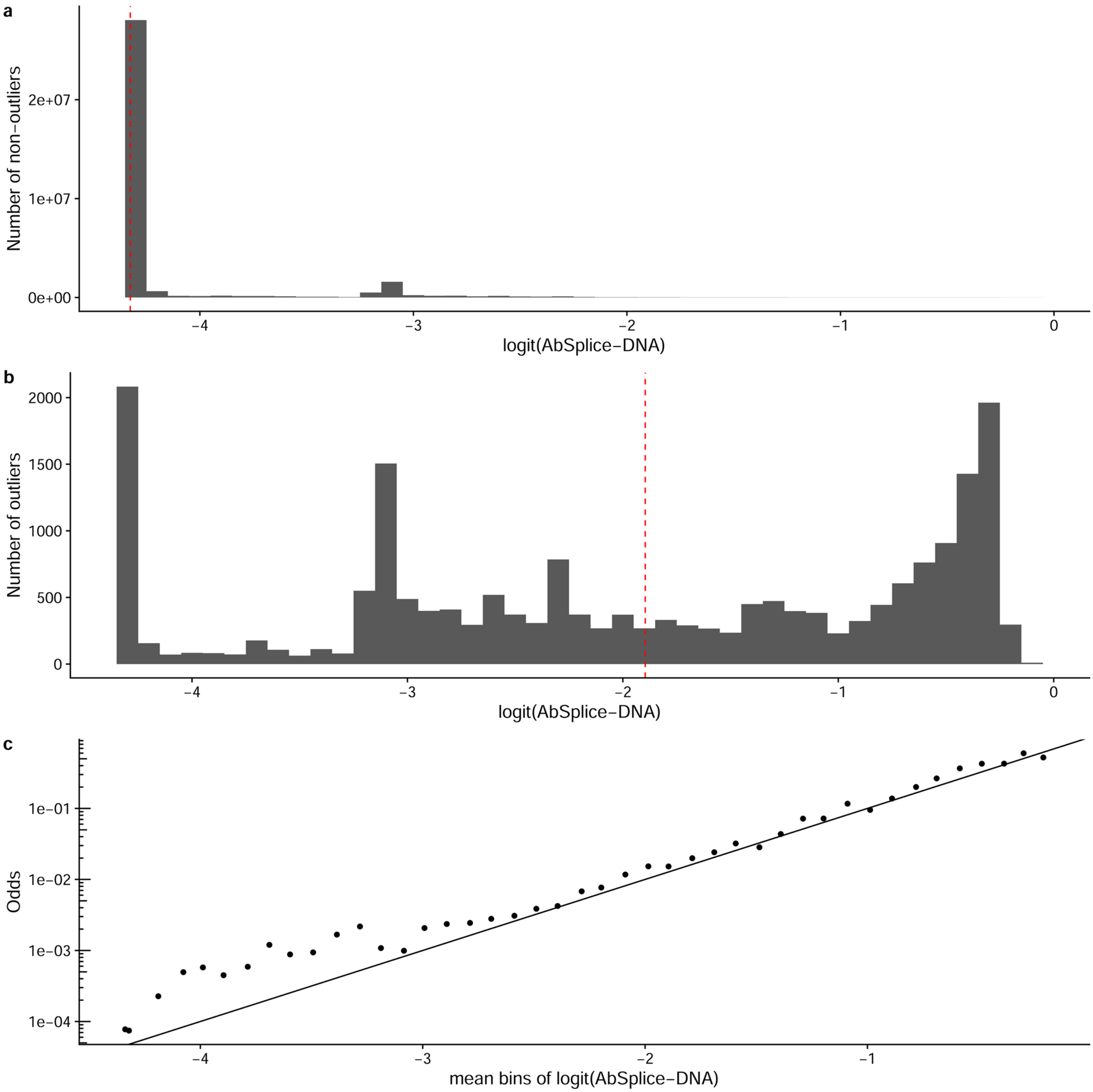
Extended Data Fig. 5 | Variant scoring of SpliceAI, MMSplice, MMSplice + SpliceMap and AbSplice-DNA. a, A gene model with 3 annotated exons in the standard annotation (1, 3 and 4) and 3 exons detected by SpliceMap (1, 2 and 4). SpliceAI scores for every bp in a 50 bp window of a variant (shown as red star) and reports the maximum score independent of the distance to a junction. MMSplice provides a score in a 100 bp window around a variant as long as there is a junction in that window. **b**, Case with a variant within 100 bp of an annotated junction in SpliceMap, but further than 100 bp from any exon in the standard annotation. MMSplice + SpliceMap is able to score the variant, while MMSplice is not. **c**, Case with a variant within 100 bp of an annotated exon in the standard annotation, but further than 100 bp from any exon in the SpliceMap. Therefore, MMSplice is able to score the variant, while MMSplice + SpliceMap is not. **d**, The variant is not within 100 bp of any annotated junction in the standard annotation or SpliceMap. Therefore neither MMSplice nor MMSplice + SpliceMap can score the variant. However, SpliceAI is always able to score a variant. Consequently, AbSplice is always able to score a variant.



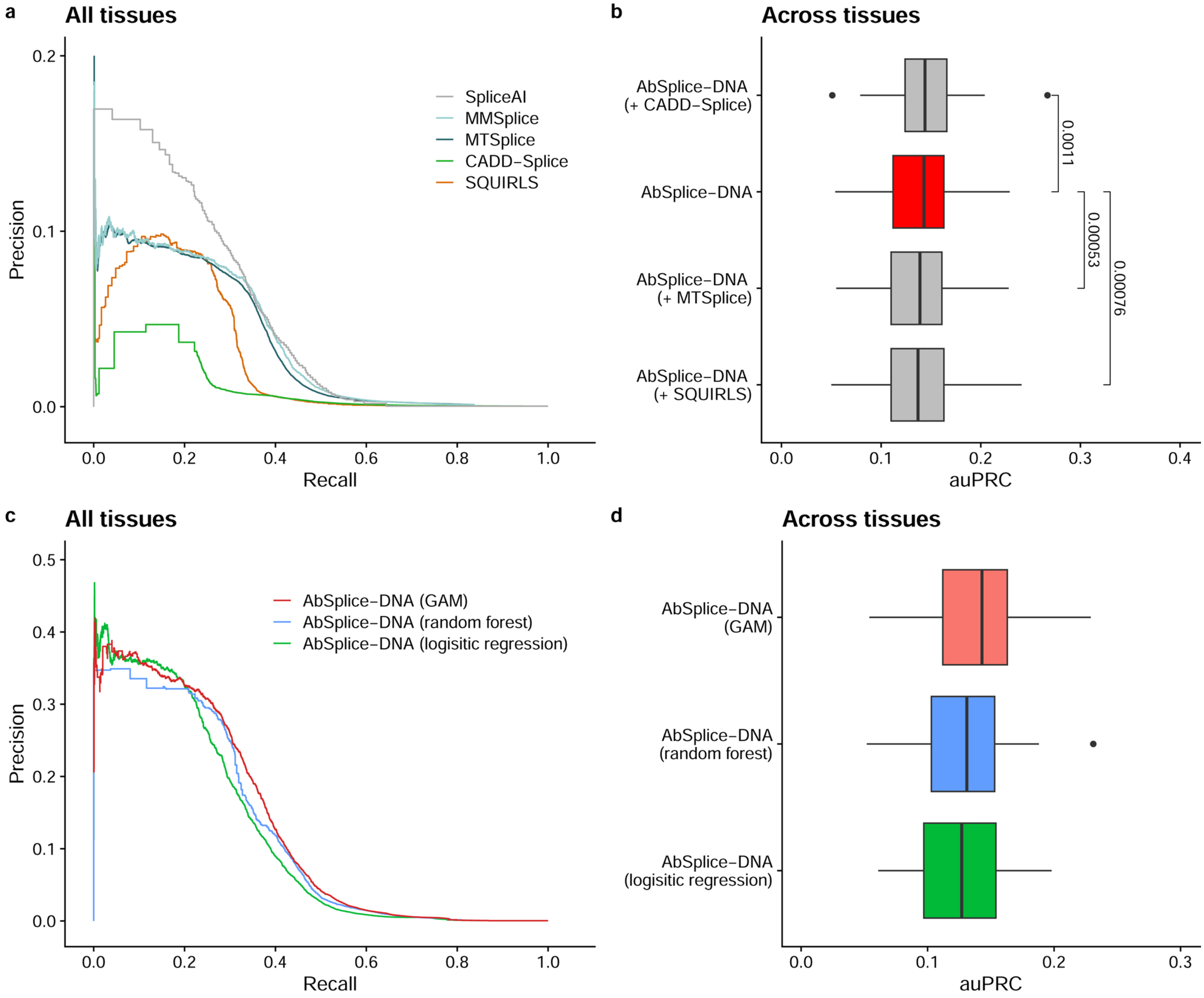
Extended Data Fig. 6 | Comparison of annotated splice-sites in SpliceMap and GENCODE. Number of introns, acceptor sites, and donor sites annotated in GENCODE and the SpliceMap of each GTEx tissue (first row), GENCODE only (second row) and SpliceMap only (third row).



Extended Data Fig. 7 | The variant effect depends on the reference isoform proportion. a, Ψ against Δlogit(Ψ) showing the non-linear splicing scaling law. The mutation effect of a variant can lead to different changes in Ψ in natural scale, depending on the reference splicing level of the intron. For example, the same variant can lead to a large change in Ψ if Ψ_{ref} is initially at an intermediate level and almost no change if Ψ_{ref} is initially at an extreme value (here low). **b**, Distribution of Ψ_{ref} in SpliceMap. Most of the introns are not alternatively spliced, so the reference level of those introns is either 0 or 1. **c**, Cumulative distribution function of the maximum difference of Ψ_{ref} (defined as: max(Ψ_{ref}) - min(Ψ_{ref})) across tissues per intron. **d**, Heatmap of the Ψ_{ref} of the most variable introns (defined as: max(Ψ_{ref}) - min(Ψ_{ref}) > 0.3) across tissues.

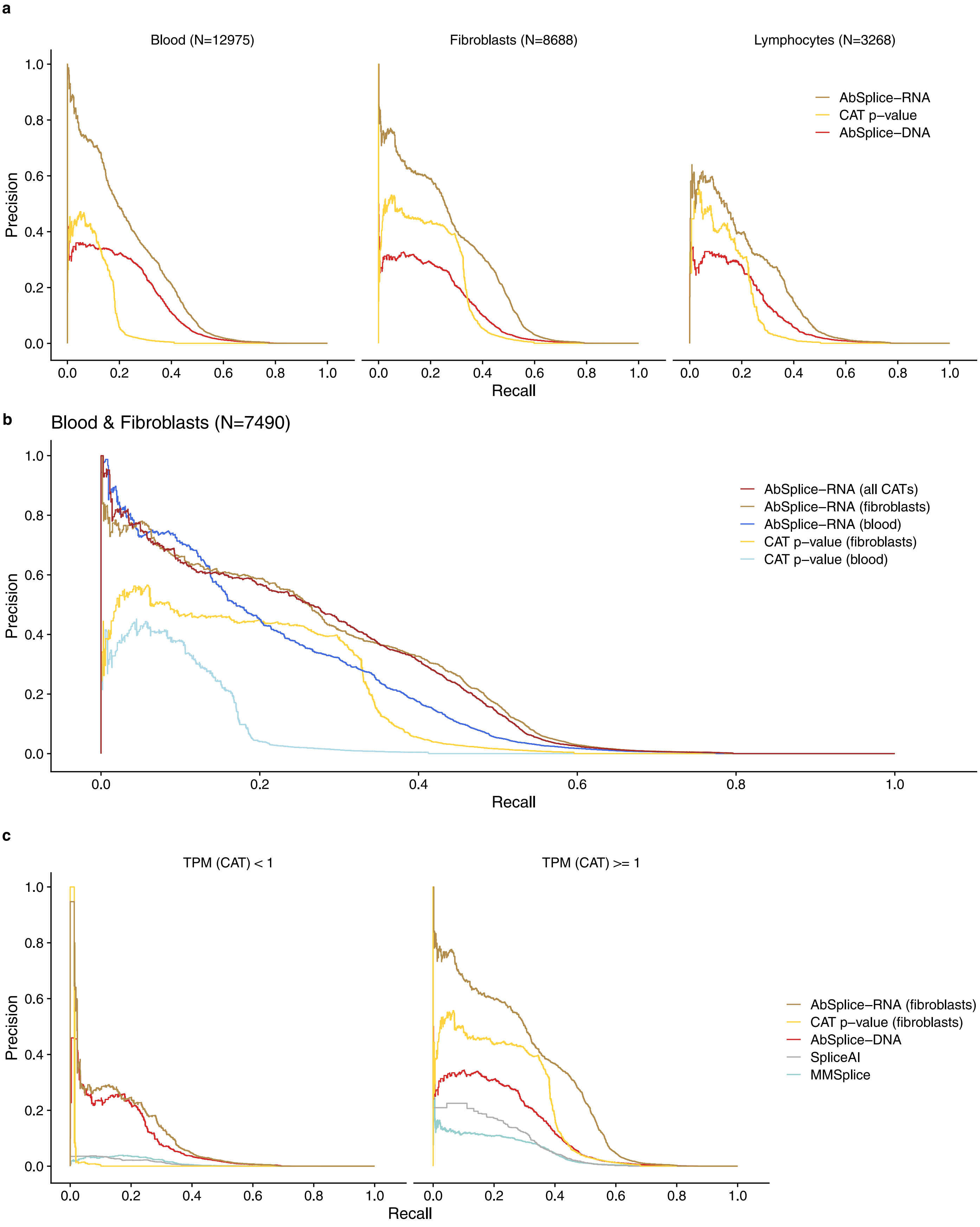


Extended Data Fig. 8 | Calibration of AbSplice-DNA. **a**, Histogram of AbSplice-DNA scores for gene, sample, tissue combinations that do not contain an aberrant splicing event. The dashed red line indicates the median. **b**, Histogram of AbSplice-DNA scores for gene, sample, tissue combinations that contain an aberrant splicing event. The peak at logit(AbSplice-DNA) ~−3.1 corresponds to AbSplice-DNA scores that are low due to small SpliceAI and MMSplice scores, but with an expressed splice site as annotated in SpliceMap. The peak at logit(AbSplice-DNA) ~−4.3 corresponds to small SpliceAI and MMSplice scores with an unused splice site as annotated in SpliceMap. **c**, Odds of aberrant splicing events as a function of logit transformed AbSplice-DNA scores (binned in bins of width 0.1). The line represents the diagonal. Note the linear relationship (especially in the high AbSplice-DNA score region) and the (extrapolated) intersection at AbSplice-DNA score of 0.5 (logit(AbSplice-DNA) = 0) corresponding to a log odds of 1, indicating a well calibrated model.



Extended Data Fig. 9 | Performance analysis of additional state-of-the art models and AbSplice-DNA trained with different model methods. **a**, Precision-recall performance of CADD-Splice, SQUIRLS, MTSplice, MMSplice and SpliceAI. **b**, Distribution of the area under the precision-recall curve (auPRC) across all GTEx tissues (n = 49) of the AbSplice-DNA models trained with varying feature sets using the models in **a**, that is 'AbSplice-DNA (+ CADD-Splice)' additionally used CADD-Splice scores during training. Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles. Shown in red is the AbSplice-DNA model used in the manuscript. Models are sorted by auPRC. *P*-values were computed using the paired two-sided Wilcoxon test. **c-d**, AbSplice-DNA was trained using a generalized additive model (GAM), random forest and logistic regression. AbSplice-DNA with GAM is the one used in the manuscript. **c**, Precision-recall curve across all GTEx tissues. **d**, Distribution of the area under the precision-recall curve of the models in **c** across tissues (n = 49). Center line, median; box limits, first and third quartiles; whiskers span all data within 1.5 interquartile ranges of the lower and upper quartiles.

https://doi.org/10.1038/s41588-023-01373-3



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | RNA-based predictions from CAT improve DNA-based scores. a, Precision-recall curves comparing the overall prediction performance on non-accessible GTEx tissues using the gene-level FRASER p-values from the CAT, AbSplice-RNA trained on a single CAT and AbSplice-DNA. Each panel shows a different CAT and the number of matching samples in the non-accessible tissues. **b**, Same as **a**, but for samples having RNA-seq from both blood and fibroblasts. AbSplice-RNA (all CATs) was trained using RNA-seq data from blood, fibroblasts and lymphocytes. Note that AbSplice-RNA (fibroblasts) gave a similar performance as AbSplice-RNA (all CATs). We did not restrict the samples to the ones also having lymphocytes as this would result in a low number of samples (N = 2,258). **c**, Model performance for genes not expressed or expressed in the clinically accessible tissue fibroblasts. The cutoff for calling a gene expressed was TPM > 1 (transcript per million). AbSplice-RNA improves for genes expressed in fibroblasts and remains on par with AbSplice-DNA for genes not expressed in fibroblasts.

nature portfolio

Corresponding author(s): Julien Gagneur

Last updated by author(s): 13.02.2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	*Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.*
Data analysis	https://github.com/gagneurlab/spliceai_rocksdb, https://github.com/gagneurlab/gnomad_rocksdb, https://github.com/gagneurlab/absplice, https://github.com/gagneurlab/splicemap, https://github.com/gagneurlab/AbSplice_analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Rare variants from gnomAD v3.1.2 are publicly available at https://gnomad.broadinstitute.org. The GTEx v8 dataset is available at (under dbGaP protection) https://gtexportal.org/home. The ALS dataset is available at dataportal.answerals.org after a registration and approval process. The mitochondrial dataset is described in

Yépez et al.29. Precomputed SpliceAI scores are publicly available at Illumina Basespace, https://basespace.illumina.com/s/otSPW8hnhaZR, after registration. SpliceMaps for all 49 GTEx tissues and iPSC-derived spinal motor neurons from ALS (hg38) are available at Zenodo DOI: 10.5281/zenodo.6387937. Precomputed AbSplice-DNA scores (hg38) in all 49 GTEx tissues are available at Zenodo DOI: 10.5281/zenodo.6408331. Due to privacy restrictions we will approach the GTEx consortium to share the benchmark dataset upon acceptance of the manuscript.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	No primary data was generated for this study. Person-related data was obtained through authorized access from primary data controllers. The study adheres to the ethical and research agreements between the Technical University of Munich and the primary data controllers. All participant informed consents were collected and rely by the primary data controllers.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to predetermine sample size. In the GTEx data we excluded tissues with less than 50 samples. In the ALS and mitochondrial disease datasets we did not exclude any samples.
Data exclusions	In the GTEx data we excluded tissues with less than 50 samples. In the ALS and mitochondrial disease datasets we did not exclude any samples.
Replication	We did not include replicates as they were not available.
Randomization	No randomization was performed as it is not relevant for this study.
Blinding	Not applicable. Blinding (in the statistical sense) is not relevant for our study. The AbSplice models are validated in 5-fold cross-validation and using independent experimental methods.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*

Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.*
Research sample	*Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.*
Sampling strategy	*Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data collection	*Describe the data collection procedure, including who recorded the data and how.*
Timing and spatial scale	*Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Reproducibility	*Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.*
Randomization	*Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.*
Blinding	*Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Antibodies

Antibodies used

Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.

Validation

Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.

Authentication

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

Mycoplasma contamination

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines
(See ICLAC register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Palaeontology and Archaeology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

□ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals

For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Reporting on sex

Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
□	□	Public health
□	□	National security
□	□	Crops and/or livestock
□	□	Ecosystems
□	□	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
□	□	Demonstrate how to render a vaccine ineffective
□	□	Confer resistance to therapeutically useful antibiotics or antiviral agents
□	□	Enhance the virulence of a pathogen or render a nonpathogen virulent
□	□	Increase transmissibility of a pathogen
□	□	Alter the host range of a pathogen
□	□	Enable evasion of diagnostic/detection modalities
□	□	Enable the weaponization of a biological agent or toxin
□	□	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

□ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

□ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

□ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

□ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

□ All plots are contour plots with outliers or pseudocolor plots.

□ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument

Identify the instrument used for data collection, specifying make and model number.

Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

□ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Indicate task or resting state; event-related or block design.

Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)

Specify: functional, structural, diffusion, perfusion.

Field strength

Specify in Tesla

Sequence & imaging parameters

Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.

Area of acquisition

State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI □ Used □ Not used

Preprocessing

Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: □ Whole brain □ ROI-based □ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
□	□ Functional and/or effective connectivity
□	□ Graph analysis
□	□ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.