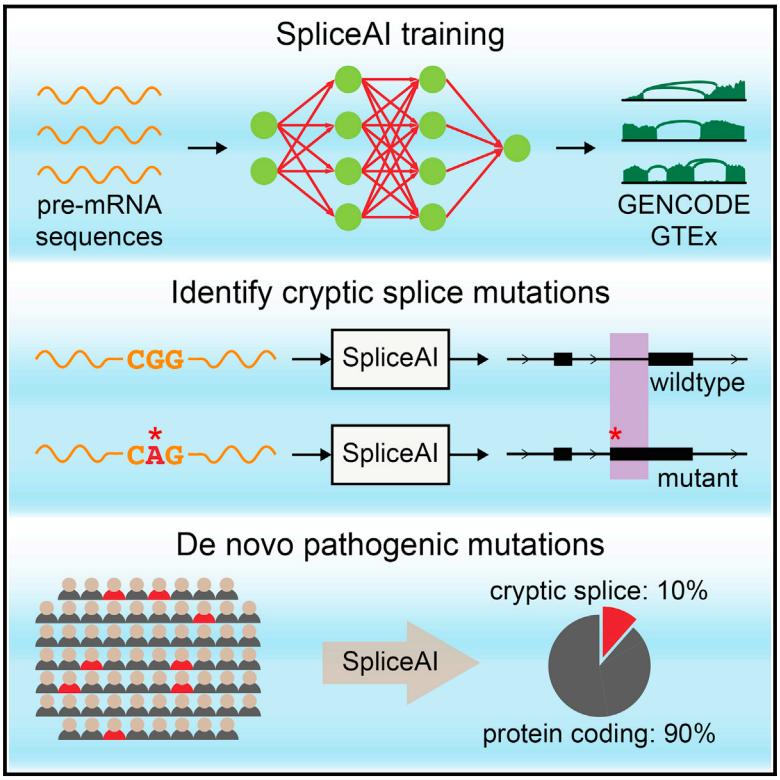
Article

Cell

Predicting Splicing from Primary Sequence with Deep Learning

Graphical Abstract



Highlights

- SpliceAI, a 32-layer deep neural network, predicts splicing from a pre-mRNA sequence
- 75% of predicted cryptic splice variants validate on RNA-seq
- Cryptic splicing may yield ~10% of pathogenic variants in neurodevelopmental disorders
- Cryptic splice variants frequently give rise to alternative splicing

Authors

Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, ..., Serafim Batzoglou, Stephan J. Sanders, Kyle Kai-How Farh

Correspondence

kfarh@illumina.com

In Brief

A deep neural network precisely models mRNA splicing from a genomic sequence and accurately predicts noncoding cryptic splice mutations in patients with rare genetic diseases.

Jaganathan et al., 2019, Cell *176*, 535–548
January 24, 2019
https://doi.org/10.1016/j.cell.2018.12.015

CellPress

Cell

Predicting Splicing from Primary Sequence with Deep Learning

Kishore Jaganathan,[1,6] Sofia Kyriazopoulou Panagiotopoulou,[1,6] Jeremy F. McRae,[1,6] Siavash Fazel Darbandi,[2] David Knowles,[3] Yang I. Li,[3] Jack A. Kosmicki,[1,4] Juan Arbelaez,[2] Wenwu Cui,[1] Grace B. Schwartz,[2] Eric D. Chow,[5] Efstathios Kanterakis,[1] Hong Gao,[1] Amirali Kia,[1] Serafim Batzoglou,[1] Stephan J. Sanders,[2] and Kyle Kai-How Farh[1,7,*]

[1]Illumina Artificial Intelligence Laboratory, Illumina, Inc., San Diego, CA, USA
[2]Department of Psychiatry, University of California, San Francisco, San Francisco, CA, USA
[3]Department of Genetics, Stanford University, Stanford, CA, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[5]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA
[6]These authors contributed equally
[7]Lead Contact
*Correspondence: kfarh@illumina.com
https://doi.org/10.1016/j.cell.2018.12.015

SUMMARY

The splicing of pre-mRNAs into mature transcripts is remarkable for its precision, but the mechanisms by which the cellular machinery achieves such specificity are incompletely understood. Here, we describe a deep neural network that accurately predicts splice junctions from an arbitrary pre-mRNA transcript sequence, enabling precise prediction of noncoding genetic variants that cause cryptic splicing. Synonymous and intronic mutations with predicted splice-altering consequence validate at a high rate on RNA-seq and are strongly deleterious in the human population. *De novo* mutations with predicted splice-altering consequence are significantly enriched in patients with autism and intellectual disability compared to healthy controls and validate against RNA-seq in 21 out of 28 of these patients. We estimate that 9%–11% of pathogenic mutations in patients with rare genetic disorders are caused by this previously underappreciated class of disease variation.

INTRODUCTION

Exome sequencing has transformed the clinical diagnosis of patients and families with rare genetic disorders and, when employed as a first-line test, significantly reduces the time and costs of the diagnostic odyssey (Tan et al., 2017). However, the diagnostic yield of exome sequencing is ~25%–30% in rare genetic disease cohorts, leaving the majority of patients without a diagnosis even after combined exome and microarray testing (Lee et al., 2014; Yang et al., 2014). Noncoding regions play a significant role in gene regulation and account for 90% of causal disease loci discovered in unbiased genome-wide association studies of human complex diseases (Farh et al., 2015; Maurano et al., 2012), suggesting that penetrant noncoding variants may also account for a significant burden of causal mutations in rare genetic diseases. Indeed, penetrant noncoding variants that disrupt the normal pattern of mRNA splicing despite lying outside the essential GT and AG splice dinucleotides, often referred to as cryptic splice variants, have long been recognized to play a significant role in rare genetic diseases (Cooper et al., 2009). However, cryptic splice mutations are often overlooked in clinical practice, due to our incomplete understanding of the splicing code and the resulting difficulty in accurately identifying splice-altering variants outside the essential GT and AG dinucleotides (Wang and Burge, 2008).

Recently, RNA sequencing (RNA-seq) has emerged as a promising assay for detecting splicing abnormalities in Mendelian disorders (Cummings et al., 2017), but thus far its utility in a clinical setting remains limited to a minority of cases where the relevant cell type is known and accessible to biopsy. High-throughput screening assays of potential splice-altering variants (Soemedi et al., 2017) have expanded the characterization of splicing variation but are less practical for evaluating random *de novo* mutations in genetic diseases, since the genomic space where splice-altering mutations may occur is extremely large. General prediction of splicing from an arbitrary pre-mRNA sequence would potentially allow precise prediction of the splice-altering consequences of noncoding variants, substantially improving diagnosis in patients with genetic diseases. To date, a general predictive model of splicing from a raw sequence that approaches the specificity of the spliceosome remains elusive, despite progress in specific applications, such as modeling the sequence characteristics of the core splicing motifs (Yeo and Burge, 2004), characterizing exonic splice enhancers and silencers (Fairbrother et al., 2002; Wang et al., 2004), and predicting cassette exon inclusion (Xiong et al., 2015).

RESULTS

Accurate Prediction of Splicing from a Primary Sequence Using Deep Learning

We constructed SpliceAI, a deep residual neural network (He et al., 2016) that predicts whether each position in a pre-mRNA

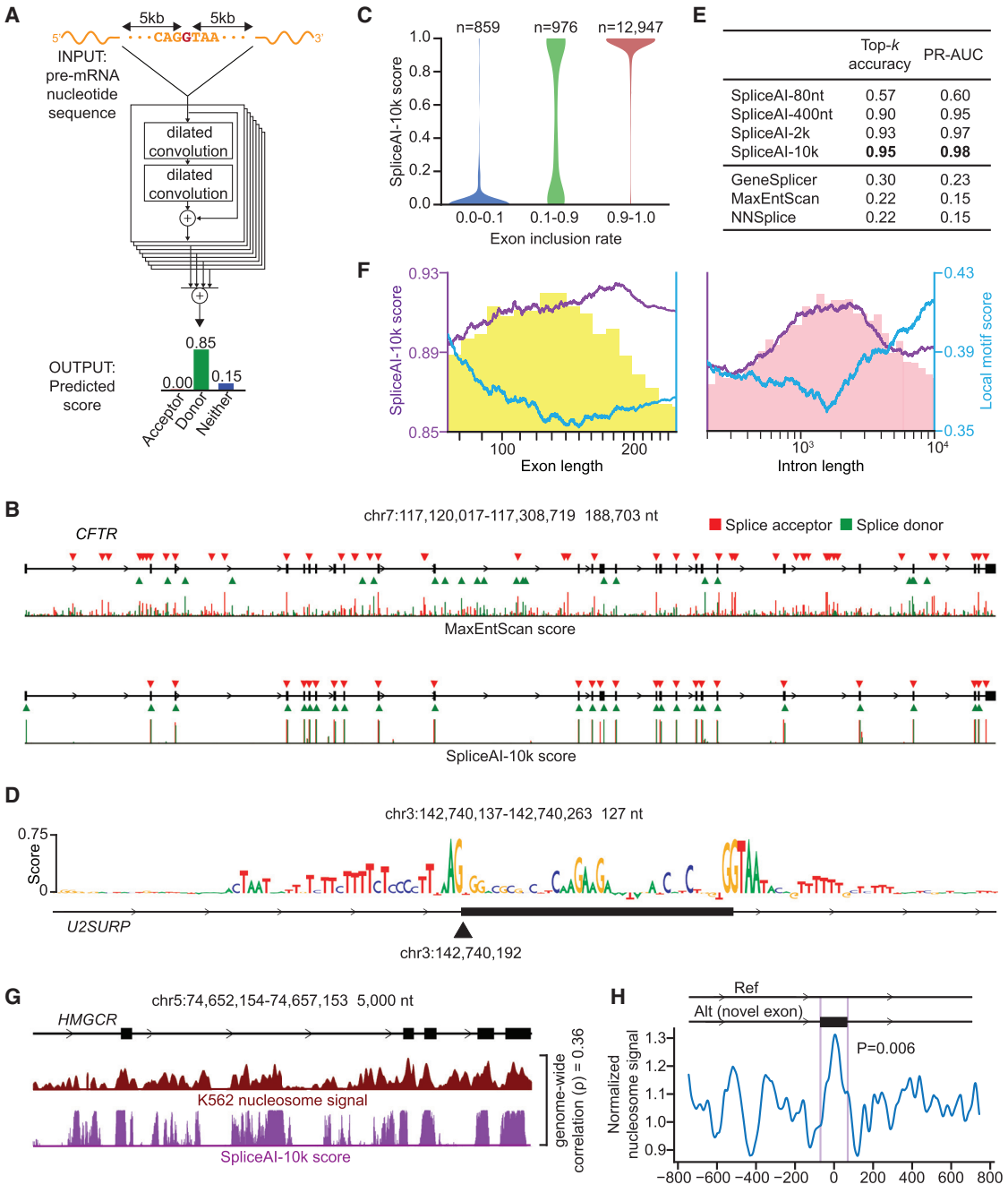


Figure 1. Predicting Splicing from Primary Sequence with Deep Learning

(A) For each position in the pre-mRNA transcript, SpliceAI-10k uses 10,000 nucleotides of flanking sequence as input and predicts whether that position is a splice acceptor, splice donor, or neither.

(B) The full pre-mRNA transcript for the *CFTR* gene scored using MaxEntScan (top) and SpliceAI-10k (bottom) is shown, along with predicted acceptor (red arrows) and donor (green arrows) sites and the actual positions of the exons (black boxes). For each method, we applied the threshold that made the number of predicted sites equal to the total number of actual sites.

(C) We measured the inclusion rate of each exon on RNA-seq and show the SpliceAI-10k score distribution for exons at different inclusion rates. Shown are the maximum of the exon's acceptor and donor scores.

(D) Impact of *in silico* mutating each nucleotide around exon 9 in the *U2SURP* gene. The vertical size of each nucleotide shows the decrease in the predicted strength of the acceptor site (black arrow) when that nucleotide is mutated (Δ score).

(E) Effect of the size of the input sequence context on the accuracy of the network. Top-*k* accuracy is the fraction of correctly predicted splice sites at the threshold where the number of predicted sites is equal to the actual number of sites present. PR-AUC is the area under the precision-recall curve. We also show the top-*k* accuracy and PR-AUC for three other algorithms for splice-site detection.

(legend continued on next page)

transcript is a splice donor, splice acceptor, or neither (Figures 1A and S1), using as input only the genomic sequence of the pre-mRNA transcript. Because splice donors and splice acceptors may be separated by tens of thousands of nucleotides, we employed a network architecture consisting of 32 dilated convolutional layers that can recognize sequence determinants spanning very large genomic distances. In contrast to previous methods that have only considered short nucleotide windows adjoining exon-intron boundaries (Yeo and Burge, 2004), or relied on human-engineered features, our neural network learns splicing determinants directly from the primary sequence by evaluating 10,000 nucleotides of the flanking context sequence to predict the splice function of each position in the pre-mRNA transcript.

We used GENCODE-annotated pre-mRNA transcript sequences (Harrow et al., 2012) on a subset of the human chromosomes to train the parameters of the neural network, and transcripts on the remaining chromosomes, with paralogs excluded, to test the network's predictions. For pre-mRNA transcripts in the test dataset, the network predicts splice junctions with 95% top-*k* accuracy, which is the fraction of correctly predicted splice sites at the threshold where the number of predicted sites is equal to the actual number of splice sites present in the test dataset (Boyd et al., 2012; Yeo and Burge, 2004). Even genes in excess of 100 kb such as *CFTR* are often reconstructed perfectly to nucleotide precision (Figure 1B). To confirm that the network is not simply relying on exonic sequence biases, we also tested the network on long noncoding RNAs. Despite the incompleteness of noncoding transcript annotations, which is expected to reduce our accuracy, the network predicts known splice junctions in long noncoding RNAs (lincRNAs) with 84% top-*k* accuracy (Figures S2A and S2B), indicating that it can approximate the behavior of the spliceosome on arbitrary sequences that are free from protein-coding selective pressures.

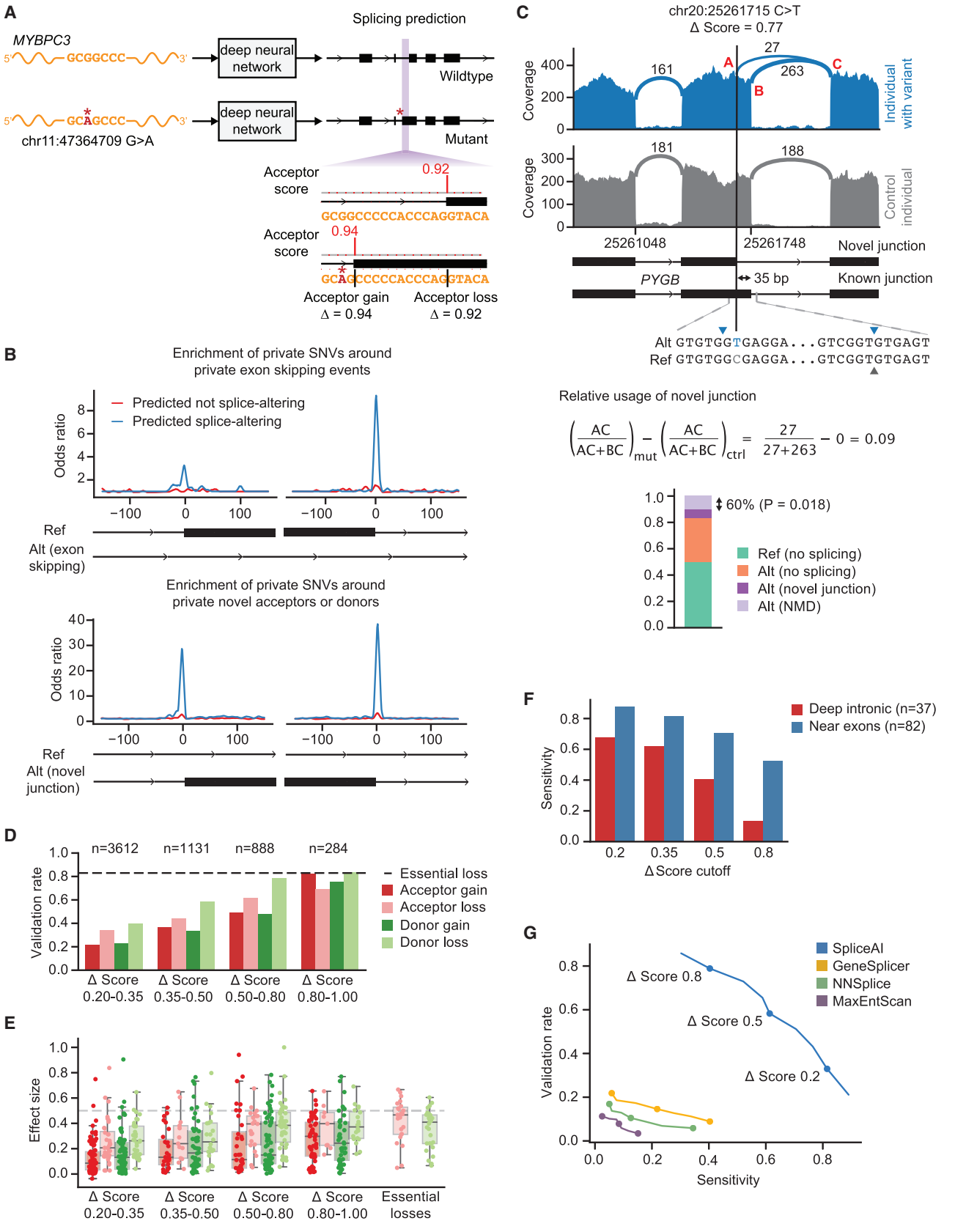
For each GENCODE-annotated exon in the test dataset (excluding the first and last exons of each gene), we also examined whether the network's prediction scores correlate with the fraction of reads supporting exon inclusion versus exon skipping, based on RNA-seq data from the Gene and Tissue Expression (GTEx) atlas (The GTEx Consortium et al., 2015; Lonsdale et al., 2013) (Figure 1C). Exons that were constitutively spliced in or spliced out across GTEx tissues had prediction scores that were close to 1 or 0, respectively, whereas exons that underwent a substantial degree of alternative splicing (between 10% and 90% exon inclusion averaged across samples) tended toward intermediate scores (Pearson correlation = 0.78, $p \approx 0$).

We next sought to understand the sequence determinants utilized by the network to achieve its remarkable accuracy. We performed systematic *in silico* substitutions of each nucleotide near annotated exons, measuring the effects on the network's prediction scores at the adjoining splice sites (Figure 1D). We found that disrupting the sequence of a splice donor motif frequently caused the network to predict that the upstream splice acceptor site will also be lost, as is observed with exon-skipping events *in vivo*, indicating that a significant degree of specificity is imparted by exon definition between a paired upstream acceptor motif and a downstream donor motif set at an optimal distance (Berget, 1995). Additional motifs that contribute to the splicing signal include the well-characterized binding motifs of the branchpoint and the SR-protein family (Figures S2C and S2D) (Fairbrother et al., 2002). The effects of these motifs are highly dependent on their position in the exon, suggesting that their roles include specifying the precise positioning of intron-exon boundaries by differentiating between competing acceptor and donor sites.

Training the network with varying input sequence context markedly impacts the accuracy of the splice predictions (Figure 1E), indicating that long-range sequence determinants thousands of nucleotides away from the splice site are essential for discerning functional splice junctions from the large number of nonfunctional sites with near-optimal motifs. To examine long-range and short-range specificity determinants, we compared the scores assigned to annotated junctions by a model trained on 80 nt of the sequence context (SpliceAI-80nt) versus the full model that is trained on 10,000 nt of context (SpliceAI-10k). The network trained on 80 nt of the sequence context assigns lower scores to junctions adjoining exons or introns of typical length (150 nt for exons, ~1,000 nt for introns) (Figure 1F), in agreement with earlier observations that such sites tend to have weaker splice motifs compared to the splice sites of exons and introns, which are unusually long or short (Amit et al., 2012). In contrast, the network trained on 10,000 nt of the sequence context shows preference for introns and exons of average length, despite their weaker splice motifs, because it can account for long-range specificity conferred by exon or intron length. The skipping of weaker motifs in long uninterrupted introns is consistent with the faster RNA polymerase II elongation experimentally observed in the absence of exon pausing, which may allow the spliceosome less time to recognize suboptimal motifs (Close et al., 2012; Jonkers et al., 2014). Our findings suggest that the average splice junction possesses favorable long-range sequence determinants that confer substantial specificity, explaining the high degree of sequence degeneracy tolerated at most splice motifs.

Because splicing occurs co-transcriptionally, interactions between chromatin state and co-transcriptional splicing might also guide exon definition (Luco et al., 2010) and have the potential to be utilized by the network to the extent that chromatin state is predictable from the primary sequence. In particular, genome-wide studies of nucleosome positioning have shown that

(F) Relationship between exon-intron length and the strength of the adjoining splice sites, as predicted by SpliceAI-80 nt (local motif score) and SpliceAI-10k. The genome-wide distributions of exon length (yellow) and intron length (pink) are shown in the background. The x axis is in log-scale.
(G) A pair of splice acceptor and donor motifs, placed 150 nt apart, are walked along the *HMGCR* gene. Shown are, at each position, K562 nucleosome signal and the likelihood of the pair forming an exon at that position, as predicted by SpliceAI-10k. The genome-wide Spearman correlation between the two tracks is shown.
(H) Average K562 and GM12878 nucleosome signal near private mutations that are predicted by the SpliceAI-10k model to create novel exons in the GTEx cohort. The p value by permutation test is shown.
See also Figures S1 and S2.



(legend on next page)

nucleosome occupancy is higher in exons (Schwartz et al., 2009; Spies et al., 2009). To test whether the network uses sequence determinants of nucleosome positioning for splice site prediction, we walked a pair of optimal acceptor and donor motifs separated by 150 nt (roughly the size of the average exon) across the genome and asked the network to predict whether the pair of motifs would result in exon inclusion at that locus (Figure 1G). We find that positions predicted to be favorable for exon inclusion correlated with positions of high nucleosome occupancy, even in intergenic regions (Spearman correlation = 0.36, $p \approx 0$), and this effect persists after controlling for GC content (Figure S2E). These results suggest that the network has implicitly learned to predict nucleosome positioning from the primary sequence and utilizes it as a specificity determinant in exon definition. Similar to exons and introns of average length, exons positioned over nucleosomes have weaker local splice motifs (Figure S2F), consistent with greater tolerance for degenerate motifs in the presence of compensatory factors (Spies et al., 2009).

Although multiple studies have reported a correlation between exons and nucleosome occupancy, a causal role for nucleosome positioning in exon definition has not been firmly established. Using data from 149 individuals with both RNA-seq and whole-genome sequencing from the Genotype-Tissue Expression (GTEx) cohort (The GTEx Consortium et al., 2015; Lonsdale et al., 2013), we identified novel exons that were private to a single individual, and corresponded to a private splice site-creating genetic mutation. These private exon-creation events were significantly associated with existing nucleosome positioning in K562 and GM12878 cells ($p = 0.006$ by permutation test, Figure 1H), even though these cell lines most likely lack the corresponding private genetic mutations. Our results indicate that genetic variants are more likely to trigger creation of a novel exon if the resulting novel exon would overlay a region of existing nucleosome occupancy, supporting a causal role for nucleosome positioning in promoting exon definition.

Verification of Predicted Cryptic Splice Mutations in RNA-Seq Data

We extended the deep learning network to the evaluation of genetic variants for splice-altering function by predicting exon-intron boundaries for both the reference pre-mRNA transcript sequence and the alternate transcript sequence containing the variant, and taking the difference between the scores (Δ score, Figure 2A). Importantly, the network was only trained on reference transcript sequences and splice junction annotations, and never saw variant data during training, making prediction of variant effects a challenging test of the network's ability to accurately model the sequence determinants of splicing.

We looked for the effects of cryptic splice variants in RNA-seq data in the GTEx cohort, comprising 149 individuals with both whole-genome sequencing and RNA-seq from multiple tissues. To approximate the scenario encountered in rare disease sequencing, we first focused on rare, private mutations (present in only one individual in the GTEx cohort). We find that private mutations predicted to have functional consequences by the neural network are strongly enriched at private novel splice junctions and at the boundaries of skipped-over exons in private exon-skipping events (Figure 2B), suggesting that a large fraction of these predictions are functional.

To quantify the effects of splice-site creating variants on the relative production of normal and aberrant splice isoforms, we measured the number of reads supporting the novel splice event as a fraction of the total number of reads covering the site (Figure 2C). We calculated both the decrease in the fraction of reads that spliced at the disrupted junction and the increase in the

Figure 2. Validation of Rare Cryptic Splice Mutations in RNA-Seq Data

(A) To assess the splice-altering impact of a mutation, SpliceAI-10k predicts acceptor and donor scores at each position in the pre-mRNA sequence of the gene with and without the mutation, as shown here for rs397515893, a pathogenic cryptic splice variant in the *MYBPC3* intron associated with cardiomyopathy. The Δ score value for the mutation is the largest change in splice prediction scores within 50 nt from the variant.

(B) We scored private genetic variants (observed in only one out of 149 individuals in the GTEx cohort) with SpliceAI-10k. Shown are the enrichment of private variants predicted to alter splicing (Δ score >0.2, blue) or to have no effect on splicing (Δ score <0.01, red) in the vicinity of private exon-skipping junctions (top) or private acceptor and donor sites (bottom). The y axis shows the number of times a private splice event and a nearby private genetic variant co-occur in the same individual, compared to expected numbers obtained through permutations.

(C) Example of a heterozygous synonymous variant in the *PYGB* gene that creates a novel donor site with incomplete penetrance. RNA-seq coverage, junction read counts, and the positions of junctions (blue and gray arrows) are shown for the individual with the variant and a control individual. The effect size is computed as the difference in the usage of the novel junction (AC) between individuals with the variant and individuals without the variant. In the stacked bar graph below, we show the number of reads with the reference or alternate allele that used the annotated or the novel junction ("no splicing" and "novel junction" respectively). The total number of reference reads differed significantly from the total number of alternate reads ($p = 0.018$, Binomial test), suggesting that 60% of transcripts splicing at the novel junction are missing in the RNA-seq data, presumably due to nonsense mediated decay (NMD).

(D) Fraction of cryptic splice mutations predicted by SpliceAI-10k that validated against the GTEx RNA-seq data. The validation rate of disruptions of essential acceptor or donor dinucleotides (dashed line) is less than 100% due to insufficient coverage and nonsense mediated decay.

(E) Distribution of effect sizes for validated cryptic splice predictions. The dashed line (50%) corresponds to the expected effect size of fully penetrant heterozygous variants. The measured effect size of essential acceptor or donor dinucleotide disruptions is less than 50% due to nonsense-mediated decay or unaccounted-for isoform changes.

(F) Sensitivity of SpliceAI-10k at detecting splice-altering private variants in the GTEx cohort at different Δ score cutoffs. Variants are split into deep intronic variants (>50 nt from exons) and variants near exons (overlapping exons or ≤50 nt from exon-intron boundaries).

(G) Validation rate and sensitivity of SpliceAI-10k and three other methods for splice site prediction at different confidence cutoffs. The three dots on the SpliceAI-10k curve show the performance of SpliceAI-10k at Δ score cutoffs of 0.2, 0.5, and 0.8. For the other three algorithms, the three dots on the curve indicate their performance at the thresholds where they predict the same number of cryptic splice variants as SpliceAI-10k at Δ score cutoffs of 0.2, 0.5, and 0.8.

See also Figures S3 and S4 and Tables S1 and S2.

fraction of reads that skipped the exon, taking the larger of the two effects (Figure S3A; STAR Methods).

Confidently predicted cryptic splice variants (Δ score $\geq$0.5) validate on RNA-seq at three-quarters the rate of essential GT or AG splice disruptions (Figure 2D). Both the validation rate and effect size of cryptic splice variants closely track their Δ scores (Figures 2D and 2E), demonstrating that the model's prediction score is a good proxy for the splice-altering potential of a variant. Validated variants, especially those with lower scores (Δ score <0.5), are often incompletely penetrant, and result in alternative splicing with production of a mixture of both aberrant and normal transcripts in the RNA-seq data (Figure 2E). Our estimates of validation rates and effect sizes are conservative and likely underestimate the true values, due to both unaccounted-for splice isoform changes and nonsense-mediated decay (Figures 2C and S3A). This is evidenced by the average effect sizes of variants that disrupt essential GT and AG splice dinucleotides being less than the 50% expected for fully penetrant heterozygous variants.

For cryptic splice variants that produce aberrant splice isoforms in at least three-tenths of the observed copies of the mRNA transcript, the network has a sensitivity of 71% when the variant is near exons, and 41% when the variant is in the deep intronic sequence (Δ score $\geq$0.5, Figure 2F). These findings indicate that deep intronic variants are more challenging to predict, possibly because deep intronic regions contain fewer of the specificity determinants that have been selected to be present near exons.

To benchmark the performance of our network against existing methods, we selected three popular classifiers that have been referenced in the literature for rare genetic disease diagnosis, GeneSplicer (Pertea et al., 2001), MaxEntScan (Yeo and Burge, 2004), and NNSplice (Reese et al., 1997), and plotted the RNA-seq validation rate and sensitivity at varying thresholds (Figure 2G). As has been the experience of others in the field (Cummings et al., 2017), we find that existing classifiers have insufficient specificity given the very large number of noncoding variants genome-wide that can possibly affect splicing, presumably because they focus on local motifs and largely do not account for long-range specificity determinants.

Given the large gap in performance compared with existing methods, we performed additional controls to exclude the possibility that our results in the RNA-seq data could be confounded by overfitting. First, we repeated the validation and sensitivity analyses separately for private variants and variants present in more than one individual in the GTEx cohort (Figures S3B–S3D) and verified that, at the same Δ score thresholds, private and common variants show no significant differences in their validation rate ($p > 0.05$, Fisher's exact test). Second, we also saw no significant differences in the validation rates of cryptic splice variants that create new GT or AG dinucleotides, variants that affect the extended acceptor or donor motif, and variants that occur in more distal regions ($p > 0.3$ χ^2 test of uniformity and $p > 0.3$ Mann-Whitney U test, respectively, Figures S3E and S3F). Third, we performed the RNA-seq validation and sensitivity analyses separately for variants on the chromosomes used for training and variants on the rest of the chromosomes (Figures S4A and S4B). Although the network was trained only on the reference genomic sequence and splice annotations, and was not exposed to variant data during training, we wanted to rule out the possibility of biases in variant predictions arising from the fact that the network has seen the reference sequence in the training chromosomes. We found that the network performs equally well on variants from the training and testing chromosomes, with no significant difference in validation rate or sensitivity ($p > 0.05$, Fisher's exact test), indicating that the network's variant predictions are unlikely to be explained by overfitting the training sequences.

Tissue-Specific Alternative Splicing Frequently Arises from Weak Cryptic Splice Variants

Alternative splicing is a major mode of gene regulation that serves to increase the diversity of transcripts in different tissues and developmental stages, and its dysregulation is associated with disease processes (Irimia et al., 2014; Wang et al., 2008). Unexpectedly, we find that the relative usage of novel splice junctions created by cryptic splice mutations can vary substantially across tissues (Figure 3A). Moreover, variants that cause tissue-specific differences in splicing are reproducible across multiple individuals (Figure 3B), indicating that tissue-specific biology likely underlies these differences, rather than stochastic effects. We find that 35% of cryptic splice variants with weak and intermediate predicted scores (Δ score 0.35–0.8) exhibit significant differences in the fraction of normal and aberrant transcripts produced across tissues (Bonferroni corrected $p < 0.01$ for a χ^2 test, Figure 3C). This contrasted with variants with high predicted scores (Δ score >0.8), which were significantly less likely to produce tissue-specific effects ($p = 0.015$). Our findings align with the earlier observation that alternatively spliced exons tend to have intermediate prediction scores (Figure 1C), compared to exons that are constitutively spliced in or spliced out, which have scores that are close to 1 or 0, respectively.

These results support a model where tissue-specific factors, such as the chromatin context and binding of RNA-binding proteins, may swing the contest between two splice junctions that are close in favorability (Luco et al., 2010; Ule et al., 2003). Strong cryptic splice variants are likely to fully shift splicing from the normal to the aberrant isoform irrespective of the epigenetic context, whereas weaker variants bring splice junction selection closer to the decision boundary, resulting in alternative junction usage in different tissue types and cell contexts. This highlights the unexpected role played by cryptic splice mutations in generating novel alternative splicing diversity, as natural selection would then have the opportunity to preserve mutations that create useful tissue-specific alternative splicing.

Predicted Cryptic Splice Variants Are Strongly Deleterious in Human Populations

Although predicted cryptic splice variants validate at a high rate in RNA-seq, in many cases the effects are not fully penetrant and a mixture of both normal and aberrant splice isoforms are produced, raising the possibility that a fraction of these cryptic splice-altering variants may not be functionally significant. To explore the signature of natural selection on predicted cryptic splice variants, we scored each variant present in 60,706 human exomes from the Exome Aggregation Consortium (ExAC)

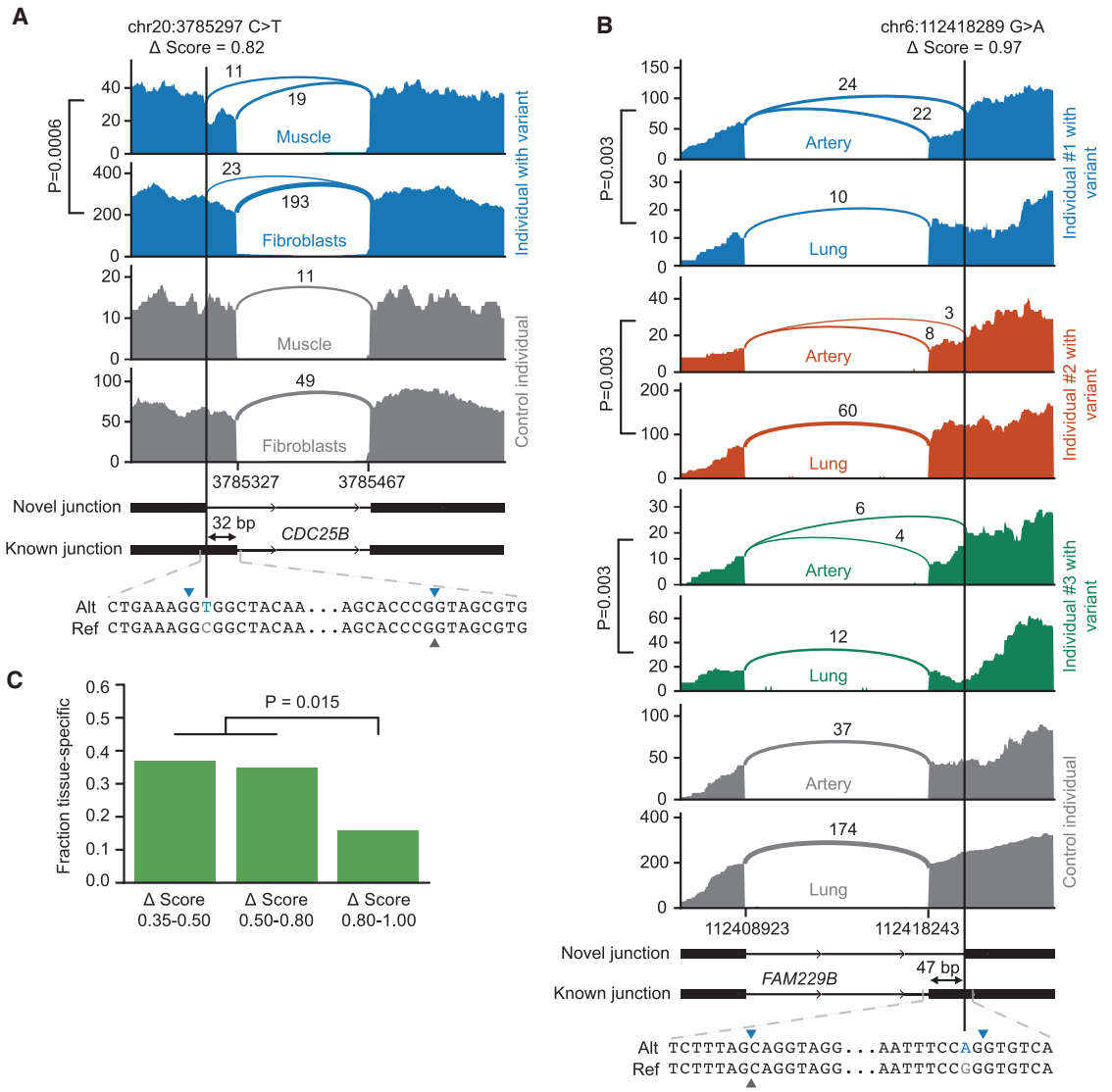


Figure 3. Cryptic Splice Variants Frequently Create Tissue-Specific Alternative Splicing

(A) Example of a heterozygous exonic variant in the *CDC25B* gene, which creates a novel donor site. The variant is private to a single individual in the GTEx cohort and exhibits tissue-specific alternative splicing that favors a greater fraction of the novel splice isoform in muscle compared to fibroblasts (p = 0.006 by Fisher's exact test). RNA-seq coverage, junction read counts, and the positions of junctions (blue and gray arrows) are shown for the individual with the variant and a control individual, in both muscle and fibroblasts.

(B) Example of a heterozygous exonic acceptor-creating variant in the *FAM229B* gene, which exhibits consistent tissue-specific effects across all three individuals in the GTEx cohort who harbor the variant. RNA-seq for artery and lung are shown for the three individuals with the variant and a control individual.

(C) Fraction of splice site-creating variants in the GTEx cohort that are associated with significantly non-uniform usage of the novel junction across expressing tissues, evaluated by the chi-square test for homogeneity. Validated cryptic splice variants with low to intermediate Δ score values were more likely to result in tissue-specific alternative splicing (p = 0.015, Fisher's exact test).

database (Lek et al., 2016) and identified variants that were predicted to alter exon-intron boundaries.

To measure the extent of negative selection acting on predicted splice-altering variants, we counted the number of predicted splice-altering variants found at common allele frequencies (≥0.1% in the human population) and compared it to the number of predicted splice-altering variants at singleton allele frequencies in ExAC (i.e., in 1 out of 60,706 individuals). Because of the recent exponential expansion in human population size, singleton variants represent recently created mutations that have been minimally filtered by purifying selection (Tennessen et al., 2012). In contrast, common variants represent a subset of neutral mutations that have passed through the sieve of purifying selection. Hence, depletion of predicted splice-altering variants in the common allele frequency spectrum relative to singleton variants provides an estimate of the fraction of

A

ExAC near exonic variants

	Singleton	Common (AF ≥ 0.1%)
SNVs with Δ score ≥ 0.8	10,369	212
SNVs with Δ score < 0.1	1,687,004	158,177

Odds Ratio (OR) = 4.58 ($P < 10^{-127}$)
Fraction of deleterious variants = (1 - 1/OR)

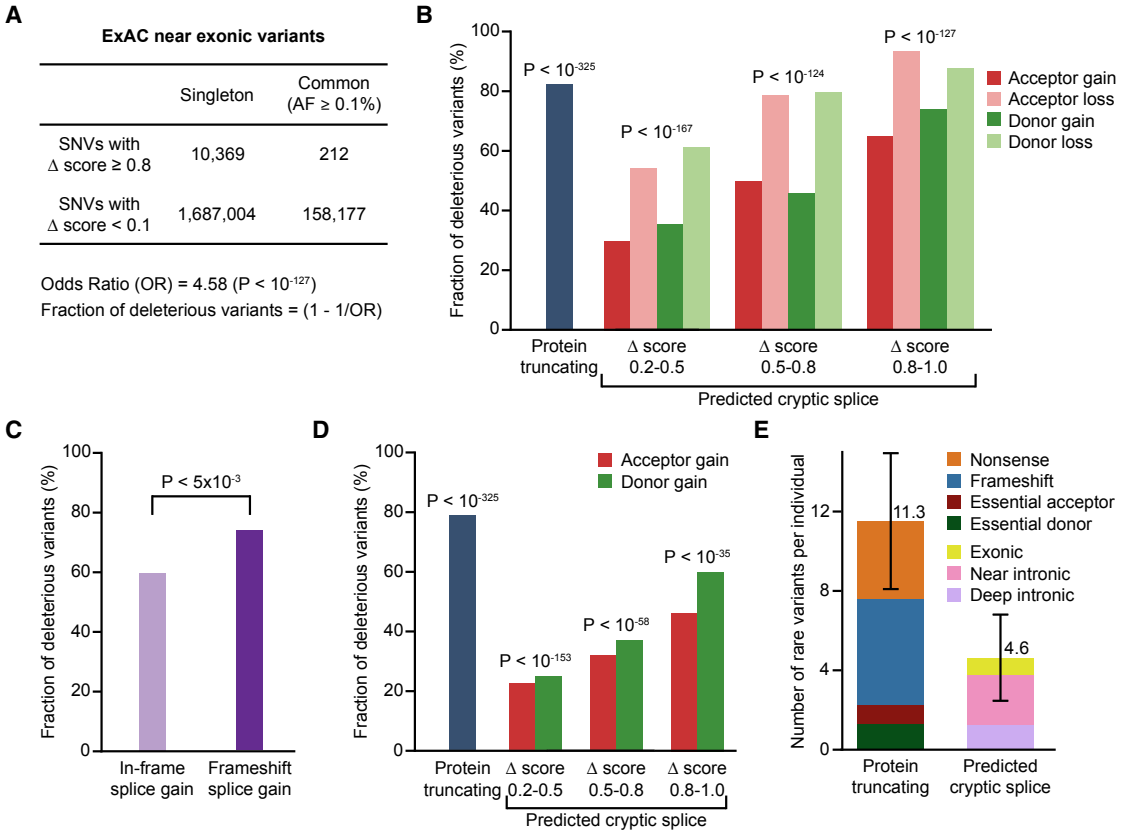


Figure 4. Predicted Cryptic Splice Variants Are Strongly Deleterious in Human Populations

(A) Synonymous and intronic variants (≤50 nt from known exon-intron boundaries and excluding the essential GT or AG dinucleotides) with confidently predicted splice-altering effects (Δ score ≥0.8) are strongly depleted at common allele frequencies (≥0.1%) in the human population relative to rare variants observed only once in 60,706 individuals. The 4.58 odds ratio ($p < 10^{-127}$ by χ^2 test) indicates that 78% of recently arising predicted cryptic splice variants are sufficiently deleterious to be removed by natural selection.

(B) Fraction of protein-truncating variants and predicted synonymous and intronic cryptic splice variants in the ExAC dataset that are deleterious, calculated as in (A).

(C) Fraction of synonymous and intronic cryptic splice gain variants in the ExAC dataset that are deleterious (Δ score ≥0.8), split based on whether the variant is expected to cause a frameshift or not ($p < 0.005$ by χ^2 test).

(D) Fraction of protein-truncating variants and predicted deep intronic (>50 nt from known exon-intron boundaries) cryptic splice variants in the gnomAD dataset that are deleterious, calculated as in (A).

(E) Average number of rare (allele frequency <0.1%) protein-truncating variants and rare functional cryptic splice variants per individual human genome. The number of cryptic splice mutations that are expected to be functional is estimated based on the fraction of predictions that are deleterious. The total number of predictions is higher. Error bars represent SD.

See also Figure S4.

predicted splice-altering variants that are deleterious, and therefore functional. To avoid confounding effects on the protein-coding sequence, we restricted our analysis to synonymous variants and intronic variants lying outside the essential GT or AG dinucleotides, excluding missense mutations that are also predicted to have splice-altering effects.

At common allele frequencies, high-scoring predicted cryptic splice variants (Δ score ≥0.8) are under strong negative selection, as evidenced by their relative depletion compared to expectation (Figure 4A). At this threshold, where most variants are expected to be close to fully penetrant in the RNA-seq data (Figure 2D), predicted synonymous and intronic cryptic splice mutations are depleted by 78% at common allele frequencies, which is comparable with the 82% depletion of frameshift, stop-gain, and essential GT or AG splice-disrupting variants (Figure 4B). The impact of negative selection is larger when considering cryptic splice variants that would cause frameshifts over those that cause in-frame changes (Figure 4C). The depletion of cryptic splice variants with frameshift consequence is nearly identical to that of other classes of protein-truncating variation, indicating that the vast majority of confidently predicted cryptic splice mutations in the exonic and near-intronic regions (≤50 nt from known exon-intron boundaries) are functional and have strongly deleterious effects in the human population.

To extend this analysis into deep intronic regions (>50 nt from known exon-intron boundaries), we used aggregated

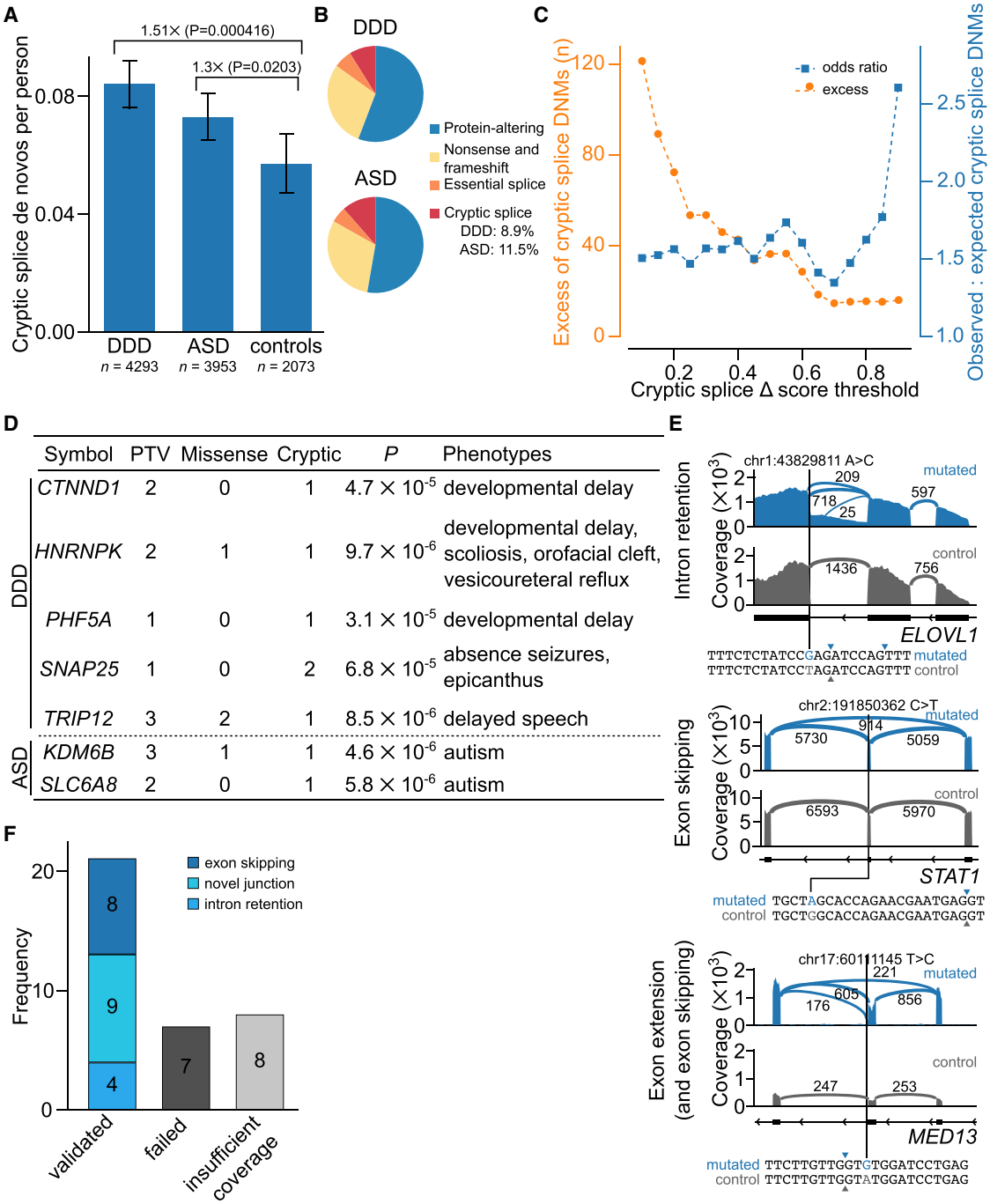


Figure 5. *De Novo* Cryptic Splice Mutations in Patients with Rare Genetic Disease

(A) Predicted cryptic splice *de novo* mutations per person for patients from the Deciphering Developmental Disorders cohort (DDD), individuals with autism spectrum disorders (ASDs) from the Simons Simplex Collection and the Autism Sequencing Consortium, as well as healthy controls. Enrichment in the DDD and ASD cohorts above healthy controls is shown, adjusting for variant ascertainment between cohorts. Error bars show 95% confidence intervals.

(B) Estimated proportion of pathogenic *de novo* mutations by functional category for the DDD and ASD cohorts, based on the enrichment of each category compared to healthy controls.

(C) Enrichment and excess of cryptic splice *de novo* mutations in the DDD and ASD cohorts compared to healthy controls at different Δ score thresholds.

(D) List of novel candidate disease genes enriched for *de novo* mutations in the DDD and ASD cohorts (FDR <0.01), when predicted cryptic splice mutations were included together with protein-coding mutations in the enrichment analysis. Phenotypes that were present in multiple individuals are shown.

(E) Three examples of predicted *de novo* cryptic splice mutations in autism patients that validate on RNA-seq, resulting in intron retention, exon skipping, and exon extension, respectively. For each example, RNA-seq coverage and junction counts for the affected individual are shown at the top, and a control individual

(legend continued on next page)

whole-genome sequencing data from 15,496 humans from the Genome Aggregation Database (gnomAD) cohort (Lek et al., 2016) to calculate the observed and expected counts of cryptic splice mutations at common allele frequencies. Overall, we observe a 56% depletion of common cryptic splice mutations (Δ score ≥0.8) at a distance >50 nt from an exon-intron boundary (Figure 4D), consistent with greater difficulty in predicting the impact of deep intronic variants, as we had observed in the RNA-seq data.

We next sought to estimate the potential for cryptic splice mutations to contribute to penetrant genetic disease, relative to other types of protein-coding variation, by measuring the number of rare cryptic splice mutations per individual in the gnomAD cohort. Based on the fraction of predicted cryptic splice mutations that are under negative selection (Figure 4A), the average human carries ~5 rare functional cryptic splice mutations (allele frequency <0.1%), compared to ~11 rare protein-truncating variants (Figure 4E). Cryptic splice variants outnumber essential GT or AG splice-disrupting variants roughly 2:1. We caution that a significant fraction of these cryptic splice variants may not fully abrogate gene function, either because they produce in-frame alterations, or because they do not completely shift splicing to the aberrant isoform.

De Novo Cryptic Splice Mutations Are a Major Cause of Rare Genetic Disorders

Large-scale sequencing studies of patients with autism spectrum disorders and severe intellectual disability have demonstrated the central role of *de novo* protein-coding mutations (missense, nonsense, frameshift, and essential splice dinucleotide) that disrupt genes in neurodevelopmental pathways (Fitzgerald et al., 2015; Iossifov et al., 2014; McRae et al., 2017; Neale et al., 2012; De Rubeis et al., 2014; Sanders et al., 2012). To assess the clinical impact of noncoding mutations that act through altered splicing, we applied the neural network to predict the effects of *de novo* mutations in 4,293 individuals with intellectual disability from the Deciphering Developmental Disorders (DDD) cohort (McRae et al., 2017), 3,953 individuals with autism spectrum disorders (ASDs) from the Simons Simplex Collection (De Rubeis et al., 2014; Sanders et al., 2012; Turner et al., 2016) and the Autism Sequencing Consortium, and 2,073 unaffected sibling controls from the Simons Simplex Collection. To control for differences in *de novo* variant ascertainment across studies, we normalized the expected number of *de novo* variants such that the number of synonymous mutations per individual was the same across cohorts.

De novo mutations that are predicted to disrupt splicing are enriched 1.51-fold in intellectual disability ($p = 0.000416$) and 1.30-fold in autism spectrum disorder ($p = 0.0203$) compared to healthy controls (Δ score ≥0.1, Figure 5A; Table S3). Splice-disrupting mutations are also significantly enriched in cases versus controls when considering only synonymous and intronic mutations (Figures S5A–S5C), excluding the possibility that the enrichment could be explained solely by mutations with dual protein-coding and splicing effects. Based on the excess of *de novo* mutations in affected versus unaffected individuals, cryptic splice mutations are estimated to comprise about 11% of pathogenic mutations in autism spectrum disorder, and 9% in intellectual disability (Figure 5B), after adjusting for the expected fraction of mutations in regions that lacked sequencing coverage or variant ascertainment in each study. Most *de novo* predicted cryptic splice mutations in affected individuals had Δ scores <0.5 (Figures 5C, S5D, and S5E) and would be expected to produce a mixture of normal and aberrant transcripts based on variants with similar scores in the GTEx RNA-seq dataset.

To estimate the enrichment of cryptic splice mutations in candidate disease genes compared to chance, we calculated the probability of calling a *de novo* cryptic splice mutation for each individual gene using trinucleotide context to adjust for mutation rate (Samocha et al., 2014) (Table S3). Combining both cryptic splice mutations and protein-coding mutations in novel gene discovery yields 5 additional candidate genes associated with intellectual disability and 2 additional genes associated with autism spectrum disorder (Figure 5D; Table S3) that would have been below the discovery threshold (false discovery rate [FDR] <0.01) when considering only protein-coding mutations (Sanders et al., 2015).

Experimental Validation of *De Novo* Cryptic Splice Mutations in Autism Patients

We obtained peripheral blood-derived lymphoblastoid cell lines (LCLs) from 36 individuals from the Simons Simplex Collection, which harbored predicted *de novo* cryptic splice mutations in genes with at least a minimal level of LCL expression (Iossifov et al., 2014; Sanders et al., 2015); each individual represented the only case of autism within their immediate family. As is the case for most rare genetic diseases, the tissue and cell type of relevance (presumably developing brain) was not accessible. Hence, we performed high-depth mRNA sequencing (~350 million × 150-bp single-end reads per sample, roughly 10 times the coverage of GTEx) to compensate for the weak expression of many of these transcripts in LCLs. To ensure that we were validating a representative set of predicted cryptic splice variants, rather than simply the top predictions, we applied relatively permissive thresholds (Δ score >0.1 for splice loss variants and Δ score >0.5 for splice gain variants; STAR Methods) and performed experimental validation on all *de novo* variants meeting these criteria.

After excluding 8 individuals who had insufficient RNA-seq coverage at the gene of interest (Figure S6), we identified unique, aberrant splicing events associated with the predicted *de novo* cryptic splice mutation in 21 out of 28 patients (Figures 5E and S6). These aberrant splicing events were absent from the other 35 individuals for whom deep LCL RNA-seq was obtained, as well as the 149 individuals from the GTEx cohort. Among the

without the mutation is shown at the bottom. Sequences are shown on the sense strand with respect to the transcription of the gene. Blue and gray arrows demarcate the positions of the junctions in the individual with the variant and the control individual, respectively.
(F) Validation status for 36 predicted cryptic splice sites selected for experimental validation by RNA-seq.
See also Figures S4, S5, and S6 and Tables S3 and S4.

21 confirmed *de novo* cryptic splice mutations, we observed 9 cases of novel junction creation (Iossifov et al., 2014; Sanders et al., 2015), 8 cases of exon skipping, and 4 cases of intron retention, as well as more complex splicing aberrations (Figure 5F; Table S4). Seven cases did not show aberrant splicing in LCLs, despite adequate expression of the transcript. Although a subset of these may represent false positive predictions, some cryptic splice mutations may result in tissue-specific alternative splicing that is not observable in LCLs under these experimental conditions.

The high validation rate of predicted cryptic splice mutations in patients with autism spectrum disorder (75%), despite the limitations of the RNA-seq assay, indicates that most predictions are functional. However, the enrichment of *de novo* cryptic splice variants in cases compared to controls (1.5-fold in DDD and 1.3-fold in ASD, Figure 5A) is only 38% of the effect size observed for *de novo* protein-truncating variants (2.5-fold in DDD and 1.7-fold in ASD) (Iossifov et al., 2014; McRae et al., 2017; De Rubeis et al., 2014). This allows us to quantify that functional cryptic splice mutations have roughly 50% of the clinical penetrance of classic forms of protein-truncating mutation (stop-gain, frameshift, and essential splice dinucleotide), on account of many of them only partially disrupting production of the normal transcript. Indeed, some of the most well-characterized cryptic splice mutations in Mendelian diseases, such as c.315-48T > C in *FECH* (Gouya et al., 2002) and c.-32-13T > G in *GAA* (Boerkoel et al., 1995), are hypomorphic alleles associated with milder phenotype or later age of onset. The estimate of clinical penetrance is calculated for all *de novo* variants meeting a relatively permissive threshold (Δ score $\geq$0.1), and variants with stronger prediction scores would be expected to have correspondingly higher penetrance.

Based on the excess of *de novo* mutations in cases versus controls across the ASD and DDD cohorts, 250 cases can be explained by *de novo* cryptic splice mutations compared to 909 cases that can be explained by *de novo* protein-truncating variants (Figure 5B). This is consistent with our earlier estimate of the average number of rare cryptic splice mutations ($\sim$5) compared to rare protein-truncating variants ($\sim$11) per person in the general population (Figure 2E), once the reduced penetrance of cryptic splice mutations is factored in. The widespread distribution of cryptic splice mutations across the genome suggests that the fraction of cases explained by cryptic splice mutations in neurodevelopmental disorders (9%–11%, Figure 5B) is likely to generalize to other rare genetic disorders where the primary disease mechanism is loss of the functional protein. To facilitate the interpretation of splice-altering mutations, we precomputed the Δ score predictions for all possible single nucleotide substitutions genome-wide and provide them as a resource to the scientific community. We believe that this resource will promote understanding of this previously under-appreciated source of genetic variation.

DISCUSSION

Despite the limited diagnostic yield of exome sequencing in patients with severe genetic disorders, clinical sequencing has focused on rare coding mutations, largely disregarding variation in the noncoding genome due to the difficulty of interpretation. Here, we introduce a deep learning network that accurately predicts splicing from the primary nucleotide sequence, thereby identifying noncoding mutations that disrupt the normal patterning of exons and introns with severe consequences on the resulting protein. We show that predicted cryptic splice mutations validate at a high rate by RNA-seq, are strongly deleterious in the human population, and are a major cause of rare genetic disease.

By using the deep learning network as an *in silico* model of the spliceosome, we were able to reconstruct the specificity determinants that enable the spliceosome to achieve its remarkable precision *in vivo*. We reaffirm many of the discoveries that were made over the past four decades of research into splicing mechanisms and show that the spliceosome integrates a large number of short- and long-range specificity determinants in its decisions. In particular, we find that the perceived degeneracy of most splice motifs is explained by the presence of long-range determinants such as exon-intron lengths and nucleosome positioning, which more than compensate and render additional specificity at the motif level unnecessary. Our findings demonstrate the promise of deep learning models for providing biological insights, rather than merely serving as black box classifiers.

Deep learning is a relatively new technique in biology and is not without potential trade-offs. By learning to automatically extract features from a sequence, deep learning models can utilize sequence determinants not well described by human experts, but there is also the risk that the model may incorporate features that do not reflect the true behavior of the spliceosome. These irrelevant features could increase the apparent accuracy of predicting annotated exon-intron boundaries but would reduce the accuracy of predicting the splice-altering effects of arbitrary sequence changes induced by genetic variation. Because accurate prediction of variants provides the strongest evidence that the model can generalize to true biology, we provide validation of predicted splice-altering variants using three fully orthogonal methods: RNA-seq, natural selection in human populations, and enrichment of *de novo* variants in case versus control cohorts. While this does not fully preclude the incorporation of irrelevant features into the model, the resulting model appears faithful enough to the true biology of splicing to be of significant value for practical applications such as identifying cryptic splice mutations in patients with genetic diseases.

Compared to other classes of protein-truncating mutations, a particularly interesting aspect of cryptic splice mutations is the widespread phenomenon of alternative splicing due to incompletely penetrant splice-altering variants, which tend to weaken canonical splice sites relative to alternative splice sites, resulting in the production of a mixture of both aberrant and normal transcripts in the RNA-seq data. The observation that these variants frequently drive tissue-specific alternative splicing highlights the unexpected role played by cryptic splice mutations in generating novel alternative splicing diversity. A potential future direction would be to train deep learning models on splice junction annotations from

RNA-seq of the relevant tissue, thereby obtaining tissue-specific models of alternative splicing. Training the network on annotations derived directly from RNA-seq data also helps to fill gaps in the GENCODE annotations, which improves the performance of the model on variant prediction (Figures S3G and S3H).

Our understanding of how mutations in the noncoding genome lead to human disease remains far from complete. The discovery of penetrant *de novo* cryptic splice mutations in childhood neurodevelopmental disorders demonstrates that improved interpretation of the noncoding genome can directly benefit patients with severe genetic disorders. Cryptic splice mutations also play major roles in cancer (Jung et al., 2015; Supek et al., 2014), and recurrent somatic mutations in splice factors have been shown to produce widespread alterations in splicing specificity (Shirai et al., 2015; Yoshida et al., 2011). Much work remains to be done to understand regulation of splicing in different tissues and cellular contexts, particularly in the event of mutations that directly impact proteins in the spliceosome. In light of recent advances in oligonucleotide therapy that could potentially target splicing defects in a sequence-specific manner (Finkel et al., 2017), greater understanding of the regulatory mechanisms that govern this remarkable process could pave the way for novel candidates for therapeutic intervention.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Deep learning for splice prediction
  - Analyses on the GTEx RNA-seq dataset
  - Analyses on the ExAC and gnomAD datasets
  - Analyses on the DDD and ASD datasets
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and four tables and can be found with this article online at https://doi.org/10.1016/j.cell.2018.12.015.

ACKNOWLEDGMENTS

We would like to acknowledge J. K. Pritchard for insightful discussions and support, the Genome Aggregation Database (gnomAD), and the groups that provided exome and genome variant data to this resource. S.J.S. was supported by a grant from the Simons Foundation (SFARI #402281 and #574598).

AUTHOR CONTRIBUTIONS

K.J. performed the deep learning and population depletion analyses with assistance from D.K., Y.I.L., H.G., A.K., and S.B. S.K.P. performed the GTEx analysis with assistance from E.K. and W.C. J.F.M. performed the disease analysis with assistance from J.A.K. S.F.D. performed the RNA-seq validation in autism samples with assistance from J.A., G.B.S., E.D.C., and S.J.S. K.K.-H.F. supervised the analyses.

DECLARATION OF INTERESTS

K.J., S.K.P., J.F.M., J.A.K., W.C., E.K., H.G., A.K., S.B., and K.K.-H.F. were employed by Illumina at the time of this study. The following patents related to this work have been filed: Deep Learning-Based Splice Site Classification, Deep Learning-Based Aberrant Splicing Detection, Aberrant Splicing Detection Using Convolutional Neural Networks (CNNS).

Received: March 29, 2018
Revised: August 31, 2018
Accepted: December 10, 2018
Published: January 17, 2019

REFERENCES

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., et al. (2012). Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep. *1*, 543–556.

Berget, S.M. (1995). Exon recognition in vertebrate splicing. J. Biol. Chem. *270*, 2411–2414.

Boerkoel, C.F., Exelbert, R., Nicastri, C., Nichols, R.C., Miller, F.W., Plotz, P.H., and Raben, N. (1995). Leaky splicing mutation in the acid maltase gene is associated with delayed onset of glycogenosis type II. Am. J. Hum. Genet. *56*, 887–897.

Boyd, S., Cortes, C., Mohri, M., and Radovanovic, A. (2012). Accuracy at the top. In Advances in Neural Processing Systems, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, eds. (NIPS), pp. 953–961.

Close, P., East, P., Dirac-Svejstrup, A.B., Hartmann, H., Heron, M., Maslen, S., Chariot, A., Söding, J., Skehel, M., and Svejstrup, J.Q. (2012). DBIRD complex integrates alternative mRNA splicing with RNA polymerase II transcript elongation. Nature *484*, 386–389.

Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell *136*, 777–793.

Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. Published online April 19, 2017. https://doi.org/10.1126/scitranslmed.aal5209.

De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al.; DDD Study; Homozygosity Mapping Collaborative for Autism; UK10K Consortium (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215.

Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297*, 1007–1013.

Farh, K.K.H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.

Finkel, R.S., Mercuri, E., Darras, B.T., Connolly, A.M., Kuntz, N.L., Kirschner, J., Chiriboga, C.A., Saito, K., Servais, L., Tizzano, E., et al.; ENDEAR Study Group (2017). Nusinersen versus sham control in infantile-onset spinal muscular atrophy. N. Engl. J. Med. *377*, 1723–1732.

Fitzgerald, T.W., Gerety, S.S., Jones, W.D., van Kogelenberg, M., King, D.A., McRae, J., Morley, K.I., Parthiban, V., Al-Turki, S., Ambridge, K., et al.; Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223–228.

Gouya, L., Puy, H., Robreau, A.-M., Bourgeois, M., Lamoril, J., Da Silva, V., Grandchamp, B., and Deybach, J.-C. (2002). The penetrance of dominant erythropoietic protoporphyria is modulated by expression of wildtype FECH. Nat. Genet. *30*, 27–28.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), L. O'Conner, ed. (IEEE), pp. 770–778.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.

Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., et al. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell *159*, 1511–1523.

Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. eLife *3*, e02407.

Jung, H., Lee, D., Lee, J., Park, D., Kim, Y.J., Park, W.-Y., Hong, D., Park, P.J., and Lee, E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat. Genet. *47*, 1242–1248.

Lee, H., Deignan, J.L., Dorrani, N., Strom, S.P., Kantarci, S., Quintero-Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., et al. (2014). Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA *312*, 1880–1887.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. *50*, 151–158.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science *327*, 996–1000.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl variant effect predictor. Genome Biol. *17*, 122.

McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., et al.; Deciphering Developmental Disorders Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433–438.

Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature *485*, 242–245.

Pertea, M., Lin, X., and Salzberg, S.L. (2001). GeneSplicer: A new computational method for splice site prediction. Nucleic Acids Res. *29*, 1185–1190.

Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. J. Comput. Biol. *4*, 311–323.

Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.

Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature *485*, 237–241.

Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al.; Autism Sequencing Consortium (2015). Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron *87*, 1215–1233.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol. *16*, 990–995.

Shirai, C.L., Ley, J.N., White, B.S., Kim, S., Tibbitts, J., Shao, J., Ndonwi, M., Wadugu, B., Duncavage, E.J., Okeyo-Owuor, T., et al. (2015). Mutant U2AF1 expression alters hematopoiesis and Pre-mRNA splicing in vivo. Cancer Cell *27*, 631–643.

Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nat. Genet. *49*, 848–855.

Spies, N., Nielsen, C.B., Padgett, R.A., and Burge, C.B. (2009). Biased chromatin signatures around polyadenylation sites and exons. Mol. Cell *36*, 245–254.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. Cell *156*, 1324–1335.

Tan, T.Y., Dillon, O.J., Stark, Z., Schofield, D., Alam, K., Shrestha, R., Chong, B., Phelan, D., Brett, G.R., Creed, E., et al. (2017). Diagnostic impact and cost-effectiveness of whole-exome sequencing for ambulant children with suspected monogenic conditions. JAMA Pediatr. *171*, 855–862.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

The GTEx Consortium, Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660.

Turner, T.N., Hormozdiari, F., Duyzend, M.H., McClymont, S.A., Hook, P.W., Iossifov, I., Raja, A., Baker, C., Hoekzema, K., Stessman, H.A., et al. (2016). Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. Am. J. Hum. Genet. *98*, 58–74.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. ArXiv, ArXiv:1609.03499.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. RNA *14*, 802–813.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831–845.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

Wu, J., Anczuków, O., Krainer, A.R., Zhang, M.Q., and Zhang, C. (2013). OLego: Fast and sensitive mapping of spliced mRNA-seq reads using small seeds. Nucleic Acids Res. *41*, 5149–5163.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. JAMA *312*, 1870–1879.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., et al. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. Nature *478*, 64–69.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
RPMI 1640	GIBCO	Cat#21870076
L-Glutamine	GIBCO	Cat#25030081
Fetal bovine serum	Atlanta Biologicals	Cat#S11150
Critical Commercial Assays		
RNeasy Plus Micro Kit	QIAGEN	Cat#74034
Agilent RNA 6000 Nano Kit	Agilent	Cat#5067-1511
TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold Set A	Illumina	Cat#RS-122-2301
Deposited Data		
RNA-seq data and variant calls for the GTEx cohort	http://www.ncbi.nlm.nih.gov/projects/gap	dbGAP accession: phs000424.v6.p1
De-novo mutations for autism patients and healthy controls	Iossifov et al., 2014	N/A
De-novo mutations from the Deciphering Developmental Disorders cohort	McRae et al., 2017	N/A
Splice junctions from GENCODE principal transcripts used to train the SpliceAI models	This study	https://basespace.illumina.com/s/5u6ThOblecrh
Splice junctions from GENCODE and GTEx data used to train the augmented SpliceAI models	This study	https://basespace.illumina.com/s/5u6ThOblecrh
Splice junctions from GENCODE principal transcripts used to test the models, with paralogs excluded	This study	https://basespace.illumina.com/s/5u6ThOblecrh
Splice junctions of lincRNAs used to test the models	This study	https://basespace.illumina.com/s/5u6ThOblecrh
Predicted delta scores for all possible SNVs	This study	https://basespace.illumina.com/s/5u6ThOblecrh
All GTEx junctions in all GTEx v6.p1 samples	This study	https://basespace.illumina.com/s/5u6ThOblecrh
List of splice-altering variants appearing in 1-4 GTEx individuals with Δ Score > 0.1 that validated against RNA-seq	This study	https://basespace.illumina.com/s/5u6ThOblecrh
Aligned BAM files for RNA-seq in 36 autism patients	This study	ArrayExpress accession: E-MTAB-7351
Experimental Models: Cell Lines		
Lymphoblastoid cell lines	Derived from patient samples from Simons Simplex Collection	Patient IDs are in Table S4
Software and Algorithms		
SpliceAI software	This study	https://github.com/Illumina/SpliceAI
SpliceAI model training	This study	https://basespace.illumina.com/s/5u6ThOblecrh
MaxEntScan	Yeo and Burge, 2004	N/A
GeneSplicer	Pertea et al., 2001	N/A
NNSplice	Reese et al., 1997	N/A

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for materials should be directed to and will be fulfilled by the Lead Contact, Kyle Farh (kfarh@illumina.com).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subject details for the 36 autism patients were previously released by Iossifov et al., Nature 2014 and are provided in Table S4 in our paper. Informed consent was obtained for all subjects included in this analysis as part of the Simons Simplex Collection (www.sfari.org). The analyses on human subjects were approved by the UCSF Institutional Review Board (IRB #: 14-14749).

Lymphoblastoid cell lines were obtained from the SSC for these probands. Cells were cultured in Culture Medium (RPMI 1640, 2mM L-glutamine, 15% fetal bovine serum) to a maximum cell density of 1×10^6 cells/ml. When cells reached maximum density, they were passaged by dissociating the cells by pipetting up and down 4 or 5 times and seeding to a density of 200,000-500,000 viable cells/ml. Cells were grown under 37°C, 5% CO_2 conditions for 10 days. Approximately 5×10^5 cells were then detached and spun down at 300 × g for 5 min at 4°C. RNA was extracted using RNeasy Plus Micro Kit (QIAGEN) following manufacturer's protocol.

METHOD DETAILS

Deep learning for splice prediction

SpliceAI architecture

We trained several ultra-deep convolutional neural network-based models to computationally predict splicing from pre-mRNA nucleotide sequence. We designed four architectures, namely, SpliceAI-80nt, SpliceAI-400nt, SpliceAI-2k, and SpliceAI-10k, which use 40, 200, 1,000, and 5,000 nucleotides on each side of a position of interest as input respectively, and output the probability of the position being a splice acceptor and donor. More precisely, the input to the models is a sequence of one-hot encoded nucleotides, where A, C, G, and T (or equivalently U) are encoded as [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], and [0, 0, 0, 1] respectively and the output of the models consists of three scores which sum to one, corresponding to the probability of the position of interest being a splice acceptor, splice donor, and neither.

The basic unit of the SpliceAI architectures is a residual block, which consists of batch-normalization layers, rectified linear units (ReLU), and convolutional units organized in a specific manner (Figure S1). Residual blocks are commonly used when designing deep neural networks. Prior to the development of residual blocks, deep neural networks consisting of many convolutional units stacked one after the other were very difficult to train due to the problem of exploding/vanishing gradients, and increasing the depth of such neural networks often resulted in a higher training error (He et al., 2016). Through a comprehensive set of computational experiments, architectures consisting of many residual blocks stacked one after the other were shown to overcome these issues (He et al., 2016).

The complete SpliceAI architectures are provided in Figure S1. The architectures consist of K stacked residual blocks connecting the input layer to the penultimate layer, and a convolutional unit with softmax activation connecting the penultimate layer to the output layer. The residual blocks are stacked such that the output of the i^{th} residual block is connected to the input of the $i+1^{th}$ residual block. Further, the output of every fourth residual block is added to the input of the penultimate layer. Such "skip connections" are commonly used in deep neural networks to increase convergence speed during training (van den Oord et al., 2016).

Each residual block has three hyper-parameters N, W, and D, where N denotes the number of convolutional kernels, W denotes the window size, and D denotes the dilation rate of each convolutional kernel. Since a convolutional kernel of window size W and dilation rate D extracts features spanning $(W-1)D$ neighboring positions, a residual block with hyper-parameters N, W, and D extracts features spanning $2(W-1)D$ neighboring positions. Hence, the total neighbor span of the SpliceAI architectures is given by $S = \sum_{i=1}^{K} 2(W_i - 1)D_i$, where N_i, W_i, and D_i are the hyper-parameters of the i^{th} residual block. For SpliceAI-80nt, SpliceAI-400nt, SpliceAI-2k, and SpliceAI-10k architectures, the number of residual blocks and the hyper-parameters for each residual block were chosen so that S is equal to 80, 400, 2,000, and 10,000 respectively.

The SpliceAI architectures only have normalization and non-linear activation units in addition to convolutional units. Consequently, the models can be used in a sequence-to-sequence mode with variable sequence length (van den Oord et al., 2016). For example, the input to the SpliceAI-10k model (S = 10,000) is a one-hot encoded nucleotide sequence of length $S/2 + I + S/2$, and the output is an $I \times 3$ matrix, corresponding to the three scores of the I central positions in the input, i.e., the positions remaining after excluding the first and last $S/2$ nucleotides. This feature can be leveraged to obtain a tremendous amount of computational saving during training as well as testing. This is due to the fact that most of the computations for positions which are close to each other are common, and the shared computations need to be done only once by the models when they are used in a sequence-to-sequence mode.

Model training and testing

We downloaded the GENCODE (Harrow et al., 2012) V24lift37 gene annotation table from the UCSC table browser and extracted 20,287 protein-coding gene annotations, selecting the principal transcript when multiple isoforms were available. We removed genes which did not have any splice junctions and split the remaining into training and test set genes as follows: The genes which belonged to chromosomes 2, 4, 6, 8, 10-22, X, and Y were used for training the models (13,384 genes, 130,796 donor-acceptor pairs). We randomly chose 10% of the training genes and used them for determining the point for early-stopping during training, and the rest were used for training the models. For testing the models, we used genes from chromosomes 1, 3, 5, 7, and 9 which did not have any paralogs (1,652 genes, 14,289 donor-acceptor pairs). To this end, we referred to the human gene paralog list from http://grch37.ensembl.org/biomart/martview.

We used the following procedure to train and test the models in a sequence-to-sequence mode with chunks of size $l = 5,000$. For each gene, the mRNA transcript sequence between the canonical transcription start and end sites was extracted from the hg19/GRCh37 assembly. The input mRNA transcript sequence was one-hot encoded as follows: A, C, G, T/U mapped to [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], [0, 0, 0, 1] respectively. The one-hot encoded nucleotide sequence was zero-padded until the length became a multiple of 5,000, and then further zero-padded at the start and the end with a flanking sequence of length $S/2$, where S is equal to 80, 400, 2,000, and 10,000 for SpliceAI-80nt, SpliceAI-400nt, SpliceAI-2k, and SpliceAI-10k models respectively. The padded nucleotide sequence was then split into blocks of length $S/2 + 5,000 + S/2$ in such a way that the i^{th} block consisted of the nucleotide positions from $5,000(i-1) - S/2 + 1$ to $5,000i + S/2$. Similarly, the splice output label sequence was one-hot encoded as follows: not a splice site, splice acceptor (first nucleotide of the corresponding exon), and splice donor (last nucleotide of the corresponding exon) were mapped to [1, 0, 0], [0, 1, 0], and [0, 0, 1] respectively. The one-hot encoded splice output label sequence was zero-padded until the length became a multiple of 5,000, and then split into blocks of length 5,000 in such a way that the i^{th} block consisted of the positions from $5,000(i-1) + 1$ to $5,000i$. The one-hot encoded nucleotide sequence and the corresponding one-hot encoded label sequence were used as inputs to the model and target outputs of the model respectively.

The models were trained for 10 epochs with a batch size of 12 on two NVIDIA GeForce GTX 1080 Ti GPUs. Categorical cross-entropy loss between the target and predicted outputs was minimized using Adam optimizer during training. The learning rate of the optimizer was set to 0.001 for the first 6 epochs, and then reduced by a factor of 2 in every subsequent epoch. For each architecture, we repeated the training procedure 5 times and obtained 5 trained models. During testing, each input was evaluated using all 5 trained models and the average of their outputs was used as the predicted output. We used this network for the analyses in Figure 1 and related supplementary figures.

For the analyses in Figures 2, 3, 4, and 5 involving identification of splice-altering variants, we augmented the training set of GENCODE annotations to also include novel splice junctions commonly observed in the GTEx cohort on chromosomes 2, 4, 6, 8, 10-22, X, Y (67,012 splice donors and 62,911 splice acceptors). This increased the number of splice junction annotations in the training set by ~50%. Training the network on the combined dataset improved the sensitivity of detecting splice-altering variants in the RNA-seq data compared to the network trained on GENCODE annotations alone (compare Figures 2D and 2F to Figures S3G and S3H), particularly for predicting deep intronic splice-altering variants, and we used this network for the analyses involving evaluation of variants (Figures 2, 3, 4, and 5 and related supplementary figures). To ensure that the GTEx RNA-seq dataset did not contain overlap between training and evaluation, we only included junctions that were present in 5 or more individuals in the training dataset, and only evaluated the performance of the network on variants that were present in 4 or fewer individuals. Details of novel splice junction identification are described in "Detection of splice junctions" under the GTEx analysis section of the methods.

To reduce problems with circularity that have become a concern for the field, the authors explicitly request that the prediction scores from the method not be incorporated as a component of other classifiers, and instead ask that interested parties employ the provided source code and data to directly train and improve upon their own deep learning models.

Top-k accuracy

An accuracy metric like the percentage of positions classified correctly is largely ineffective due to the fact that most of the positions are not splice sites. We instead evaluated the models using two metrics that are more appropriate in such settings, namely, top-*k* accuracy and area under the precision-recall curve. The top-*k* accuracy for a particular class is defined as follows: Suppose the test set has *k* positions that belong to the class. We choose the threshold so that exactly *k* test set positions are predicted as belonging to the class. The fraction of these *k* predicted positions that truly belong to the class is reported as the top-*k* accuracy. Indeed, this is equal to the precision when the threshold is chosen so that precision and recall have the same value. The top-*k* accuracies and the area under the precision-recall curves reported in Figures 1E and S2A are the average results across the splice acceptor and donor classes.

Model evaluation on lincRNAs

We obtained a list of all lincRNA transcripts based on the GENCODE V24lift37 annotations. Unlike protein-coding genes, lincRNAs are not assigned a principal transcript in the GENCODE annotations. To minimize redundancy in the validation set, we identified the transcript with the longest total exonic sequence per lincRNA gene, and called this the canonical transcript for the gene. Since lincRNA annotations are expected to be less reliable than annotations for protein-coding genes, and such misannotations would affect our estimates of top-*k* accuracy, we used the GTEx data to eliminate lincRNAs with potential annotation issues (see section "Analyses on the GTEx dataset" below for details on these data). For each lincRNA, we counted all split reads that mapped across the length of the lincRNA across all GTEx samples (see "Detection of splice junctions" below for details). This was an estimate of the total junction-spanning reads of the lincRNA that used either annotated or novel junctions. We also counted the number of reads that spanned junctions of the canonical transcript. We only considered lincRNAs for which at least 95% of junction-spanning reads across all GTEx samples corresponded to the canonical transcript. We also required all junctions of the canonical transcript to be observed at least once in the GTEx cohort (excluding junctions that spanned introns of length < 10 nt). For computing top-*k* accuracy, we only considered the junctions of the canonical transcripts of the lincRNAs that passed the filters above (781 transcripts, 1047 junctions).

Identifying splice junctions from pre-mRNA sequence

In Figure 1B, we compare the performance of MaxEntScan and SpliceAI-10k with respect to identifying the canonical exon boundaries of a gene from its sequence. We used the *CFTR* gene, which is in our test set and has 26 canonical splice acceptors and donors, as a case study and obtained an acceptor and donor score for each of the 188,703 positions from the canonical transcription start site

(chr7:117,120,017) to the canonical transcription end site (chr7:117,308,719) using MaxEntScan and SpliceAI-10k. A position was classified as a splice acceptor or donor if its corresponding score was greater than the threshold chosen while evaluating the top-*k* accuracy. MaxEntScan predicted 49 splice acceptors and 22 splice donors, out of which 9 and 5 are true splice acceptors and donors respectively. For the sake of better visualization, we show the pre-log scores of MaxEntScan (clipped to a maximum of 2,500). SpliceAI-10k predicted 26 splice acceptors and 26 splice donors, all of which are correct. For Figure S2B, we repeated the analysis using the *LINC00467* gene.

Estimation of exon inclusion at GENCODE-annotated splice junctions

We computed the inclusion rate of all GENCODE-annotated exons from the GTEx RNA-seq data (Figure 1C). For each exon, excluding the first and last exon of each gene, we computed the inclusion rate as:

$$\frac{(L+R)/2}{S+(L+R)/2}$$

where L is the total read count of the junction from the previous canonical exon to the exon under consideration across all GTEx samples, R is the total read count of the junction from the exon under consideration to the next canonical exon, and S is the total read count of the skipping junction from the previous to the next canonical exon.

Significance of various nucleotides toward splice site recognition

In Figure 1D, we identify the nucleotides that are considered important by SpliceAI-10k toward the classification of a position as a splice acceptor. To this end, we considered the splice acceptor at chr3:142,740,192 in the *U2SURP* gene, which is in our test set. The ''importance score'' of a nucleotide with respect to a splice acceptor is defined as follows: Let s_{ref} denote the acceptor score of the splice acceptor under consideration. The acceptor score is recalculated by replacing the nucleotide under consideration with A, C, G, and T. Let these scores be denoted by s_A, s_C, s_G, and s_T respectively. The importance score of the nucleotide is estimated as

$$s_{ref} - \frac{s_A + s_C + s_G + s_T}{4}$$

This procedure is often referred to as in-silico mutagenesis. We plot 127 nucleotides from chr3:142,740,137 to chr3:142,740,263 in such a way that the height of each nucleotide is its importance score with respect to the splice acceptor at chr3:142,740,192.

Effect of TACTAAC and GAAGAA motifs on splicing

In order to study the impact of the position of the branch point sequence on acceptor strength, we first obtained the acceptor scores of the 14,289 test set splice acceptors using SpliceAI-10k. Let y_{ref} denote the vector containing these scores. For each value of i ranging from 0 to 100, we did the following: For each test set splice acceptor, we replaced the nucleotides from i to $i-6$ nt before the splice acceptor by TACTAAC and recomputed the acceptor score using SpliceAI-10k. The vector containing these scores is denoted by $y_{alt,\,i}$.

We plot the following quantity as a function of i in Figure S2C:

$$mean\left(y_{alt,i} - y_{ref}\right)$$

For Figure S2D, we repeated the same procedure using the SR-protein motif GAAGAA. In this case, we also studied the impact of the motif when present after the splice acceptor as well as the impact on donor strength. GAAGAA and TACTAAC were the motifs with the greatest impact on acceptor and donor strength, based on a comprehensive search in the k-mer space.

Role of exon and intron lengths in splicing

To study the effect of exon length on splicing, we filtered out the test set exons which were either the first or last exon. This filtering step removed 1,652 out of the 14,289 exons. We sorted the remaining 12,637 exons in the order of increasing length. For each of them, we calculated a splicing score by averaging the acceptor score at the splice acceptor site and the donor score at the splice donor site using SpliceAI-80nt. We plot the splicing scores as a function of exon length in Figure 1F. Before plotting, we applied the following smoothing procedure: Let x denote the vector containing the lengths of the exons, and y denote the vector containing their corresponding splicing scores. We smoothed both x and y using an averaging window of size 2,500.

We repeated this analysis by calculating the splicing scores using SpliceAI-10k. In the background, we show the histogram of the lengths of the 12,637 exons considered for this analysis. We applied a similar analysis to study the effect of intron length on splicing, with the main difference being that it was not necessary to exclude the first and last exons.

Role of nucleosomes in splicing

We downloaded nucleosome data for the K562 cell line from the UCSC genome browser. We used the *HMGR* gene, which is in our test set, as an anecdotal example to demonstrate the impact of nucleosome positioning on SpliceAI-10k score. For each position p in the gene, we calculated its ''planted splicing score'' as follows:

- The 8 nucleotides from positions p+74 to p+81 were replaced by a donor motif AGGTAAGG.
- The 4 nucleotides from positions p-78 to p-75 were replaced by an acceptor motif TAGG.
- The 20 nucleotides from positions p-98 to p-79 were replaced by a poly-pyrimidine tract CCTCCTTTTTCCTCGCCCTC.

- The 7 nucleotides from positions p-105 to p-99 were replaced by a branch point sequence CACTAAC.
- The average of the acceptor score at p-75 and donor score at p+75 predicted by SpliceAI-10k is used as the planted splicing score.

The K562 nucleosome signal as well as the planted splicing score for the 5,000 positions from chr5:74,652,154 to chr5:74,657,153 is shown in Figure 1G.

To calculate the genome-wide Spearman correlation between these two tracks, we randomly chose one million intergenic positions which were at least 100,000 nt away from all canonical genes. For each of these positions, we calculated its planted splicing score as well as its average K562 nucleosome signal (window size of 50 was used for averaging). The correlation between these two values across the one million positions is shown in Figure 1G. We further sub-classified these positions based on their GC content (estimated using the nucleotides in between the planted acceptor and donor motifs) with a bin size of 0.02. We show the genome-wide Spearman correlation for each bin in Figure S2E.

For each of the 14,289 test set splice acceptors, we extracted nucleosome data within 50 nucleotides on each side and calculated its nucleosome enrichment as the average signal on the exon side divided by the average signal on the intron side. We sorted the splice acceptors in the increasing order of their nucleosome enrichment and calculated their acceptor scores using SpliceAI-80nt. The acceptor scores are plotted as a function of nucleosome enrichment in Figure S2F. Before plotting, the smoothing procedure used in Figure 1F was applied. We repeated this analysis using SpliceAI-10k and also for the 14,289 test set splice donors.

Enrichment of nucleosome signal at novel exons

For Figure 1H, we wanted to look at the nucleosome signal around predicted novel exons. To ensure that we were looking at highly confident novel exons, we only selected singleton variants (variants present in a single GTEx individual) where the predicted gained junction was entirely private to the individual with the variant. Additionally, to remove confounding effects from nearby exons, we only looked at intronic variants at least 750 nt away from annotated exons. We downloaded nucleosome signals for the GM12878 and K562 cell lines from the UCSC browser and extracted the nucleosome signal within 750 nt from each of the predicted novel acceptor or donor sites. We averaged the nucleosome signal between the two cell lines and flipped the signal vectors for variants that overlapped genes on the negative strand. We shifted the signal from acceptor sites by 70 nt to the right and the signal from donor sites by 70 nt to the left. After shifting, the nucleosome signal for both acceptor and donor sites was centered at the middle of an idealized exon of length 140 nt, which is the median length of exons in the GENCODE annotations. We finally averaged all shifted signals and smoothed the resulting signal by computing the mean within an 11 nt window centered at each position.

To test for an association, we selected random singleton SNVs, that were at least 750 nt away from annotated exons and were predicted by the model to have no effect on splicing (Δ Score < 0.01). We created 1000 random samples of such SNVs, each sample having as many SNVs as the set of splice-site gain sites that were used for Figure 1H (128 sites). For each random sample, we computed a smoothed average signal as described above. Since the random SNVs were not predicted to create novel exons, we centered the nucleosome signal from each SNV at the SNV itself and randomly shifted either 70 nt to the left or 70 nt to the right. We then compared the nucleosome signal at the middle base of Figure 1H to the signals obtained from the 1000 simulations at that base. An empirical p value was computed as the fraction of simulated sets that had a middle value greater or equal to that observed for the splice-site gain variants.

Analyses on the GTEx RNA-seq dataset

Δ Score of a single nucleotide variant

We quantified the splicing change due to a single nucleotide variant as follows: We first used the reference nucleotide and calculated the acceptor and donor scores for 101 positions around the variant (50 positions on each side). Suppose these scores are denoted by the vectors a_{ref} and d_{ref} respectively. We then used the alternate nucleotide and recalculated the acceptor and donor scores. Let these scores be denoted by the vectors a_{alt} and d_{alt} respectively.

We evaluated the following four quantities:

$$\Delta \text{ Score } (\text{acceptor gain}) = \max(a_{alt} - a_{ref})$$

$$\Delta \text{ Score } (\text{acceptor loss}) = \max(a_{ref} - a_{alt})$$

$$\Delta \text{ Score } (\text{donor gain}) = \max(d_{alt} - d_{ref})$$

$$\Delta \text{ Score } (\text{donor loss}) = \max(d_{ref} - d_{alt})$$

The maximum of these four scores is referred to as the Δ Score of the variant.

Criteria for quality control and filtering of variants

We downloaded the GTEx VCF and RNA-seq data from dbGaP (study accession phs000424.v6.p1; https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v6.p1).

We evaluated the performance of SpliceAI on autosomal SNVs that appeared in at most 4 individuals in the GTEx cohort. In particular, a variant was considered if it satisfied the following criteria in at least one individual A:

1. The variant was not filtered (the FILTER field of the VCF was PASS).
2. The variant was not marked as MULTI_ALLELIC and there was indeed a single ALT allele for that record.
3. Individual A was heterozygous for the variant.
4. The ratio alt_depth / (alt_depth + ref_depth) was between 0.25 and 0.75, where alt_depth and ref_depth are the number of reads supporting the alternative and reference allele in individual A respectively.
5. The total depth, alt_depth + ref_depth, was between 20 and 300 for individual A.
6. The variant overlapped a gene body region. Gene bodies were defined as the regions between the transcription starts and ends of canonical transcripts from GENCODE V24lift37.

For variants satisfying these criteria in at least one individual, we considered all individuals where the variant appeared (even if it did not satisfy the above criteria) as having the variant. We refer to variants appearing in a single individual as singleton and variants appearing in 2-4 individuals as common. We did not evaluate variants appearing in 5 or more individuals, in order to avoid considering variants near junctions used for training.

RNA-seq read alignment

We used OLego (Wu et al., 2013) to map the reads of the GTEx samples against the hg19 reference, allowing an edit distance of at most 4 between the query read and the reference (parameter -M 4). Note that OLego can operate completely *de novo* and does not require any gene annotations. Since OLego looks for the presence of splicing motifs at the ends of split reads, its alignments can be biased toward or against the reference around SNVs that disrupt or create splice sites respectively. To eliminate such biases, we further created an alternative reference sequence for each GTEx individual, by inserting into the hg19 reference all the SNVs of the individual with a PASS filter. We used OLego with the same parameters to map all samples from each individual against that individual's alternative reference sequence. For each sample, we then combined the two sets of alignments (against the hg19 reference and against the individual's alternative reference), by picking the best alignment for each read pair. To choose the best alignment for a read pair P, we used the following procedure:

1. If both reads of P were unmapped in both sets of alignments, we chose either the hg19 or the alternative alignments of P at random.
2. If P had more unmapped ends in one set of alignments than in the other (e.g., both ends of P were mapped against the alternative reference but only one end was mapped against hg19), we chose the alignment with both ends of P mapped.
3. If both ends of P were mapped in both sets of alignments, we chose the alignment with the fewest total mismatches, or a random one, if the number of mismatches was the same.

Detection of splice junctions

We used leafcutter_cluster, a utility in the leafcutter package (Li et al., 2018), to detect and count splice junctions in each sample. We required a single split read to support a junction and assumed a maximum intron length of 500Kb (parameters -m 1 -l 500000). To get a high-confidence set of junctions for training the deep learning model, we compiled the union of all leafcutter junctions across all samples and then removed from consideration junctions that met any of the following criteria:

1. Either end of the junction overlapped an ENCODE blacklist region (table wgEncodeDacMapabilityConsensusExcludable in hg19 from the UCSC genome browser) or a simple repeat (Simple Repeats track in hg19 from the UCSC genome browser).
2. Both ends of the junction were on non-canonical exons (based on the canonical transcripts from GENCODE version V24lift37).
3. The two ends of the junction were on different genes or either end was in a non-genic region.
4. Either end lacked the essential GT/AG dinucleotides.

Junctions that were present in 5 or more individuals were used to augment the list of GENCODE annotated splice junctions for the analyses on variant prediction (Figures 2, 3, 4, and 5). Links to the files containing the list of splice junctions used to train the model are provided in the Key Resources Table.

Although we used junctions detected by leafcutter for augmenting the training dataset, we noticed that, despite the use of relaxed parameters, leafcutter was filtering many junctions with good support in the RNA-seq data. Thus, for the GTEx RNA-seq validation analyses (Figures 2 and 3), we recomputed the set of junctions and junction counts directly from the RNA-seq read data. We counted all non-duplicate split-mapped reads with MAPQ at least 10 and with at least 5 nt aligned on each side of the junction. A read was allowed to span more than two exons, in which case the read was counted toward each junction with at least 5 nt of mapped sequence on both sides.

Definition of private junctions

A junction was considered private in individual A if it satisfied at least one of the following criteria:

1. The junction had at least 3 reads in at least one sample from A and was never observed in any other individual.
2. There were at least two tissues that satisfied both of the following two criteria:
   a. The average read count of the junction in samples from individual A in the tissue was at least 10.
   b. Individual A had at least twice as many normalized reads on average than any other individual in that tissue. Here, the normalized read count of a junction in a sample was defined as the number of reads of the junction normalized by the total number of reads across all junctions for the corresponding gene.

Tissues with fewer than 5 samples from other individuals (not A) were ignored for this test.

Enrichment of singleton SNVs around private junctions

If a private junction had exactly one end annotated, based on the GENCODE annotations, we considered it a candidate for an acceptor or donor gain and searched for singleton SNVs (SNVs appearing in a single GTEx individual) that were private in the same individual within 150 nt from the unannotated end. If a private junction had both ends annotated, we considered it a candidate for a private exon skipping event if it skipped at least one but no more than 3 exons of the same gene based on the GENCODE annotations. We then searched for singleton SNVs within 150 nt from the ends of each of the skipped exons. Private junctions with both ends absent from the GENCODE exon annotations were ignored, as a substantial fraction of these were alignment errors.

To compute the enrichment of singleton SNVs around novel private acceptors or donors (Figure 2B, bottom), we aggregated the counts of singleton SNVs at each position relative to the private junction. If the overlapping gene was on the negative strand, relative positions were flipped. We split SNVs into two groups: SNVs that were private in the individual with the private junction and SNVs that were private in a different individual. To smooth the resulting signals, we averaged counts in a 7 nt window centered at each position. We then computed the ratio of smoothed counts from the first group (private in the same individual) to the smoothed counts of the second group (private in a different individual). For novel private exon skips (Figure 2B, top), we followed a similar procedure, aggregating the counts of singleton SNVs around the ends of skipped exons.

Validation of model predictions in GTEx RNA-seq data

For either private variants (appearing in one individual in the GTEx cohort) or common variants (appearing in two to four individuals in the GTEx cohort), we obtained the predictions of the deep learning model for both the reference and the alternate alleles and computed the Δ Score. We also obtained the location where the model predicted the aberrant (novel or disrupted) junction to be. We then sought to determine whether there was evidence in the RNA-seq data supporting a splicing aberration in the individuals with the variant at the predicted location. In many cases, the model can predict multiple effects for the same variant, e.g., a variant disrupting an annotated splice donor could also increase usage of a suboptimal donor as in Figure S3A, in which case the model might predict both a donor loss at the annotated splice site and a donor gain at the suboptimal site. For this reason, we considered predicted splice site-creating and splice site-disrupting effects of each variant separately. Note that junctions appearing in less than five individuals were excluded during model training, to avoid evaluating the model on novel junctions it was trained on.

Validation of predicted cryptic splice mutations based on private splice junctions

For each singleton variant predicted to cause novel junction formation, we looked in the RNA-seq data to check if the predicted novel junction appeared only in the individual with the SNV and in no other GTEx individuals. Similarly, for a singleton variant predicted to cause loss of a splice site of exon X, we looked for novel exon skipping events from the previous canonical exon (the one upstream of X based on GENCODE annotations) to the next canonical exon (the one downstream of X) that appeared only in the individual with the variant and in no other individuals in GTEx. We excluded predicted losses if the splice site predicted to be lost by the model was not annotated in GENCODE or never observed in GTEx individuals without the variant. We also excluded predicted gains if the splice site predicted to be gained was already annotated in GENCODE. To extend this analysis to common variants (present in two to four individuals), we checked whether the affected junction was present in at least half the individuals with the variant, and absent in all individuals without the variant.

Using the requirement that the predicted aberrant splice event is private to the individuals with the variant, we could validate 40% of predicted high-scoring (Δ Score $\geq$ 0.5) acceptor and donor gains, but only 3.4% of predicted high-scoring losses and 5.6% of essential GT or AG disruptions (at a false validation rate of $<$ 0.2% based on permutations – see section ''Estimating false validation rates''). The reason for the discrepancy in the validation rates of gains and losses is twofold. First, unlike gains, exon-skipping events are rarely entirely private to the individuals with the variant, because exons are often skipped at a low baseline level, which can be observed with sufficiently deep RNA-seq. Second, splice-site losses can have other effects besides increasing exon skipping, such as increasing intron retention or increasing the usage of alternative suboptimal splice sites. For these reasons, we did not rely entirely on private novel junctions for validating the model's predictions and also validated variants based on quantitative evidence for the increase or decrease of the usage of the junction predicted to be affected in the individuals with the variant.

Validation of predicted cryptic splice mutations through quantitative criteria

For a junction j from sample s, we obtained a normalized junction count c_{js}:

$$c_{js} = \text{asinh}\left(\frac{r_{js}}{\sum_g r_{gs}}\right) \quad (1)$$

Here, r_{js} is the raw junction count for junction j in sample s, and the sum in the denominator is taken over all other junctions between GENCODE annotated acceptors and donors of the same gene as j. The asinh transformation is defined as $asinh(x) = \ln(x + \sqrt{x^2+1})$. It is similar to the log transformation often used to transform RNA-seq data, however, it is defined at 0, thus eliminating the need for pseudocounts, which would have distorted values substantially, since many junctions, especially novel ones, have low or zero counts in many samples. The asinh transformation behaves like a log transformation for large values but is close to linear for small values. For this reason, it is often used in datasets (such as RNA-seq or ChIP-seq datasets) with a large number of near zero values. As described below, in the section "Consideration criteria for validation," samples where the denominator in Equation (1) was below 200 were excluded for all validation analyses, thus avoiding numerical issues.

For each gained or lost junction j predicted to be caused by an SNV appearing in a set of individuals I, we computed the following z-score in each tissue t separately:

$$z_{jt} = \frac{mean_{s\in A_t}(c_{js}) - mean_{s'\in U_t}(c_{js'})}{std_{s'\in U_t}(c_{js'})} \quad (2)$$

where A_t is the set of samples from individuals in I in tissue t and U_t is the set of samples from all other individuals in tissue t. Note that there might be multiple samples in the GTEx dataset for the same individual and tissue. As before, c_{js} is the count for junction j in sample s. For predicted losses, we also computed a similar z-score for the junction k skipping the putatively affected exon:

$$z_{kt} = \frac{mean_{s'\in U_t}(c_{ks'}) - mean_{s\in A_t}(c_{ks})}{std_{s'\in U_t}(c_{ks'})} \quad (3)$$

Note that a loss that resulted in skipping would lead to a relative decrease of the lost junction and a relative increase in skipping. This justifies the reversion of the difference in the numerators of z_{jt} and z_{kt}, so both of these scores would tend to be negative for a real splice site loss.

Finally, we computed the median z-score across all considered tissues. An acceptor or donor gain prediction was considered validated if the median of z-scores from Equation (2) across tissues was positive or the predicted gained junction appeared in at least half of the individuals with the variant and in no other individuals (as described in the section "Validation of predicted cryptic splice mutations based on private splice junctions" above). For losses, we computed the median of each of the z-scores from Equations (2) and (3) separately. An acceptor or donor loss prediction was considered validated if any of the following was true:

1. The median of the z-scores from Equation (2), quantifying the relative loss of the junction was less than the 5th percentile of the corresponding value in permuted data (−1.46) and the median of the z-scores from Equation (3), quantifying the relative change in skipping was non-positive (i.e., zero, negative, or missing, which would be the case if the skipping junction was not observed in any individual). In other words, there was strong evidence for a reduction in the usage of the affected junction and no evidence suggesting a decrease in skipping in the affected individual.
2. The median of z-scores from Equation (3) was less than the 5th percentile of the corresponding value in permuted data (−0.74) and the median of z-scores from Equation (3) was non-positive.
3. The median of z-scores from Equation (2) was less than the 1st percentile of the corresponding values in permuted data (−2.54).
4. The median of z-scores from Equation (3) was less than the 1st percentile of the corresponding values in permuted data (−4.08).
5. The junction skipping the affected exon was observed in at least half of the individuals with the variant and in no other individuals (as described in the section "Validation of predicted cryptic splice mutations based on private splice junctions" above).

A description of the permutations used to get the above cutoffs is given in the section "Estimating false validation rates."

Empirically, we observed that we needed to apply stricter validation criteria for losses compared to gains, since, as explained in the section "Validation of predicted cryptic splice mutations based on private splice junctions," losses tend to result in more mixed effects than gains. Observing a novel junction near a private SNV is very unlikely to occur by chance, so even minor evidence of the junction should be sufficient for validation. In contrast, most predicted losses resulted in weakening of an existing junction, and such weakening is harder to detect than the on-off change caused by gains and more likely to be attributed to noise in the RNA-seq data.

Inclusion criteria for validation analysis

To avoid computing z-scores in the presence of low counts or poor coverage, we used the following criteria to filter variants for the validation analysis:

1. Samples were considered for the above z-score calculation only if they expressed the gene ($\sum_g r_{gs} > 200$ in Equation (1)).

2. A tissue was not considered for a loss or gain z-score calculation if the average count of the lost or "reference" junction respectively in individuals without the variant was less than 10. The "reference" junction is the canonical junction used prior to the gain of the novel junction, based on GENCODE annotations (see section on effect size calculation for details). The intuition is that we should not try to validate a splice-loss variant that affects a junction not expressed in control individuals. Similarly, we should not try to validate a splice-gain variant if control individuals did not sufficiently express transcripts spanning the affected site.
3. In the case of a predicted splice-site loss, samples from individuals without the variant were only considered if they had at least 10 counts of the lost junction. In the case of a predicted acceptor or donor gain, samples from control individuals were only considered if they had at least 10 counts of the "reference" junction. The intuition is that even in a tissue with large average expression of the affected junction (i.e., passing criterion 2.), different samples could have vastly different sequencing depths, so only control samples with sufficient expression should be included.
4. A tissue was considered only if there was at least one sample passing the above criteria from individuals with the variant, and at least 5 samples passing the above criteria from at least 2 distinct control individuals.

Variants for which there were no tissues satisfying the above criteria for consideration were deemed non-ascertainable and were excluded when calculating the validation rate. We excluded from validation splice site gain predictions at GENCODE-annotated splice sites. Similarly, for splice-loss variants, we only considered those that decrease the scores of existing GENCODE-annotated splice sites. Overall, 55% and 44% of high-scoring (Δ Score $\geq$ 0.5) predicted gains and losses respectively were considered ascertainable and used for the validation analysis.

Estimating false validation rates

To ensure that the above procedure had reasonable true validation rates, we first looked at SNVs that appear in 1-4 GTEx individuals and disrupt essential GT/AG dinucleotides. We argued that such mutations almost certainly affect splicing so their validation rate should be close to 100%. Among such disruptions, 39% were ascertainable based on the criteria described above, and among the ascertainable ones, the validation rate was 81%. To estimate the false validation rate, we permuted the individual labels of the SNV data. For each SNV that appeared in k GTEx individuals, we picked a random subset of k GTEx individuals and assigned the SNV to them. We created 10 such randomized datasets and repeated the validation process on them. The validation rate in the permuted datasets was 1.7%–2.1% for gains and 4.3%–6.9% for losses, with a median of 1.8% and 5.7% respectively. The higher false validation rate for losses and the relatively low validation rate of essential disruptions are due to the difficulty in validating splice-site losses as highlighted in the section "Validation of predicted cryptic splice mutations based on private splice junctions."

Calculating the effect size of cryptic splice variants in RNA-seq data

We defined the effect size of a variant as the fraction of transcripts of the affected gene that changed splicing patterns due to the variant (e.g., the fraction that switched to a novel acceptor or donor). As a reference example for a predicted splice-gain variant, consider the variant in Figure 2C. For a predicted gained donor A, we first identified the junction (AC) to the closest annotated acceptor C. We then identified a "reference" junction (BC), where B $\neq$ A is the annotated donor closest to A. In each sample s, we then computed the relative usage of the novel junction (AC) compared to the reference junction (BC):

$$u_{(AB)s} = \frac{r_{(AC)s}}{r_{(AC)s} + r_{(BC)s}} \tag{4}$$

Here, $r_{(AC)s}$ is the raw read count of junction (AC) in sample s. For each tissue, we computed the change in the usage of the junction (AC) between the individuals with the variant and all other individuals:

$$mean_{s \in A_t} u_{(AC)s} - mean_{s' \in U_t} u_{(AC)s'} \tag{5}$$

where A_t is the set of samples from individuals with the variant in tissue t and U_t is the set of samples from other individuals in tissue t. The final effect size was computed as the median of the above difference across all considered tissues. The computation was similar in the case of a gained acceptor or in the case where the splice-site creating variant was intronic. A simplified version of the effect size computation (assuming a single sample from individuals with and without the variant) is shown in Figure 2C.

For a predicted loss, we first computed the fraction of transcripts that skipped the affected exon. The computation is demonstrated on Figure S3A. For a predicted loss of a donor C, we identified the junction (CE) to the next downstream annotated exon, as well as the junction (AB) from the upstream exon to the putatively affected one. We quantified the fraction of transcripts that skipped the affected exon as follows:

$$k_{(AE)s} = \frac{r_{(AE)s}}{r_{(AE)s} + mean\left(r_{(AB)s} + r_{(CE)s}\right)} \tag{6}$$

As for gains, we then computed the change in the skipped fraction between samples from individuals with the variant and samples from individuals without the variant:

$$mean_{s\in A_t} k_{(AE)s} - mean_{s'\in U_t} k_{(AE)s'} \quad (7)$$

The fraction of skipped transcripts as computed above does not fully capture the effects of an acceptor or donor loss, as such a disruption could also lead to increased levels of intron retention or usage of suboptimal splice sites. To account for some of these effects, we also computed the usage of the lost junction (CE) relative to the usage of other junctions with the same acceptor E:

$$I_{(CE)s} = \frac{r_{(CE)s}}{\sum r_{(\cdot E)s}} \quad (8)$$

Here, $\sum r_{(\cdot E)s}$ is the sum of all junctions from any (annotated or novel) donor to the acceptor E. This includes the affected junction (CE), the skipping junction (AE), as well as potential junctions from other suboptimal donors that compensated for the loss of C, as illustrated in the example in Figure S3A. We then computed the change in the relative usage of the affected junction:

$$mean_{s'\in U_t} I_{(CE)s'} - mean_{s\in A_t} I_{(CE)s} \quad (9)$$

Note that, unlike (5) and (7), which measure the increase in usage of the gained or skipping junction in individuals with the variant, in (9) we want to measure the decrease in usage of the lost junction, hence the reversion of the two parts of the difference. For each tissue, the effect size was computed as the maximum of (7) and (9). As for gains, the final effect size for the variant was the median effect size across tissues.

Inclusion criteria for effect size analysis

A variant was considered for effect size computation only if it was deemed validated based on the criteria described in the previous section. To avoid calculating the fraction of aberrant transcripts on very small numbers, we only considered samples where the counts of the aberrant and reference junctions were both at least 10. Because most cryptic splice variants were in the intron, the effect size could not be computed directly by counting the number of reference and alternate reads overlapping the variant. Hence, the effect size of losses is calculated indirectly from the decrease in the relative usage of the normal splice junction. For the effect size of novel junction gains, the aberrant transcripts can be impacted by nonsense mediated decay, attenuating the observed effect sizes. Despite the limitations of these measurements, we observe a consistent trend toward smaller effect sizes for lower-scoring cryptic splice variants across both gain and loss events.

Expected effect size of fully penetrant heterozygous private SNVs

For a fully penetrant splice-site creating variant that causes all transcripts from the variant haplotype of the individuals with the variant to switch to the novel junction, and assuming that the novel junction does not occur in control individuals, the expected effect size would be 0.5 by Equation (5).

Similarly, if a heterozygous SNV causes a novel exon skipping event, and all transcripts of the affected haplotype switched to the skipping junction, the expected effect size in Equation (7) is 0.5. If all transcripts from individuals with the variant switched to a different junction (either the skipping junction, or another compensating one), the ratio in Equation (8) would be 0.5 in samples from individuals with the variant and 1 in samples from other individuals, so the difference in Equation (9) would be 0.5. This assumes that there was no skipping or other junctions into acceptor E in individuals without the variant. It also assumes that the splice site disruption does not trigger intron retention. In practice, at least low levels of intron retention are often associated with splice site disruptions. Furthermore, exon skipping is widespread, even in the absence of splice-altering variants. This explains why the measured effect sizes are below 0.5, even for variants disrupting essential GT/AG dinucleotides.

The expectation of effect sizes of 0.5 for fully penetrant heterozygous variants also assumes that the variant did not trigger nonsense-mediated decay (NMD). In the presence of NMD, both the numerator and the denominator of Equations (4), (6), and (8) would drop, thus diminishing the observed effect size.

Fraction of transcripts degraded through nonsense-mediated decay (NMD)

For Figure 2C since the variant was exonic, we could count the number of reads that spanned the variant and had the reference or the alternate allele (“Ref (no splicing)” and “Alt (no splicing)” respectively). We also counted the number of reads that spliced at the novel splice site, and that presumably carried the alternate allele (“Alt (novel junction)”). In the example of Figure 2C and in many other cases we looked at, we observed that the total number of reads coming from the haplotype with the alternate allele (the sum of “Alt (no splicing)” and “Alt (novel junction)”) was less than the number of reads with the reference allele (“Ref (no splicing)”). Since we believe that we have eliminated reference biases during read mapping, by mapping to both the reference and alternate haplotypes, and assuming that the number of reads is proportional to the number of transcripts with each allele, we were expecting that the reference allele would account for half of the reads at the variant locus. We assume that the “missing” alternate allele reads correspond to transcripts from the alternate allele haplotype that spliced at the novel junction and were degraded through nonsense mediated decay (NMD). We called this group “Alt (NMD).”

To determine whether the difference between the observed number of reference and alternate reads was significant we computed the probability of observing Alt (no splicing) + Alt (novel junction) (or fewer) reads under a binomial distribution with success

probability 0.5 and a total number of trials of Alt (no splicing) + Alt (novel junction) + Ref (no splicing). This is a conservative p value since we are underestimating the total number of "trials" by not counting the potentially degraded transcripts. The fraction of NMD transcripts in Figure 2C was computed as the number of "Alt (NMD)" reads over the total number of reads splicing at the novel junction (Alt (NMD) + Alt (novel junction)).

Sensitivity of the network at detecting cryptic splice junctions

For evaluating the sensitivity of the SpliceAI model (Figure 2F), we used SNVs that were within 20 nt from the affected splice site (i.e., the novel or disrupted acceptor or donor) and not overlapping the essential GT/AG dinucleotide of an annotated exon, and had an estimated effect size of at least 0.3 (see section "Effect size calculation"). In all sensitivity plots, SNVs were defined as being "near exons" if they overlapped an annotated exon or were within 50 nt of the boundaries of an annotated exon. All other SNVs were considered "deep intronic." Using this truth dataset of strongly supported cryptic splice sites, we evaluated our model at varying Δ Score thresholds and report the fraction of cryptic splice sites in the truth dataset that are predicted by the model at that cutoff.

Comparison with existing splicing prediction models

We performed a head-to-head comparison of SpliceAI-10k, MaxEntScan (Yeo and Burge, 2004), GeneSplicer (Pertea et al., 2001) and NNSplice (Reese et al., 1997) with respect to various metrics. We downloaded the MaxEntScan and GeneSplicer software from http://genes.mit.edu/burgelab/maxent/download/ and http://www.cs.jhu.edu/~genomics/GeneSplicer/ respectively. NNSplice is not available as a downloadable software, so we downloaded the training and testing sets from http://www.fruitfly.org/data/seq_tools/datasets/Human/GENIE_96/splicesets/, and trained models with the best performing architectures described in (Reese et al., 1997). As a sanity check, we reproduced the test set metrics reported in (Reese et al., 1997).

To evaluate the top-*k* accuracies and the area under the precision-recall curves of these algorithms, we scored all the positions in the test set genes and lincRNAs with each algorithm.

MaxEntScan and GeneSplicer outputs correspond to log odds ratios, whereas NNSplice and SpliceAI-10k outputs correspond to probabilities. To ensure that we gave MaxEntScan and GeneSplicer the best chance of success, we calculated Δ Scores using them with the default output as well as with a transformed output where we first transform their outputs so that they correspond to probabilities. More precisely, the default output of MaxEntScan corresponds to

$$x = \log_2 \frac{p(\textit{splice site})}{p(\textit{not a splice site})}$$

which, after the transformation $(2^x/(2^x+1))$, corresponds to the desired quantity. We compiled the GeneSplicer software twice, once by setting the RETURN_TRUE_PROB flag to 0 and once by setting it to 1. We picked the output strategy that led to the best validation rate against RNA-seq data (MaxEntScan: transformed output, GeneSplicer: default output).

To compare the validation rate and sensitivity of the various algorithms (Figure 2G), we found cutoffs at which all algorithms predicted the same number of gains and losses genome-wide. That is, for each cutoff on the SpliceAI-10k Δ Score values, we found the cutoffs at which each competing algorithm would make the same number of gain predictions and the same number of loss predictions as SpiceNet-10k. The chosen cutoffs are given in Table S2.

Comparison of variant prediction for singleton versus common variants

We performed the validation and sensitivity analysis (as described in sections "Sensitivity analysis" and "Validation of model predictions") separately for singleton SNVs and SNVs appearing in 2-4 GTEx individuals (Figures S3B and S3C). To test whether the validation rate differed significantly between singleton and common variants, we performed a Fisher Exact test, comparing the validation rates in each Δ Score group (0.2 – 0.35, 0.35 – 0.5, 0.5 – 0.8, 0.8 – 1) and for each predicted effect (acceptor or donor gain or loss). After Bonferroni correction to account for 16 tests, all p values were greater than 0.05. We similarly compared the sensitivity for detecting singleton or common variants. We used a Fisher Exact test to test whether the sensitivity differed significantly between the two groups of variants. We considered deep-intronic variants and variants near exons separately and performed Bonferroni correction for two tests. None of the p values were significant using a 0.05 cutoff. We therefore combined singleton and common GTEx variants and considered them together for the analyses presented in Figures 2 and 3.

Comparison of variant prediction on the training versus testing chromosomes

We compared the validation rate on RNA-seq and sensitivity of SpliceAI-10k between variants on the chromosomes used during training and variants on the rest of the chromosomes (Figures S4A and S4B). All p values were greater than 0.05 after Bonferroni correction. We also computed the fraction of deleterious variants separately for variants on the training and test chromosomes, as described in the section "Fraction of deleterious variants" below (Figure S4C). For each Δ Score group and each type of variant, we used a Fisher Exact test to compare the number of common and rare variants between training and test chromosomes. After Bonferroni correction for 12 tests, all p values were greater than 0.05. Finally, we computed the number of cryptic splice *de novo* variants on the training and test chromosomes (Figure S4D) as described in the section "Enrichment of *de novo* mutations per cohort."

Comparison of variant prediction across different types of cryptic splice variants

We split predicted splice site-creating variants into three groups: variants creating a novel GT or AG splice dinucleotide, variants overlapping the rest of the splicing motif (positions around the exon-intron boundary up to 3 nt into the exon and 8 nt into the intron), and variants outside the splice motif (Figures S3E and S3F). For each Δ Score group (0.2 – 0.35, 0.35 – 0.5, 0.5 – 0.8, 0.8 – 1), we

performed a χ^2 test to test the hypothesis that the validation rate is uniform across the three types of splice site-creating variants. All tests yielded p values greater than 0.3 even before multiple hypothesis correction. To compare the effect size distribution between the three types of variants, we used a Mann-Whitney *U* test and compared all three pairs of variant types for each Δ Score group (for a total of 4 × 3 = 12 tests). After Bonferroni correction for 12 tests, all p values were greater than 0.3.

Detection of tissue-specific splice-gain variants

For Figure 3C, we wanted to test whether the usage rate of novel junctions was uniform across tissues expressing the affected gene. We focused on SNVs that created novel private splice sites, that is, SNVs resulting in a gained splice junction which only appeared in at least half of the individuals with the variant and in no other individuals. For each such novel junction j, we computed, in each tissue t, the total counts of the junction across all samples from individuals with the variant in the tissue: $\sum_{s\in A_t} r_{js}$. Here A_t is the set of samples from individuals with the variant in tissue t. Similarly, we computed the total counts of all annotated junctions of the gene for the same samples $\sum_{s\in A_t}\sum_{g} r_{gs}$, where g indexes the annotated junctions of the gene. The relative usage of the novel junction in tissue t, normalized against the gene's background counts, can then be measured as:

$$m_t = \frac{\sum_{s\in A_t} r_{js}}{\sum_{s\in A_t}\left(r_{js} + \sum_g r_{gs}\right)}$$

We also computed the average usage of the junction across tissues:

$$m = \frac{\sum_t \sum_{s\in A_t} r_{js}}{\sum_t \sum_{s\in A_t}\left(r_{js} + \sum_g r_{gs}\right)}$$

We wanted to test the hypothesis that the relative usage of the junction is uniform across tissues and equal to m. We thus performed a χ^2 test comparing the observed tissue counts $\sum_{s\in A_t} r_{js}$ to the expected counts under the assumption of a uniform rate, $m\sum_{s\in A_t}\left(r_{js} + \sum_g r_{gs}\right)$. A splice-site creating variant was considered tissue-specific if the Bonferroni-corrected χ^2 p value was less than 10^{-2}. The degrees of freedom for the test are T – 1, where T is the number of considered tissues. Only tissues that satisfied the consideration criteria described in the validation section were used in the test. Further, to avoid cases with low counts, where the uniformity test was underpowered, we only tested for uniformity variants with at least three considered tissues, at least one aberrant read per tissue on average (i.e., $m > 1$), and at least 15 aberrant reads in total across all considered tissues (i.e., $\sum_t \sum_{s\in A_t} r_{js} > 15$). We ignored all variants with Δ Score less than 0.35, since this class of variants has generally low effect sizes and low junction counts. We observed that the fraction of tissue-specific variants was very low for this class, but we believe that this was due to power issues.

Analyses on the ExAC and gnomAD datasets

Variant filtering

We downloaded the Sites VCF release 0.3 file (60,706 exomes) from the ExAC browser (Lek et al., 2016) and the Sites VCF release 2.0.1 file (15,496 whole genomes) from the gnomAD browser. We created a filtered list of variants from them in order to evaluate SpliceAI-10k. In particular, variants which satisfied the following criteria were considered:

- The FILTER field was PASS.
- The variant was a single nucleotide variant, and there was only one alternate nucleotide.
- The AN field (total number of alleles in called genotypes) had a value at least 10,000.
- The variant was in between the transcription start and end site of a canonical GENCODE transcript.

A total of 7,615,051 and 73,099,995 variants passed these filters in the ExAC and gnomAD datasets respectively.

Fraction of deleterious variants

For this analysis, we considered only those variants in the ExAC and gnomAD filtered lists which were singleton or common (allele frequency (AF) ≥ 0.1%) in the cohort. We sub-classified these variants based on their genomic position according to the GENCODE canonical annotations:

- Exonic: This group consists of synonymous ExAC variants (676,594 singleton and 66,524 common). Missense variants were not considered here to ensure that most of the deleteriousness of the variants in this group was due to splicing changes.
- Near intronic: This group consists of intronic ExAC variants which are between 3 and 50 nt from a canonical exon boundary. More precisely, for the analysis of acceptor gain/loss and donor gain/loss variants, only those variants which were 3-50 nt from a splice acceptor and donor respectively were considered (575,636 singleton and 48,362 common for acceptor gain/loss, 567,774 singleton and 50,614 common for donor gain/loss).

- Deep intronic: This group consists of intronic gnomAD variants which are more than 50 nt away from a canonical exon boundary (34,150,431 singleton and 8,215,361 common).

For each variant, we calculated its Δ Scores for the four splice types using SpliceAI-10k. Then, for each splice type, we constructed a 2×2 chi-square contingency table where the two rows corresponded to predicted splice-altering variants (Δ Score in the appropriate range for the splice type) versus predicted not splice-altering variants (Δ Score < 0.1 for all splice types), and the two columns corresponded to singleton versus common variants. For splice-gain variants, we filtered those that occur at already existing GENCODE-annotated splice sites. Similarly, for splice-loss variants, we only considered those that decrease the scores of existing GENCODE-annotated splice sites. The odds ratio was calculated, and the fraction of deleterious variants was estimated as

$$\left(1 - \frac{1}{\text{Odds ratio}}\right) \times 100\%$$

The protein-truncating variants in the ExAC and gnomAD filtered lists were identified as follows:

- Nonsense: VEP (McLaren et al., 2016) consequence was 'stop_gained' (44,046 singleton and 722 common in ExAC, 20,660 singleton and 970 common in gnomAD).
- Frameshift: VEP consequence was 'frameshift_variant'. The single nucleotide variant criterion during variant filtering was relaxed in order to create this group (48,265 singleton and 896 common in ExAC, 30,342 singleton and 1,472 common in gnomAD).
- Essential acceptor/donor loss: The variant was in the first or last two positions of a canonical intron (29,240 singleton and 481 common in ExAC, 12,387 singleton and 746 common in gnomAD).

The 2×2 chi-square contingency table for protein-truncating variants was constructed for the ExAC and gnomAD filtered lists, and used to estimate the fraction of deleterious variants. Here, the two rows corresponded to protein-truncating versus synonymous variants, and the two columns corresponded to singleton versus common variants as before.

The results for the ExAC (exonic and near intronic) and gnomAD (deep intronic) variants are shown in Figures 4B and 4D respectively.

Frameshift versus in-frame splice gain

For this analysis, we focused our attention on the ExAC variants which were exonic (synonymous only) or near intronic, and were singleton or common (AF ≥ 0.1%) in the cohort. To classify an acceptor gain variant as in-frame or frameshift, we measured the distance between the canonical splice acceptor and the newly created splice acceptor, and checked whether it was a multiple of 3 or not. We classified donor gain variants similarly by measuring the distance between the canonical splice donor and the newly created splice donor.

The fraction of deleterious in-frame splice gain variants was estimated from a 2×2 chi-square contingency table where the two rows corresponded to predicted in-frame splice gain variants (Δ Score ≥ 0.8 for acceptor or donor gain) versus predicted not splice-altering variants (Δ Score < 0.1 for all splice types), and the two columns corresponded to singleton versus common variants. This procedure was repeated for frameshift splice gain variants by replacing the first row in the contingency table with predicted frameshift splice gain variants.

To calculate the p value shown in Figure 4C, we constructed a 2×2 chi-square contingency table using only the predicted splice gain variants. Here, the two rows corresponded to in-frame versus frameshift splice gain variants, and the two columns corresponded to singleton versus common variants as before.

Number of cryptic splice variants per individual

To estimate the number of rare functional cryptic splice variants per individual (Figure 4E), we first simulated 100 gnomAD individuals by including each gnomAD variant in each allele with a probability equal to its allele frequency. In other words, each variant was sampled twice independently for each individual to mimic diploidy. We counted the number of rare (AF < 0.1%) exonic (synonymous only), near intronic, and deep intronic variants per person which had a Δ Score greater than or equal to 0.2, 0.2, and 0.5 respectively. These are relatively permissive Δ Score thresholds which optimize sensitivity while ensuring that at least 40% of the predicted variants are deleterious. At these cutoffs, we obtained an average of 7.92 synonymous/near intronic and 3.03 deep intronic rare cryptic splice variants per person. Because not all of these variants are functional, we multiplied the counts by the fraction of variants that are deleterious at these cutoffs.

Analyses on the DDD and ASD datasets

Cryptic splicing de novo *mutations*

We obtained published *de novo* mutations (DNMs). These included 3,953 probands with autism spectrum disorder (Iossifov et al., 2014; De Rubeis et al., 2014), 4,293 probands from the Deciphering Developmental Disorders cohort (McRae et al., 2017), and 2,073 healthy controls (Iossifov et al., 2014). Low quality DNMs were excluded from analyses (ASD and healthy controls: Confidence = = lowConf, DDD: PP(DNM) < 0.00781, (McRae et al., 2017)). The DNMs were evaluated with the network, and we used Δ Score values (see methods above) to classify cryptic splice mutations depending on the context. We only considered

mutations annotated with VEP consequences of synonymous_variant, splice_region_variant, intron_variant, 5_prime_UTR_variant, 3_prime_UTR_variant, or missense_variant. We used sites with Δ Score > 0.1 for Figures 5, S5D, and S5E, and sites with Δ Score > 0.2 for Figures S5A–S5C.

Enrichment of de novo *mutations per cohort*

Candidate cryptic splice DNMs were counted in each of the three cohorts. The DDD cohort did not report intronic DNMs > 8 nt away from exons and so regions > 8 nt from exons were excluded from all cohorts for the purposes of the enrichment analysis to enable equivalent comparison between the DDD and ASD cohorts (Figure 5A). We also performed a separate analysis which excluded mutations with dual cryptic splicing and protein-coding function consequences to demonstrate that the enrichment is not due to the enrichment of mutations with protein-coding effects within the affected cohorts (Figures S5A and S5B). Counts were scaled for differing ascertainment of DNMs between cohorts by normalizing the rate of synonymous DNMs per individual between cohorts, using the healthy control cohort as the baseline. We compared the rate of cryptic splice DNMs per cohort using an E-test to compare two Poisson rates.

The plotted rates for enrichment over expectation (Figure 5C) were adjusted for the lack of DNMs > 8 nt from exons by scaling upward by the proportion of all cryptic splice DNMs expected to occur between 9-50 nt away from exons using a trinucleotide sequence context model (see below, "Enrichment of *de novo* mutations per gene"). The silent-only diagnostic proportion and excess of cryptic sites (Figures S5B and S5C) were also adjusted for the lack of missense sites by scaling the cryptic count by the proportion of cryptic splice sites expected to occur at missense sites versus synonymous sites. The impact of Δ Score threshold on enrichment was assessed by calculating the enrichment of cryptic splice DNMs within the DDD cohort across a range of cutoffs. For each of these the observed: expected odds ratio was calculated, along with the excess of cryptic splice DNMs.

Proportion of pathogenic DNMs

The excess of DNMs compared to baseline mutation rates can be considered the pathogenic yield within a cohort. We estimated the excess of DNMs by functional type within ASD and DDD cohorts, against the background of the healthy control cohort (Figure 5B). The DNM counts were normalized to the rate of synonymous DNMs per individual as described above. The DDD cryptic splice count was adjusted for the lack of DNMs 9-50 nt away from introns as described above. For both ASD and DDD cohorts, we also adjusted for the missing ascertainment of deep intronic variants > 50 nt away from exons, using the ratio of near-intronic (< 50 nt) versus deep intronic (> 50 nt) cryptic splice variants from the negative selection analysis (Figure 2G).

Enrichment of de novo *mutations per gene*

We determined null mutation rates for every variant in the genome using a trinucleotide sequence context model (Samocha et al., 2014). We used the network to predict the Δ Score for all possible single nucleotide substitutions within exons and up to 8 nt into the intron. Based on the null mutation rate model, we obtained the expected number of *de novo* cryptic splice mutations per gene (using Δ Score > 0.2 as a cutoff).

As per the DDD study (McRae et al., 2017), genes were assessed for enrichment of DNMs compared to chance under two models, one considering only protein-truncating (PTV) DNMs, and one considering all protein-altering DNMs (PTVs, missense, and in-frame indels). For each gene, we selected the most significant model, and adjusted the p value for multiple hypothesis testing. These tests were run once where we didn't consider cryptic splice DNMs or cryptic splice rates (the default test, used in the original DDD study), and once where we also counted cryptic splice DNMs and their mutation rates. We report additional candidate genes that were identified as genes with FDR-adjusted p value < 0.01 when including cryptic splice DNMs, but FDR-adjusted p value > 0.01 when not including cryptic splice DNMs (the default test). Enrichment tests were performed similarly for the ASD cohort.

Validation of predicted cryptic splice sites in lymphoblastoid cell lines

We selected high confidence *de novos* from affected probands in the Simons Simplex Collection, with at least RPKM > 1 RNA-seq expression in lymphoblastoid cell lines. We selected *de novo* cryptic splice variants for validation based on a Δ Score threshold > 0.1 for splice loss variants and a Δ Score threshold > 0.5 for splice gain variants. Because the cell lines needed to be procured far in advance, these thresholds reflect an earlier iteration of our methods, compared to the thresholds we adopted elsewhere in the paper (Figures 2G and 5A–5D), and the network did not include GTEx novel splice junctions for model training.

Lymphoblastoid cell lines were obtained from the SSC for these probands. Cells were cultured in Culture Medium (RPMI 1640, 2mM L-glutamine, 15% fetal bovine serum) to a maximum cell density of 1×10^6 cells/ml. When cells reached maximum density, they were passaged by dissociating the cells by pipetting up and down 4 or 5 times and seeding to a density of 200,000-500,000 viable cells/ml. Cells were grown under 37°C, 5% CO_2 conditions for 10 days. Approximately 5×10^5 cells were then detached and spun down at 300 × g for 5 min at 4°C. RNA was extracted using RNeasy Plus Micro Kit (QIAGEN) following manufacturer's protocol. RNA quality was assessed using Agilent RNA 6000 Nano Kit (Agilent Technologies) and ran on Bioanalyzer 2100 (Agilent Technologies). RNA-seq libraries were generated by TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold Set A (Illumina). Libraries were sequenced on HiSeq 4000 instruments at Center for Advanced Technology (UCSF) using 150-nt single-end sequencing at a coverage of 270-388 million reads (median 358 million reads).

Sequencing reads for each patient were aligned against a reference created from hg19 by substituting *de novo* variants of the patient (Iossifov et al., 2014) with the corresponding alternative allele. Sequencing coverage, splice junction usage, and transcript locations were plotted. We evaluated the predicted cryptic splice sites as described above in the validation of model predictions section. Thirteen novel splice sites (9 novel junction, 4 exon skipping) were confirmed as they were only observed in the sample containing the cryptic splice site and not observed in any of the 149 GTEx samples or in the other 35 sequenced samples. For 4 additional

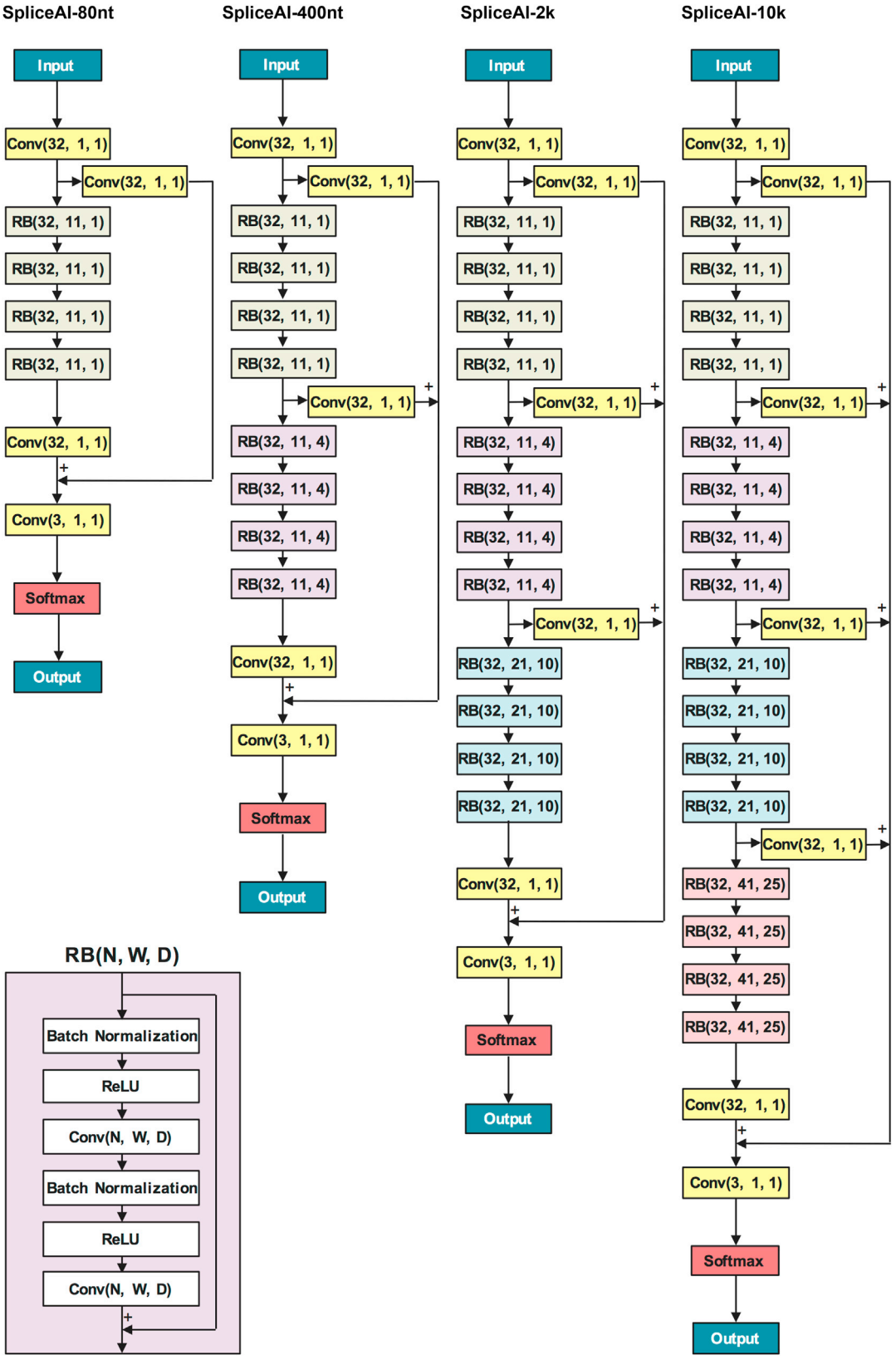
exon skipping events, low levels of exon-skipping were often observed in GTEx. In these cases, we computed the fraction of reads that used the skipping junction and verified that this fraction was highest in the cryptic splice site containing sample compared to other samples. Four additional cases were validated on the basis of prominent intron retention that was absent or much lower in other samples. Modest intron retention in control samples prevented us from resolving events in *DDX11* and *WDR4*. Two events (in *CSAD*, and *GSAP*) were classified as failing validation because the variant was not present in sequencing reads.

DATA AND SOFTWARE AVAILABILITY

Training and testing data, prediction scores for all possible single nucleotide substitutions in the reference genome, RNA-seq validation results, RNA-seq junctions, and source code are publicly hosted at https://basespace.illumina.com/s/5u6ThOblecrh and https://github.com/Illumina/SpliceAI.

RNA-seq data for the 36 lymphoblastoid cell lines are being deposited in the ArrayExpress database at EMBL-EBI (https://www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-7351.

Prediction scores and source code are publicly released under GPL v3 and are free for use for academic and non-commercial applications.



(legend on next page)

Figure S1. Detailed Description of the SpliceAI-80nt, SpliceAI-400nt, SpliceAI-2k, and SpliceAI-10k Architectures, Related to Figure 1

The four architectures use flanking nucleotide sequence of lengths 40, 200, 1,000, and 5,000 respectively on each side of the position of interest as input, and output the probability of the position being a splice acceptor, splice donor, and neither. The architectures mainly consist of convolutional layers Conv(N, W, D), where N, W, and D are the number of convolutional kernels, window size, and dilation rate of each convolutional kernel in the layer respectively.

A

	Top-*k* accuracy	PR-AUC
SpliceAI-80nt	0.51	0.52
SpliceAI-400nt	0.70	0.77
SpliceAI-2k	0.82	0.89
SpliceAI-10k	**0.84**	**0.90**
GeneSplicer	0.33	0.26
MaxEntScan	0.33	0.26
NNSplice	0.32	0.24

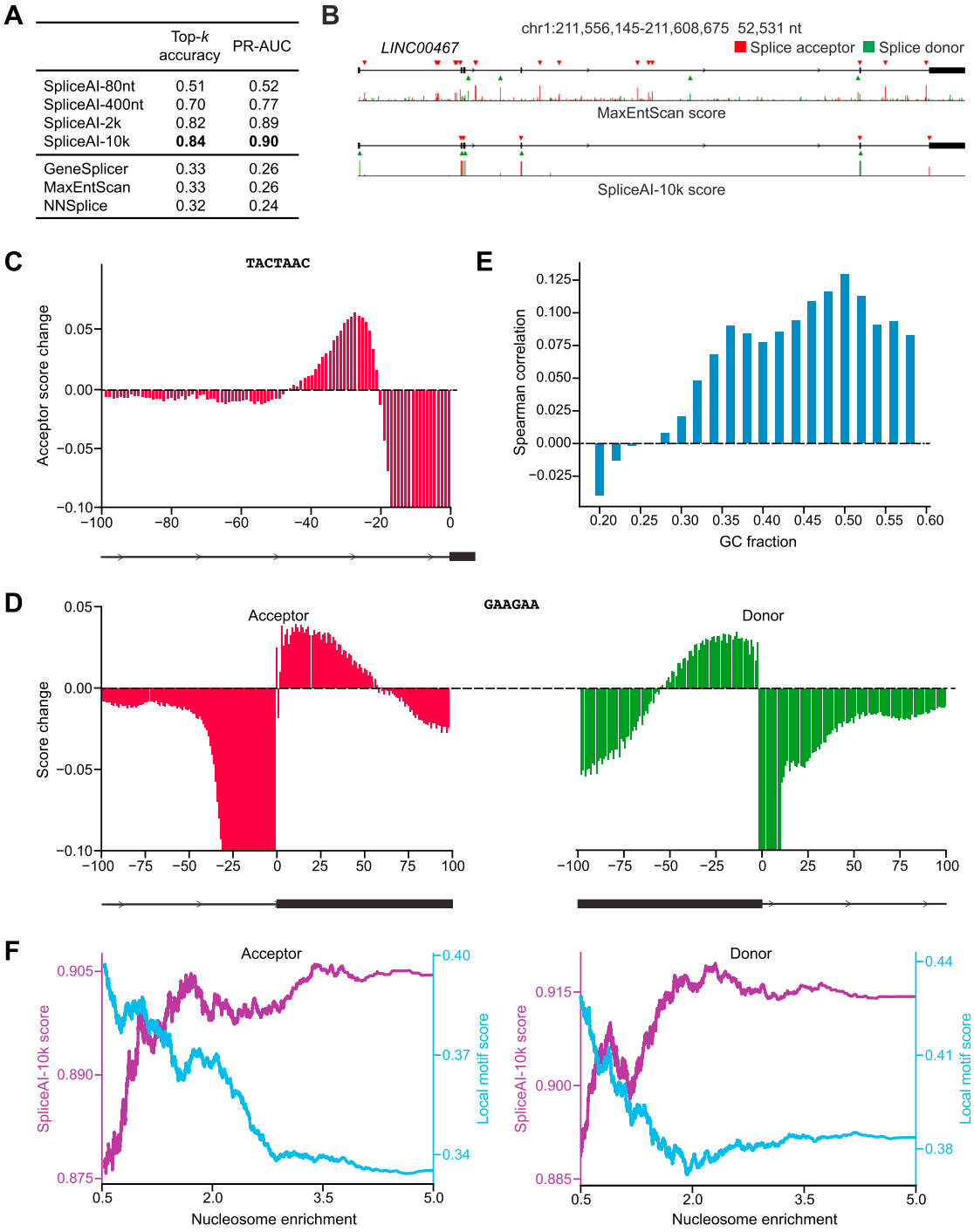


Figure S2. Evaluation of Various Splicing Prediction Algorithms on lincRNAs and the Effects of Various Factors on Splicing, Related to Figure 1

(A) The top-*k* accuracies and the area under the precision-recall curves of various splicing prediction algorithms when evaluated on lincRNAs are shown.

(B) The full pre-mRNA transcript for the *LINC00467* gene scored using MaxEntScan and SpliceAI-10k is shown, along with predicted acceptor (red arrows) and donor (green arrows) sites and the actual positions of the exons.

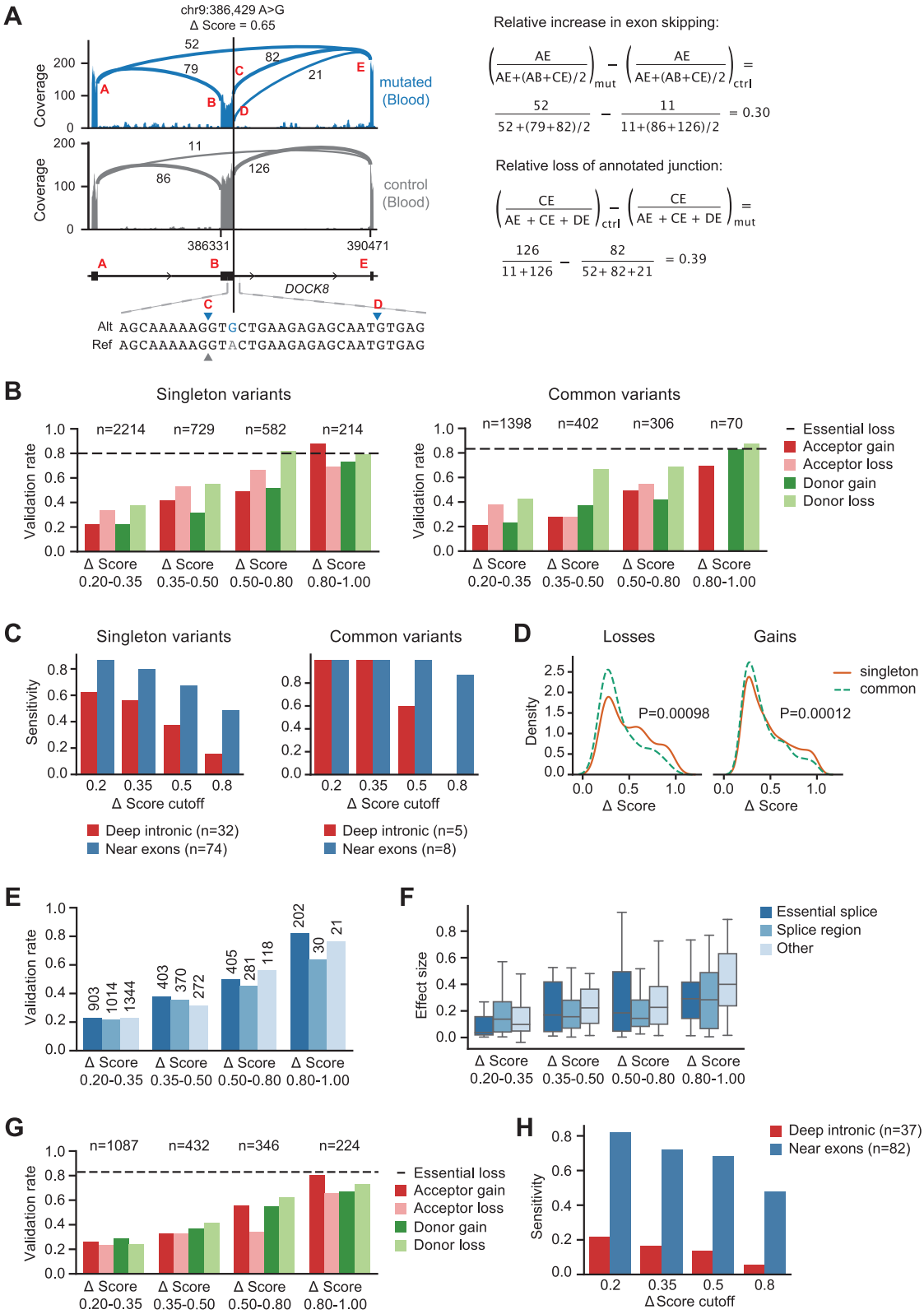
(C) The optimal branch point sequence TACTAAC was introduced at various distances from each of the 14,289 test set splice acceptors and the acceptor scores were calculated using SpliceAI-10k. The average change in the predicted acceptor score is plotted as a function of the distance from the splice acceptor. The predicted scores increase when the distance from the splice acceptor is in between 20 and 45 nt; at less than 20 nt distance, TACTAAC disrupts the poly-pyrimidine tract due to which the predicted acceptor scores are very low.

(legend continued on next page)

(D) The SR-protein hexamer motif GAAGAA was similarly introduced at various distances from each of the 14,289 test set splice acceptors and donors. The average change in the predicted SpliceAI-10k acceptor and donor scores are plotted as a function of the distance from the splice acceptor and donor respectively. The predicted scores increase when the motif is on the exonic side and less than ~50 nt from the splice site. At larger distances into the exon, the GAAGAA motif tends to disfavor the usage of the splice acceptor or donor site under consideration, presumably because it now preferentially supports a more proximal acceptor or donor motif. The very low acceptor and donor scores when GAAGAA is placed at positions very close to the intron is due to disruption of the extended acceptor or donor splice motifs.

(E) At 1 million randomly chosen intergenic positions, strong acceptor and donor motifs spaced 150 nt apart were introduced and the probability of exon inclusion was calculated using SpliceAI-10k. To show that the correlation between SpliceAI-10k predictions and nucleosome positioning occurs independent of GC composition, the positions were binned based on their GC content (calculated using the 150 nucleotides between the introduced splice sites) and the Spearman correlation between SpliceAI-10k predictions and nucleosome signal is plotted for each bin.

(F) Splice acceptor and donor sites from the test set were scored using SpliceAI-80nt (referred to as local motif score) and SpliceAI-10k, and the scores are plotted as a function of nucleosome enrichment. Nucleosome enrichment is calculated as the nucleosome signal averaged across 50 nt on the exonic side of the splice site divided by the nucleosome signal averaged across 50 nt on the intronic side of the splice site. SpliceAI-80nt score, which is a surrogate for motif strength, is negatively correlated with nucleosome enrichment, whereas SpliceAI-10k score is positively correlated with nucleosome enrichment. This suggests that nucleosome positioning is a long-range specificity determinant that can compensate for weak local splice motifs.



(legend on next page)

Figure S3. Illustration of Effect Size Computation and SpliceAI-10k Performance Benchmarks, Related to Figure 2

(A) The intronic variant chr9:386429 A > G disrupts the normal donor site (C) and activates a previously suppressed intronic downstream donor (D). Shown are the RNA-seq coverage and junction read counts in whole blood from the individual with the variant and a control individual. The donor sites in the individual with the variant and the control individual are marked with blue and gray arrows respectively. Bold red letters correspond to junction endpoints. For visibility, exon lengths have been exaggerated 4-fold compared to intron lengths. To estimate the effect size, we compute both the increase in the usage of the exon skipping junction (AE) and the decrease in the usage of the disrupted junction (CE) relative to all other junctions with the same donor E. The final effect size is the maximum of the two values (0.39). An increased amount of intron retention is also present in the mutated sample. These variable effects are common at exon skipping events and increase the complexity of validating rare variants that are predicted to cause acceptor or donor site loss.

(B) Fraction of cryptic splice mutations predicted by SpliceAI-10k that validated against GTEx RNA-seq data. The model was evaluated on all variants appearing in at most four GTEx individuals. Variants with predicted splice-altering effects were validated against RNA-seq data. The validation rate is shown separately for variants appearing in a single individual (left) and variants appearing in two to four individuals (right). Predictions are grouped by their Δ Score. We compared the validation rate between singleton and common variants for each of the four classes of variants (gain or loss of acceptor or donor) in each Δ Score group. The differences are not significant ($p > 0.05$, Fisher Exact test with Bonferroni correction for 16 tests).

(C) SpliceAI-10k sensitivity at detecting splice-altering variants in the GTEx cohort at different Δ Score cutoffs. The model's sensitivity is shown separately for singleton (left) and common (right) variants. The differences in sensitivity between singleton and common variants at a Δ Score cutoff of 0.2 are not significant for either variants near exons or deep intronic variants ($p > 0.05$, Fisher Exact test with Bonferroni correction for two tests).

(D) Distribution of Δ Score values for validated singleton and common variants. p values are for Mann-Whitney U tests comparing the scores of singleton and common variants. Common variants have significantly weaker Δ Score values, due to natural selection filtering out splice-disrupting mutations with large effects.

(E) Predicted splice site-creating variants were grouped based on whether the variant created a new essential GT or AG splice dinucleotide, whether it overlapped the rest of the splice motif (all positions around the exon-intron boundary up to 3 nt into the exon and 8 nt into the intron, excluding the essential dinucleotide), or whether it was outside the splice motif. Shown is the validation rate for each of the three categories of splice site-creating variants. The total number of variants in each category is shown above the bars. Within each Δ Score group, the differences in validation rates between the three groups of variants are not significant ($p > 0.3$, χ^2 test of uniformity).

(F) Distribution of effect sizes for each of the three categories of splice site-creating variants. Within each Δ Score group, the differences in effect sizes between the three groups of variants are not significant ($p > 0.3$, Mann-Whitney U test with Bonferroni correction).

(G) We trained SpliceAI-10k using only junctions from canonical GENCODE transcripts and compared the performance of this model to a model trained on both canonical junctions and splice junctions appearing in at least five individuals in the GTEx cohort (Figure 2). We compared the validation rates of the two models for each of the four classes of variants (gain or loss of acceptor or donor) in each Δ Score group. The differences in validation rates between the two models are not significant ($p > 0.05$, Fisher Exact test with Bonferroni correction for 16 tests).

(H) Sensitivity of the model that was trained on canonical junctions at detecting splice-altering variants in the GTEx cohort at different Δ Score cutoffs. The sensitivity of this model in deep intronic regions is lower than that of the model on Figure 2 ($p < 0.001$, Fisher Exact test with Bonferroni correction). The sensitivity near exons is not significantly different.

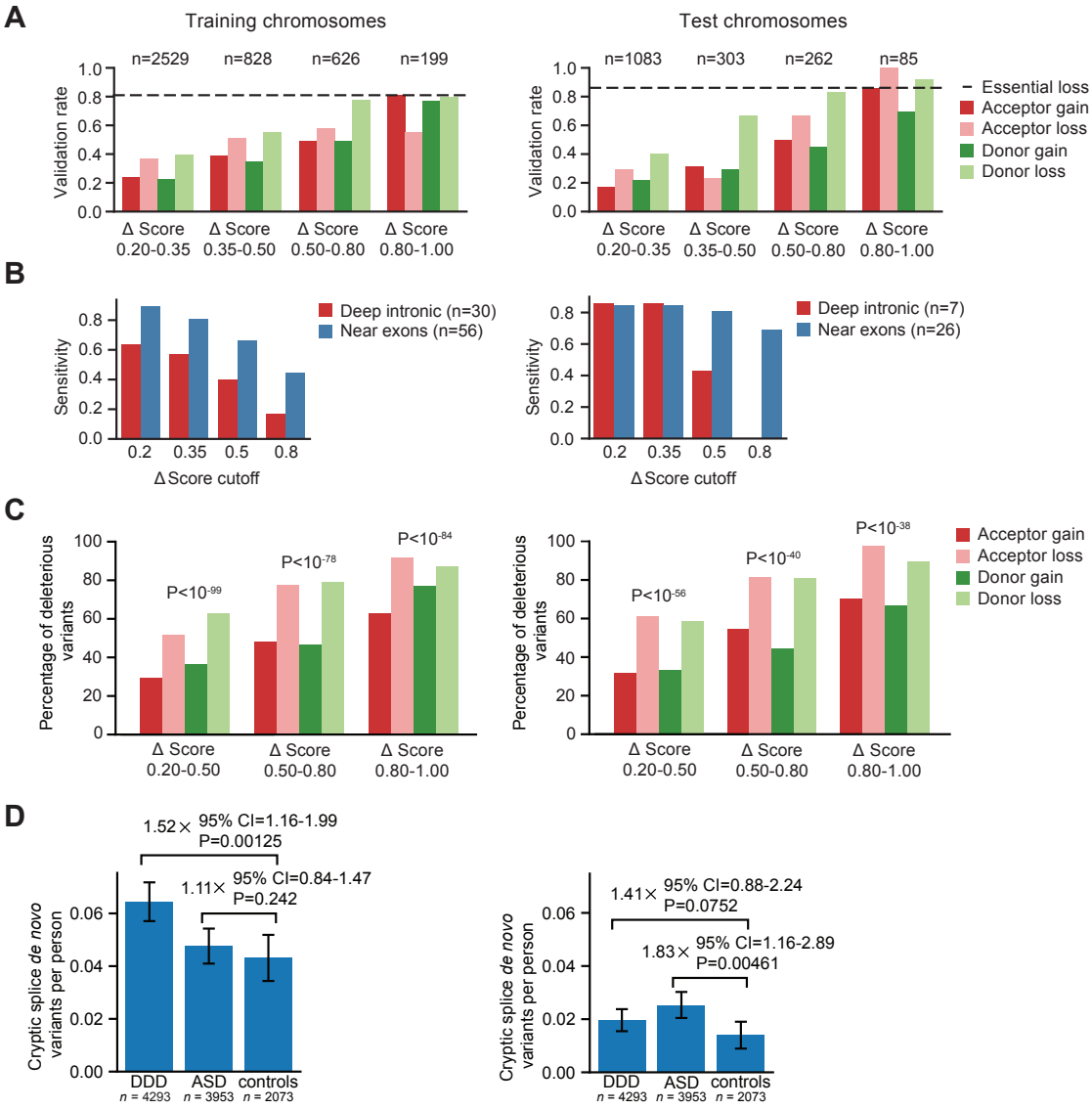


Figure S4. Evaluation of SpliceAI-10k on Train and Test Chromosomes, Related to Figures 2, 4, and 5

(A) Fraction of cryptic splice mutations predicted by SpliceAI-10k that validated against GTEx RNA-seq data. The validation rate is shown separately for variants on chromosomes used during training (all chromosomes except chr1, chr3, chr5, chr7, and chr9; left) and the rest of the chromosomes (right). Predictions are grouped by their Δ Score. We compared the validation rate between train and test chromosomes for each of the four classes of variants (gain or loss of acceptor or donor) in each Δ Score group. This accounts for potential differences in the distribution of predicted Δ Score values between train and test chromosomes. The differences in validation rates are not significant ($p > 0.05$, Fisher Exact test with Bonferroni correction for 16 tests).

(B) Sensitivity of SpliceAI-10k at detecting splice-altering variants in the GTEx cohort at different Δ Score cutoffs. The model's sensitivity is shown separately for variants on the chromosomes used for training (left) and on the rest of the chromosomes (right). We used a Fisher Exact test to compare the model's sensitivity at a Δ Score cutoff of 0.2 between train and test chromosomes. The differences are not significant for either variants near exons or deep intronic variants ($p > 0.05$ after Bonferroni correction for two tests).

(C) Fraction of predicted synonymous and intronic cryptic splice variants in the ExAC dataset that are deleterious, calculated separately for variants on chromosomes used for training (left) and the rest of the chromosomes (right). Fractions and p values are computed as shown in Figure 4A. We compared the number of common and rare variants between training and test chromosomes for each of the four classes of variants (gain or loss of acceptor or donor) in each Δ Score group. The differences are not significant ($p > 0.05$, Fisher Exact test with Bonferroni correction for 12 tests).

(D) Predicted cryptic splice *de novo* mutations (DNMs) per person for DDD, ASD, and control cohorts, shown separately for variants on the chromosomes used for training (left) and the rest of the chromosomes (right). Error bars show 95% confidence intervals (CI). The number of cryptic splice *de novo* variants per person is smaller for the test set because it is roughly half the size of the training set. Numbers are noisy due to small sample size.

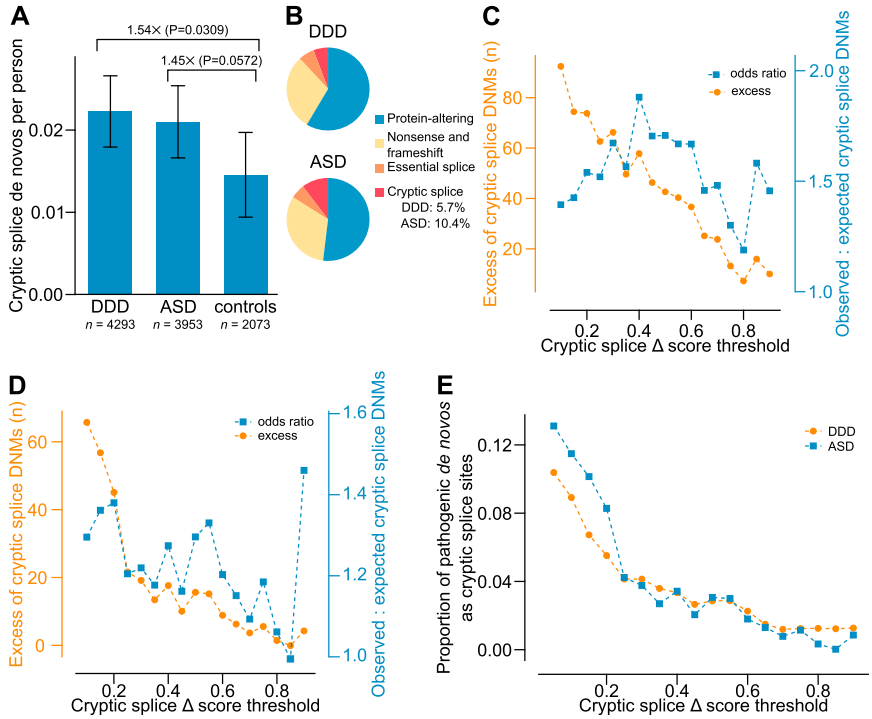


Figure S5. De Novo Cryptic Splice Mutations in Patients with Rare Genetic Disease, Related to Figure 5

(A) Predicted cryptic splice *de novo* mutations (DNMs), only from synonymous, intronic, or untranslated region sites with cryptic splice Δ score > 0.2 per person for patients from the Deciphering Developmental Disorders cohort (DDD), individuals with autism spectrum disorders (ASD) from the Simons Simplex Collection and the Autism Sequencing Consortium, as well as healthy controls. Enrichment in the DDD and ASD cohorts above healthy controls is shown, adjusting for variant ascertainment between cohorts. Error bars show 95% confidence intervals.

(B) Estimated proportion of pathogenic DNMs by functional category for the DDD and ASD cohorts, based on the enrichment of each category compared to healthy controls. The cryptic splice proportion (only from synonymous, intronic, or untranslated region sites) is adjusted for the lack of missense and deeper intronic sites.

(C) Enrichment and excess of cryptic splice DNMs (only from synonymous, intronic, or untranslated region sites) in the DDD and ASD cohorts compared to healthy controls at different Δ score thresholds. The cryptic splice excess is adjusted for the lack of missense and deeper intronic sites.

(D) Enrichment and excess of cryptic splice DNMs within ASD probands at different Δ score thresholds for predicting cryptic splice sites.

(E) Proportion of pathogenic DNMs attributable to cryptic splice sites as a fraction of all classes of pathogenic DNMs (including protein-coding mutations), using different Δ score thresholds for predicting cryptic splice sites. More permissive Δ score thresholds increase the number of cryptic splice sites identified over background expectation, at the trade-off of having a lower odds ratio.

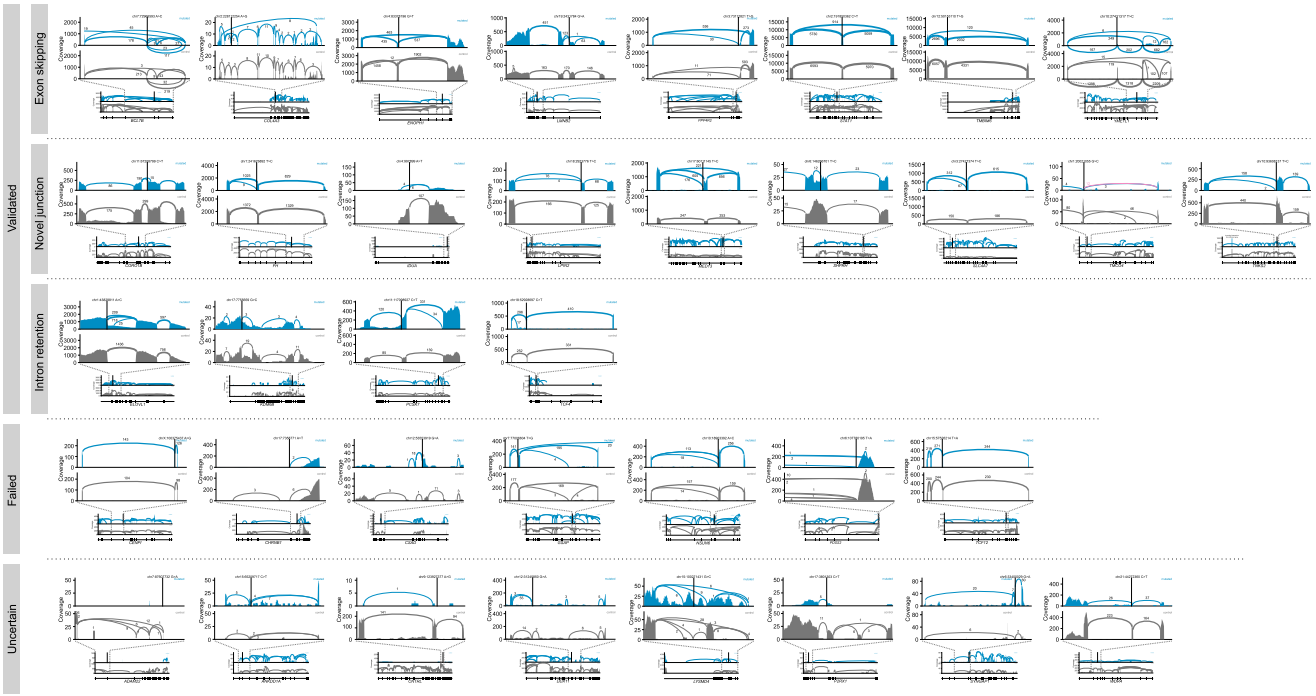


Figure S6. RNA-Seq Validation of Predicted Cryptic Splice *De Novo* Mutations in ASD Patients, Related to Figure 5

Coverage and splice junction counts of RNA expression from 36 predicted cryptic splice sites selected for experimental validation by RNA-seq. For each sample, RNA-seq coverage and junction counts for the affected individual are shown at the top, and a control individual without the mutation is shown at the bottom. The plots are grouped by validation status and splice aberration type.