nature structural & molecular biology

Article

https://doi.org/10.1038/s41594-022-00844-1

Mechanism and modeling of human disease-associated near-exon intronic variants that perturb RNA splicing

Received: 6 September 2021

Accepted: 23 August 2022

Published online: 27 October 2022

Check for updates

Hung-Lun Chiang[1,5], Yi-Ting Chen[1,5], Jia-Ying Su[1,2,3,4,5], Hsin-Nan Lin[1], Chen-Hsin Albert Yu[1], Yu-Jen Hung[1], Yun-Lin Wang[1], Yen-Tsung Huang[2] and Chien-Ling Lin[1]

It is estimated that 10%–30% of disease-associated genetic variants affect splicing. Splicing variants may generate deleteriously altered gene product and are potential therapeutic targets. However, systematic diagnosis or prediction of splicing variants is yet to be established, especially for the near-exon intronic splice region. The major challenge lies in the redundant and ill-defined branch sites and other splicing motifs therein. Here, we carried out unbiased massively parallel splicing assays on 5,307 disease-associated variants that overlapped with branch sites and collected 5,884 variants across the 5′ splice region. We found that strong splice sites and exonic features preserve splicing from intronic sequence variation. Whereas the splice-altering mechanism of the 3′ intronic variants is complex, that of the 5′ is mainly splice-site destruction. Statistical learning combined with these molecular features allows precise prediction of altered splicing from an intronic variant. This statistical model provides the identity and ranking of biological features that determine splicing, which serves as transferable knowledge and out-performs the benchmarking predictive tool. Moreover, we demonstrated that intronic splicing variants may associate with disease risks in the human population. Our study elucidates the mechanism of splicing response of intronic variants, which classify disease-associated splicing variants for the promise of precision medicine.

RNA splicing is a fundamental process to ligate exons for translation and excise introns for nucleic acid recycling. Alternative RNA splicing greatly expands the coding capacity of a genome, and its temporal and spatial regulation contribute to the transcriptomic complexity of an organism[1]. With the aid of splicing regulatory elements and structural composition, three essential elements within introns or near intron–exon boundaries act as splicing signals: the 5′ splice site (5′ss), the branch site (BS) and the 3′ splice site (3′ss)[2]. A polypyrimidine tract downstream of the BS and an AG dinucleotide exclusion zone facilitate 3′ss recognition[3]. Together, the BS, polypyrimidine tract and 3′ss stabilize U2 small nuclear ribonucleoproteins (snRNPs), U2 auxiliary factor 2 and U2 auxiliary factor 1 for 3′ss recognition[2]. Mechanistically, the 5′ss base pairs with U1 small nuclear RNA (snRNA), later replaced by U6 snRNA over spliceosome rearrangement, as does the BS with the BS recognition sequence of U2 snRNA. Partial base pairing with the U2 snRNA causes the branchpoint to bulge out, and the interaction between the

[1]Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan. [2]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan. [3]Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan. [4]Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan. [5]These authors contributed equally: Hung-Lun Chiang, Yi-Ting Chen, Jia-Ying Su. e-mail: mbcllin@gate.sinica.edu.tw

U6 and U2 snRNPs positions the branchpoint for nucleophilic attack on the 5′ss, representing the first catalytic event of splicing. Large-scale BS mapping studies have revealed that BS are frequently located between nucleotides −40 to −18 upstream of the 3′ss, and multiple BS have been detected within a given intron[4–7]. By contrast to the well-defined 5′ss and 3′ss, the BS appears to lack a definitive sequence motif or position in higher organisms. The ambiguity of the BS hinders interpretations of intronic sequence variation near intron–exon boundaries.

It has been estimated that 10%–30% of the pathogenic mutations in patients suffering rare genetic disorders alter splicing[8,9]. The resulting misspliced transcripts may generate aberrant proteins or trigger nonsense-mediated decay pathways that eliminate the gene product[10]. Auxiliary elements to the splice sites act as notable secondary sites within introns that can cause splicing defects[11]. However, detailed RNA sequencing data to support splicing phenotypes of disease-associated intronic variants are lacking, considering their low frequency and the limited likelihood of deriving their splicing outcomes from existing data. The desire to interpret splicing variants has prompted the development of high-throughput splicing assays and respective modeling. Massively parallel splicing reporter assays (MaPSy) have been used to test the effect of variable sequences on splice-site choice. Variable sequences near the 5′ss and 3′ss, or encompassing a full intron/exon, have been tested in the context of fixed backbones to examine their effect on splicing. The results demonstrated that the first 5′ss adjacent to the exon is preferred, given two 5′ss of the same strength, whereas the 3′ss choice is more ambiguous, suggesting that 3′ss selection involves more sophisticated regulation. MaPSy with random sequences at the BS showed degenerate BS recognition and confined dependence on the U2 core proteins[12]. Moreover, MaPSy with a split-GFP (green fluorescent protein) construct design was applied to examine aberrant exon skipping induced by genetic variations using fluorescence-activated cell sorting[13]. The results showed that a total of 54% of splicing-disrupting variants were intronic (including splice sites), demonstrating a considerable contribution of intronic splicing regulation. Taken together, these results suggest that intronic signals contribute significantly to splicing regulation, as does variation to the splicing defect, which prompted us to develop MaPSy directly on the disease-associated intronic variants.

Recently, deep learning has been implemented in some studies to decipher the contribution of primary sequences to splice-site selection[8,14,15]. These studies have revealed splicing variant enrichment at the splice sites and a sparse distribution in the exon that extends to the 3′-end of introns, consistent with the notion that 5′ss choice is made strictly by the splice-site consensus, whereas 3′ss choice engages more intronic regulatory elements, namely the BS and polypyrimidine tracts. However, the respective models were trained on differentiating constitutive splice sites versus alternative or mock splice sites, and not on intronic variants. Furthermore, the derived tools performed only moderately, mostly predicting splice site and exonic splicing variants[13,16–19]. Moreover, deep learning cannot establish the significance of each input factor, limiting the scope for model advancement.

Functional assignment of genetic variation has never been more urgent given the rapid growth in screening for the genetic basis of disease and precision medicine. The ultimate goal of modeling splicing variants is to establish the functional impact of variants and to recognize the contributions of splicing regulatory factors. MaPSy of exonic disease-relevant variants indicated that ~10% of them affected splicing, with splice-site strength and exonic splicing regulatory elements contributing to predictions of the variants' effects[20]. In the absence of reliable detection of disease-relevant intronic variants, efforts have been made to computationally classify the impact of the respective mutations with respect to categories of pathogenicity derived from databases, but imperfect categorization limits the predictive power for splicing variants[21,22]. To systematically study the effect of near-exon intronic variants on splicing, we performed unbiased MaPSy on disease-relevant and/or rare near-exon intronic variants, assessed the features determining the splicing defect and integrated our data into predictive models. Direct sequencing of the spliced product enabled the identification of splicing errors beyond exon skipping. We further extended the models into an interactive web-based tool, Splice-Alternative Profile Predictor (SpliceAPP; https://bc.imb.sinica.edu.tw/SpliceAPP/), for splicing variant prediction. Furthermore, we show that predicted splicing mutations are correlated with disease phenotype and a skewed biochemical index in the Taiwan Biobank (TWB). Importantly, we provide models explaining and predicting near-exon intronic mutations and demonstrate their functional deficiency, linking the noncoding variants to their potential defect in protein production.

Results

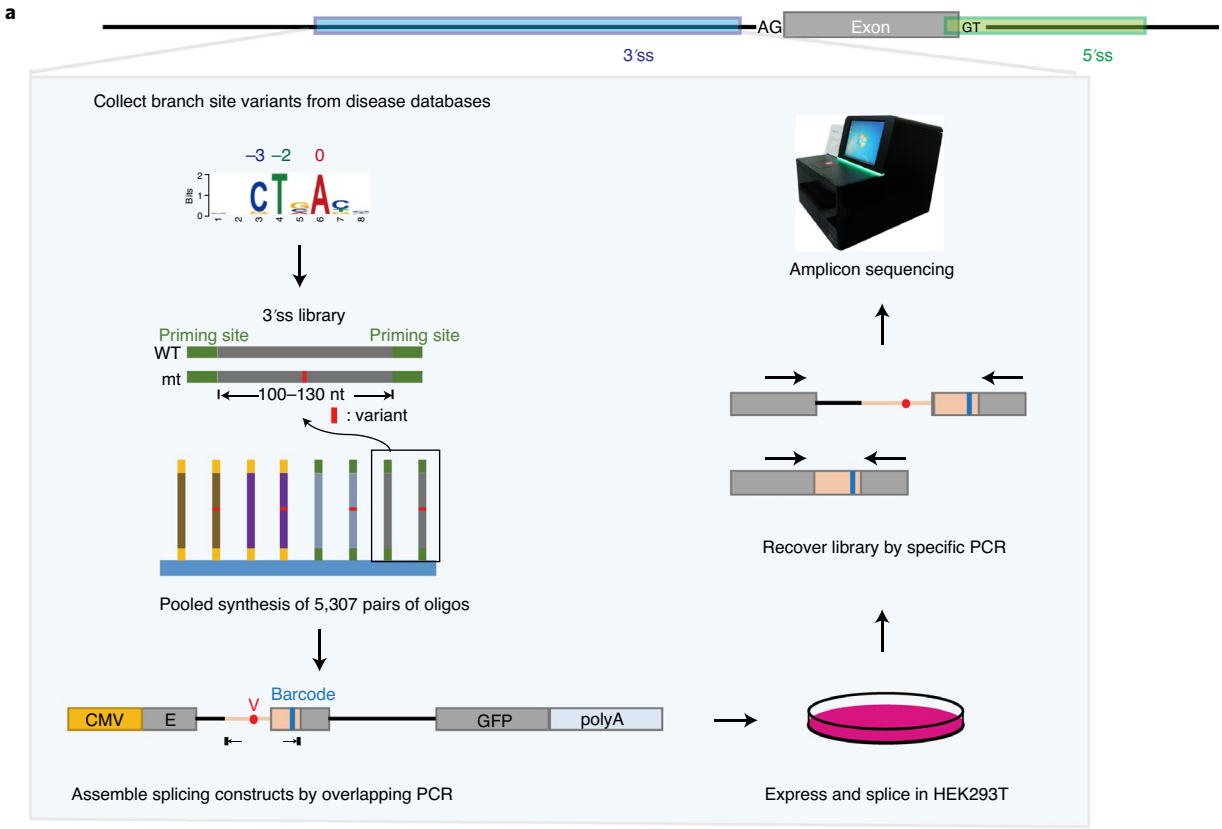
MaPSy of disease-relevant human BS and 5′ss variants

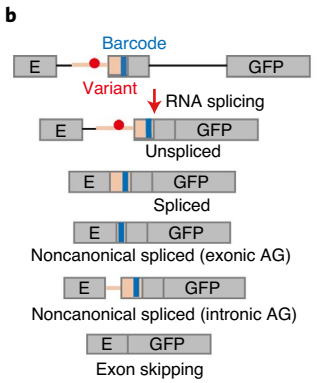
Splice-site recognition requires base-pairing of 5′ss with U1 snRNA and BS with U2 snRNA. Regardless of their functional significance, human BS motifs are degenerate and it has been reported previously that almost all human introns contain multiple BS[6]. It is unclear whether each BS is indispensable for splicing and how genetic variation in the BS contributes to splicing defects in disease contexts. In addition, the contextual influence of the 5′ss and BS choice remains to be elucidated. Therefore, we collected MaPSy on disease-relevant and/or rare variants near BS and 5′ss to systematically test their effect on splicing. A total of 5,307 BS variants are reported in the Human Gene Mutation Database[23], the public archive of interpretations of clinically relevant variants (ClinVar)[24], the database of single nucleotide polymorphisms[25] or the Catalogue of Somatic Mutations in Cancer[26], all of which provide overlapping data on the most informative BS position 0 (the branchpoint), as well as the −2 or −3 positions of experimentally discovered branchpoints[4–6] (Fig. 1a). We synthesized paired oligos containing wild-type (WT) or mutant (mt) BS and their affiliated exons in bulk, ligated them into three-exon splicing minigenes, and then transfected

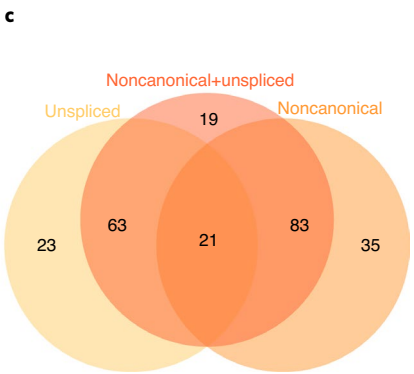
Fig. 1 | MaPSy of near-exon intronic mutations. a, Experimental design of MaPSy[13]. BS mutations (0, −2, −3 to branchpoint position) documented in databases of human disease were collected and synthesized as 5,307 pairs of oligos. Each oligo pair contains a WT and a BS mt variant across the 78-nucleotide (nt) intronic and 35-nucleotide exonic regions. The oligos are flanked by common priming sites for amplification and ligation into 3-exon splicing minigenes. Accordingly, the synthesized region comprises the 3′ss of the second exon of the minigene. After minigene assembly, pooled minigenes were spliced in HEK293T cells. The resulting spliced isoforms were harvested and resolved by amplicon sequencing. **b**, Possible spliced outcomes for splicing minigenes. **c**, Numbers of HC significant splicing-altering mutations that exhibited a more than twofold change in four repeated experiments. Unspliced, significant change in unspliced reads; noncanonical, significant use of noncanonical splice sites; noncanonical + unspliced, significant change considering both unspliced and noncanonical splicing. **d**, Volcano plot of the MaPSy results. The *x* axis shows the WT/mt ratio for canonical splicing and the *y* axis shows the maximal *f* value for the four experiments on each WT/mt oligo pair. Non-sig. refers to variants that did not alter splicing in MaPSy. HC-sig. refers to high-confidence significant variants that altered the splicing. Addition or deletion of a potential 3′ss (add AG/reduce AG, respectively) at the mutation is labeled. **e**, Validation of MaPSy. Individual WT/mt oligo pairs were synthesized and transfected into HEK293T cells before visualizing the spliced product using electrophoresis (left). Of 39 pairs, the results for 38 matched amplicon sequencing data, as visualized using Integrative Genome Viewer (IGV, right, see also Extended Data Fig. 4). NGS, next generation sequencing. Scale bar, 100 µm. **f**, Correlation of semiquantitative RT–PCR and amplicon sequencing results in splicing efficiency. The gray area displays the 95% confidence interval for predictions from the linear model.

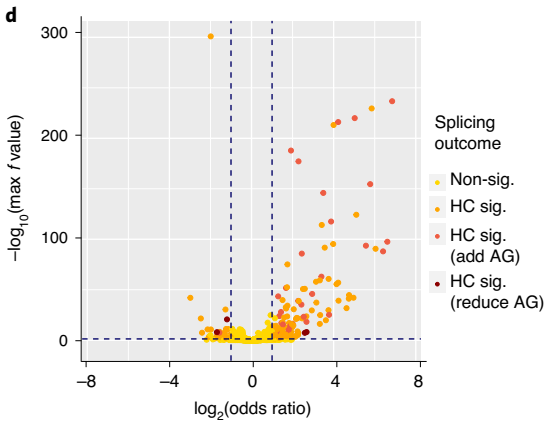
https://doi.org/10.1038/s41594-022-00844-1

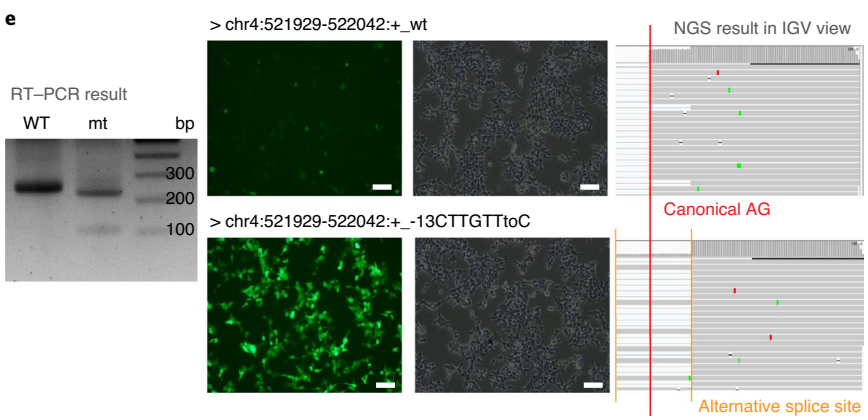
them into human embryonic kidney cells (HEK293T) to examine the splicing outcome. An exonic barcode was associated with the intronic variant for genotype identification (Fig. 1a,b). The spliced product was recovered by reverse transcription and library-specific amplification, and then resolved by amplicon sequencing (Fig. 1a and Extended Data Figs. 1 and 2). Biased splicing efficiency and use of noncanonical

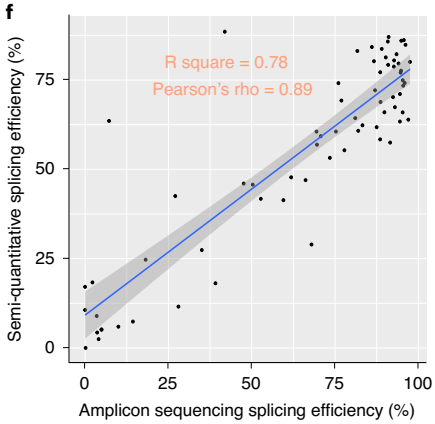












https://doi.org/10.1038/s41594-022-00844-1

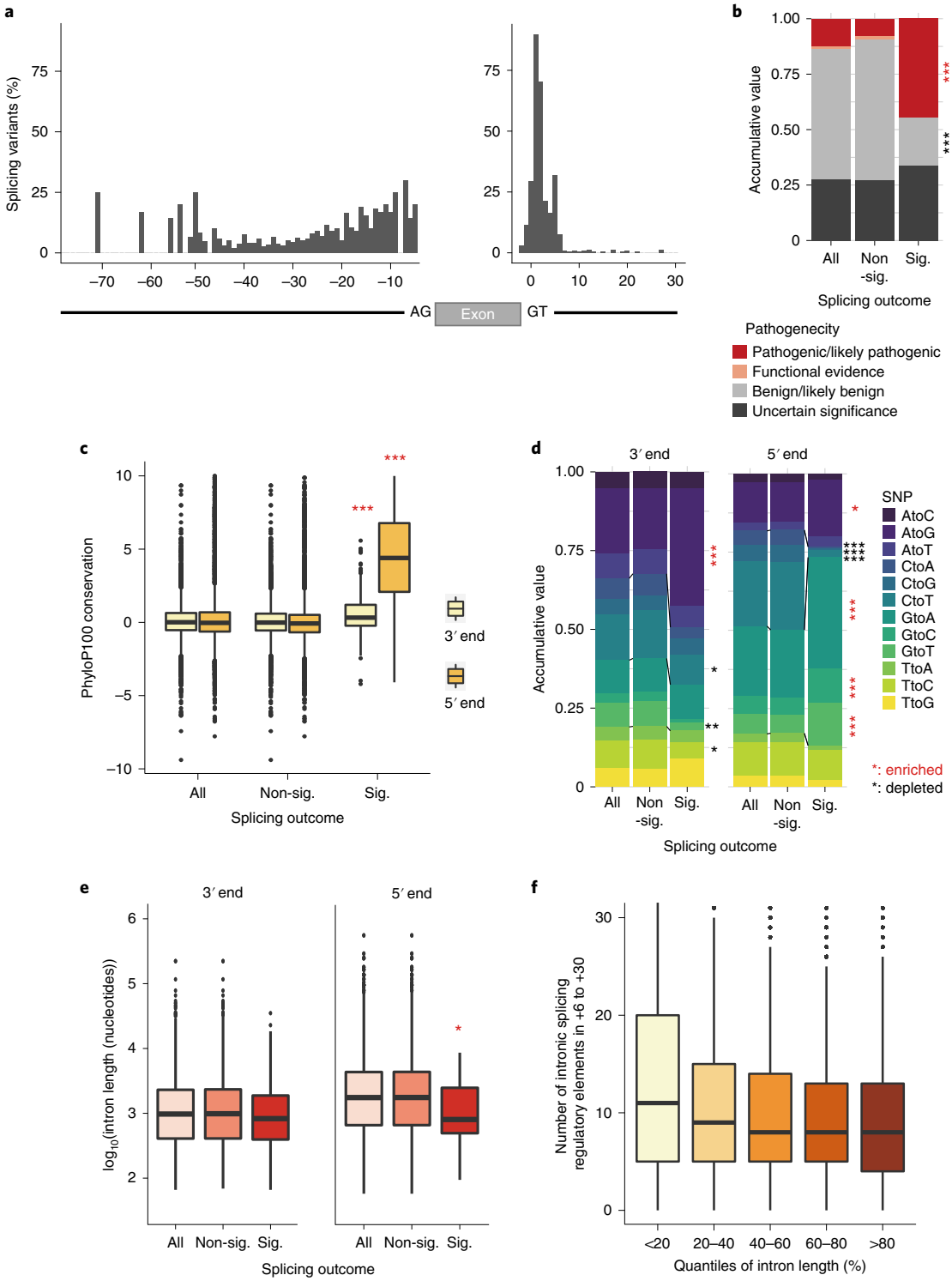


Fig. 2 | Significant splicing mutations demonstrate distinctive features. **a**, The 3′ splicing variants are enriched toward the 3′ss, whereas the 5′ splicing variants are specifically enriched at the 5′ss. **b**, Compared with the total library, significant splicing mutations are enriched for pathogenic/likely pathogenic mutations and are depleted in benign/likely benign mutations, as categorized in the ClinVar and Human Gene Mutation databases. Statistical significance is determined by two-sided Fisher's exact test. ***$P = 1 \times 10^{-10}$ between nonsignificant and significant for pathogenic/likely pathogenic and $P = 3 \times 10^{-13}$ for benign/likely benign. **c**, Significant splicing mutations are evolutionarily more conserved. Statistical significance is determined by two-sided Wilcoxon's test. $n = 5{,}052$ 3′ss variants and 5,884 5′ss variants. ***$P = 1 \times 10^{-9}$ for the 3′-end and $P < 2 \times 10^{-16}$ for the 5′-end. **d**, Nucleotide changes for SNPs. Statistical significance is determined by two-sided Fisher's exact test. For the 3′-end: $P = 2 \times 10^{-9}$ (AtoG), $P = 0.01$ (CtoT), $P = 0.002$ (GtoT) and $P = 0.04$ (TtoC). For the 5′-end: $P = 0.007$ (AtoG), $P = 2 \times 10^{-4}$ (CtoA), $P = 2 \times 10^{-6}$ (CtoG), $P = 2 \times 10^{-21}$ (CtoT), $P = 2 \times 10^{-7}$ (GtoA), $P = 1 \times 10^{-4}$ (GtoC) and $P = 4 \times 10^{-6}$ (GtoT). **e**, 5′ splicing variants outside the splice region are associated with shorter introns. $n = 3{,}970$ 3′ss variants and 4,831 5′ss variants. *$P = 0.01$. **f**, Genome-wide association between the number of intronic splicing regulatory elements near the 5′ss and the intron length. $n = 704{,}953$ introns. Statistical significance is determined by two-sided Wilcoxon's test. All, total library; Non-sig., variants that do not affect splicing outcome; Sig., significant splicing mutations. The boxes in box plots represent median (central line) and interquartile range (25th to 75th percentile). Whiskers indicate ±1.5× the interquartile range from the box or the last data point within that and the dots show the outliers (**c**,**e**,**f**).

splice sites by variants relative to their WT counterparts was identified by Fisher's exact test. We observed that 11.0% (455 of 4,154 valid comparisons) of candidate variants showed a consistent pattern of altered splicing across four experimental replicates, among which 244 candidates (6.1%) showed more than twofold change in the use of noncanonical splice sites. Variants that significantly altered splicing are named 'significant splicing variants', among which those with over twofold change are named 'high-confidence (HC) significant variants' hereafter (Fig. 1c,d, Extended Data Fig. 3 and Supplementary Table 1). MaPSy of 5,884 near 5′ss variants (−3 to +30) were collected from the published dataset[13].

Consistent with the BS MaPSy, 38 of 39 candidate pairs (97%) showed a consistent change in splicing upon independent semiquantitative reverse transcription-polymerase chain reaction (RT–PCR) (Fig. 1e). Overall, the 70 independent splicing assays revealed a high correlation between splicing efficiency and amplicon sequencing results (Pearson's correlation = 0.89) (Fig. 1f and Extended Data Fig. 4). Thus, our MaPSy confidently identified disease-relevant BS variants that caused splicing defects.

Complex and interconnected features of splicing variants

To deduce features underlying the effect of variants on splicing, we investigated the characteristics of the variants that caused significant splicing defects. Unlike the distinct enrichment of splicing variants at the 5′ss, splicing variants at the 3′-end of introns spread 50 nucleotides into the intron with slight enrichment toward the 3′ss (Fig. 2a and Extended Data Fig. 5a), reflecting the degenerate features of BS and other splicing signals in the intronic 3′-end. The same trend has also been observed in other MaPSy experiments[13,27]. Moreover, compared with the overall MaPSy library or nonsignificant variants, significant splicing variants were enriched among pathogenic and likely pathogenic mutations (Fig. 2b and Extended Data Fig. 5b). It is worth noting that although benign/likely benign variants are depleted in the pool of significant splicing variants, some benign/likely benign variants cause splicing alterations, possibly due to a misclassification arising from their noncoding nature (Fig. 2b). In addition, the MaPSy demonstrated that the significant splicing variants were highly conserved (Fig. 2c), potentially because they reside in the imperative splicing signals. Strikingly, A and G transition mutations correlated best with significant splicing defects in which A-to-G mutations of the 3′-end could represent branchpoint mutations, whereas G-to-A mutations of the 5′-end could be 5′ss mutations (Fig. 2d). Furthermore, we found that significant splicing variants outside the splice site were associated with shorter introns (Fig. 2e), suggesting enrichment of splicing regulatory elements in the shorter introns. To test the abovementioned hypothesis, we examined the number of intronic splicing regulatory elements +6 to +30 to 5′ss in relation to intron length in the whole human genome. The results showed that shorter introns are associated with enriched regulatory elements near the 5′ss (Fig. 2f and Extended Data Fig. 5c). Altogether, we identified general features underlying splicing variants using unbiased MaPSy screening. These features are complex and interconnected, implying that the combined features of the variants and their context determine the functional impact of mutations.

BS dysfunction and 3′ss competition alters 3′ splicing

Because of the excessive complexity of the 3′ splicing signal, we looked further into the specific characteristics of 3′ intronic splicing variants. We found that variants at branchpoints (bp 0) were enriched among significant splicing variants, whereas variants mutated at three nucleotides upstream of the branchpoints (bp −3) were depleted, revealing the essentiality of the branchpoint nucleotide (Fig. 3a). Remarkably, a considerable proportion of significant splicing variants created novel 3′ss AG dinucleotides (denoted 'add AG' in Fig. 1d and Extended Data Fig. 5d), and the mutation effect—the change in splicing efficiency of the mutant allele—was greater for novel AG variants (Fig. 3b). We reasoned that the impact of these novel AG dinucleotides could be a consequence of both BS dysfunction and 3′ss competition, so we isolated novel AG mutations for further investigation. A critical feature of the significant splicing variants was a weak 3′ss score (Fig. 3c), especially for add-AG mutations. The scores of the novel yet functional 3′ss were similar to their respective canonical 3′ss and, moreover, were frequently predicted as having a strong upstream BS (Fig. 3d, significant novel AGs represented by red dots are segregated in the first quadrant). At the sequence level, we identified a YAG 3′ss (Y = pyrimidine) with a polypyrimidine tract resembling a canonical 3′ss among the functional novel 3′ss (Fig. 3e). In addition to the 3′ss property, the novel exons exhibited a greater difference in exon-to-intron GC content (mean △GC = 3.5%), thus resembling canonical exons (mean △GC = 7.3%), whereas the mean △GC of nonfunctional novel AG variants was only 0.11% (Fig. 3f and Extended Data Fig. 5e). A previous study reported that this substantial difference in GC content enhanced splicing and likely promoted exon definition for spliceosome recognition[28]. Similar to all variants (Fig. 2a and Extended Data Fig. 5a), the functional novel AG tended to be proximal to the 3′ss (Fig. 3g). These results support a complex mechanism involving both BS dysfunction and 3′ss competition for 3′ intronic variants and a contribution to exon definition in the splicing reaction.

Modeling determining factors of splicing defects

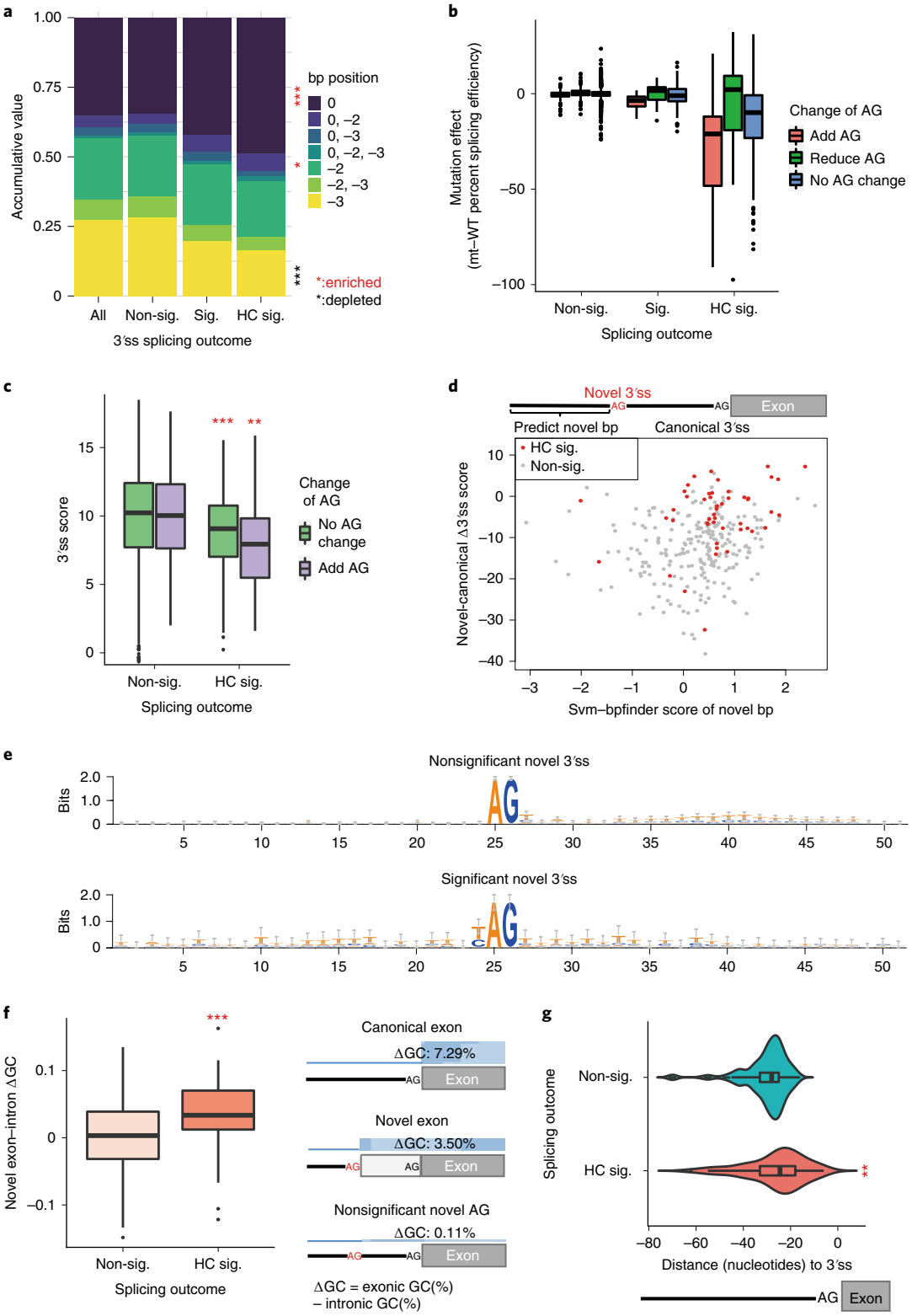
Despite identifying several traits specifically describing significant splicing variants (Figs. 2 and 3), none were absolute; even though the distribution of each trait diverged, there was overlap between nonsignificant and significant splicing variants. Therefore, we hypothesized that splicing defects could be explained by multiple additive features. To model their combined effects, we first collected potential explanatory features based on our current knowledge of the RNA splicing mechanism. These features included GC content, splice-site scores, BS and polypyrimidine tract score, pathogenicity level reported in the databases, alternative splicing of associated exons, splicing efficiency, evolutionary conservation score, mutation location, SNP changes, folding energy, openness of the splice region, predicted RNA-binding protein

Fig. 3 | Distinctive features of 3′ splicing variants. a, Significant splicing mutations display enrichment for variants at bp 0 but depletion for variants at bp −3. *P* values from two-sided Fisher's exact tests between the nonsignificant and HC significant variants are as follows: $P = 9 \times 10^{-6}$ (bp0), $P = 0.04$ (0, −2) and $P = 3 \times 10^{-5}$ (−3). **b**, Mutation effect (change in splicing efficiency of the mutant allele) of variants classified according to change in AG dinucleotide. $n = 301$, 282 and 3,558 for add AG, reduce AG and no AG change, respectively. **c**, The 3′ss score (weight matrix model), as determined by MaxEntScan[42], for the nonsignificant and HC significant variants. $n = 3{,}390$ and 281 for no AG change and add AG. *P* values from a two-sided Wilcoxon test between Non-sig. and HC sig. are 1×10^{-6} and 1×10^{-4} for no AG change and add AG, respectively. **d**, Significant add-AG novel 3′ss (red dots) have relatively strong novel 3′ss and predicted novel BS. The *x* axis represents novel BS score, as assessed by SVM-BPfinder[43], and the *y* axis is the difference between the novel 3′ss score and the canonical 3′ss score, as determined by MaxEntScan[42]. **e**, Information content of base composition around the novel 3′ss. Compared with the nonsignificant novel AGs (above), the significant novel AGs (below) are accompanied by a canonical 3′ss-like sequence structure, including a polypyrimidine tract and a YAG 3′ss (Y:T/C). **f**, Differential exon–intron GC content for nonsignificant and significant add-AG variants. The schematic (right) illustrates average differential exon–intron GC content for all exons, novel exons, unused novel AGs and their associated exons in the library. $n = 281$ add-AG variants. $^{***}P = 3 \times 10^{-5}$, two-sided Wilcoxon test. **g**, Distance between the mutation and the 3′ss of significant and nonsignificant add-AG variants. $n = 281$ add-AG variants. $^{**}P = 0.001$, two-sided Wilcoxon test. Boxes in box plots represent medians (central line) and interquartile ranges (25th to 75th percentile). Whiskers indicate ±1.5× the interquartile range from the box or the last data point within that and the dots show the outliers (**b**,**c**,**f**,**g**).

https://doi.org/10.1038/s41594-022-00844-1

(RBP) binding and predicted splicing regulatory elements. We then used least absolute shrinkage and selection operator (LASSO) regression for variable selection and regularization[29], and independently modeled 3′ (novel AG and non-AG) and 5′ near-exon intronic variants based on distinct sets of contributing features (Fig. 4a). We further normalized all features by their own standard deviation for standardized coefficients. The independent contribution and robustness (against data perturbations) of each feature was confirmed in Extended Data Fig. 6.

The 3′ novel AG model identified a high novel 3′ss score as the most important factor promoting the use of novel 3′ss. By contrast, a large distance between the variant and the 3′ss, as well as a high canonical 3′ss score, was most effective at suppressing the use of novel 3′ss (Fig. 4b). Other contributory factors included the level of conservation at the variant position (promoting novel 3′ss use), differences in novel exon-to-intron GC content (promoting) and the binding sites of RBPs (varied effects). It is worth noting that the binding sites of the potential



https://doi.org/10.1038/s41594-022-00844-1

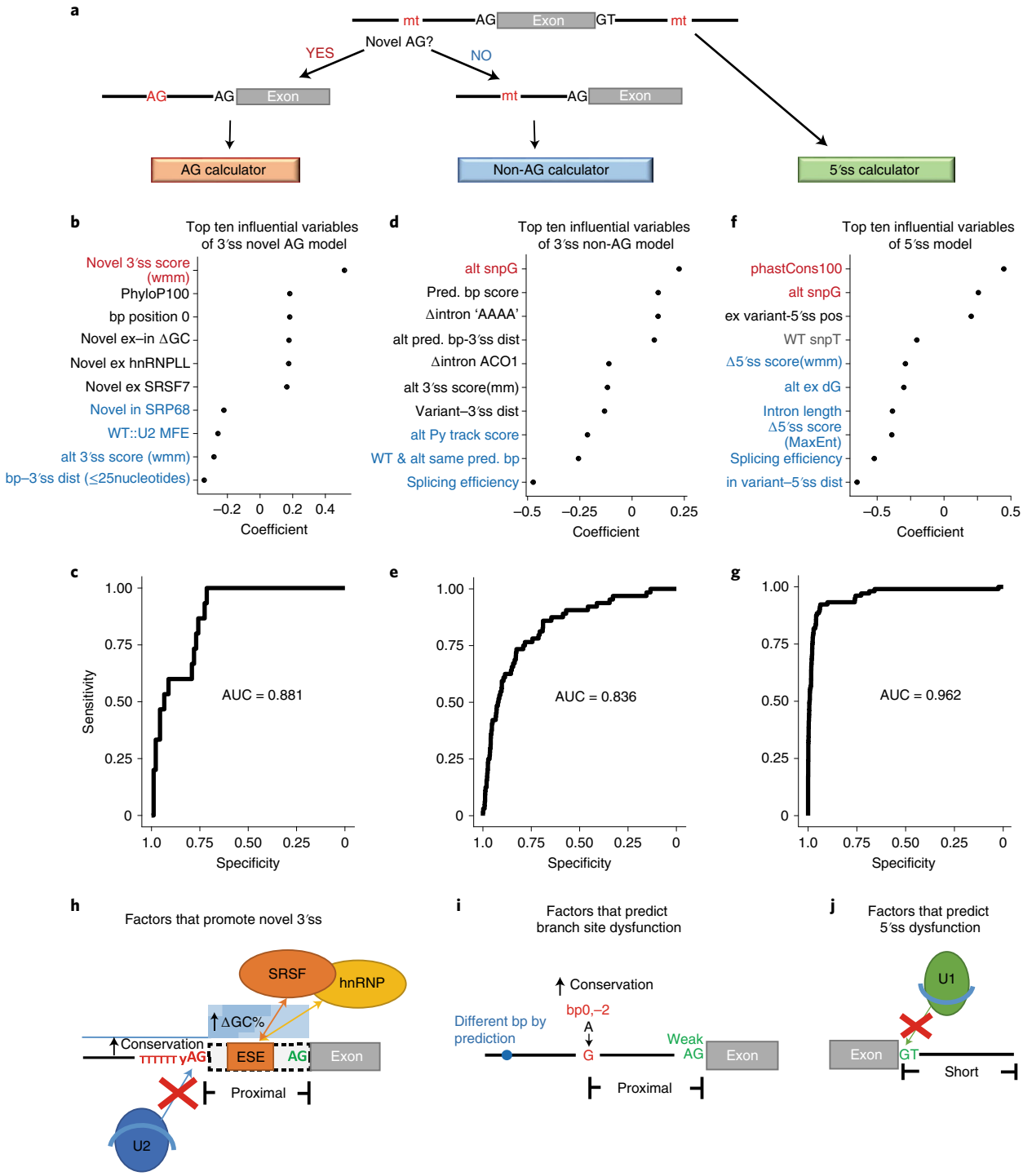
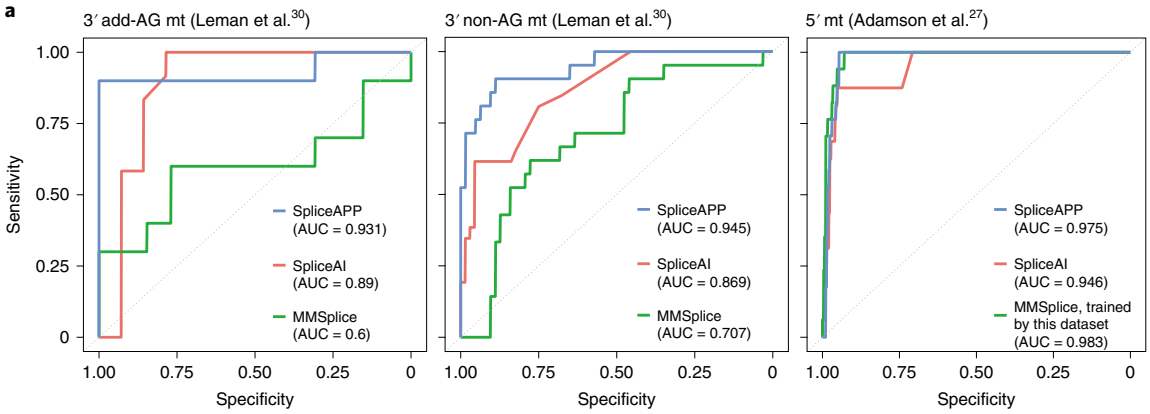


Fig. 4 | Generalized linear model to summarize predictors of splice-altering intronic variants. a, Segregation of intronic variants into two models based on intronic location and AG addition. **b**,**d**,**f**, Top ten contributory factors predicting variants in each category that affect splicing. Factors with a positive coefficient promote altered splicing, whereas those with a negative coefficient suppress it. **b**, Top ten contributory factors predicting 3' variants creating novel 3'ss. A 'novel' factor refers to a new property associated with the novel 3′ss AG. An 'alt' factor refers to a canonical property in the context of sequence variation. **d**, Top ten contributory factors predicting 3' non-AG variants that affect splicing. A delta '△' factor refers to the difference of scores/motifs between the alt and WT sequence. **f**, Top ten contributory factors predicting 5' variants that affect splicing. More detailed descriptions of the factors can be found in Supplementary Table 2. **c**,**e**,**g**, ROC curve for each model applied to one-third of the experimental data not included in the model training. **c**, ROC curve for 3' novel AG model. **e**, ROC curve for 3' non-AG model. **g**, ROC curve for 5' model. **h**–**j**, Models predicting splice-altering intronic mutations. **h**, Models for 3′ novel AG mutations. The strength of the canonical and novel 3′ss, the distance between the mutation and the canonical 3′ss, the pairing energy with the U2 BS recognition region and the exonic features of the novel exons all contribute to the likelihood of creating a novel 3′ss. **i**, Models for 3′ non-AG mutations. The strength of the canonical 3′ss, the distance between the mutation and its 3′ss, the position relative to a BS, SNP change, conservation level, and if a predicted BS is affected by the mutation contributes to the likelihood of causing splicing defects. **j**, Models for 5′-end mutations. Destruction of the U1 recognition splice-site region is the major splice-altering mechanism. Splicing variants outside the splice region associate with smaller intron length, suggesting an enrichment of splicing regulatory elements of shorter introns.

https://doi.org/10.1038/s41594-022-00844-1

a



b

Prediction and performance of the 3′ predictive models

	SIGNIFICANT*	NO EFFECT*	SENSITIVITY	SPECIFICITY	F-MEASURE
SPLICEAPP (107)	28/31	69/76	90.3%	90.8%	0.85
ADD AG (16)	6/7	9/9	85.7%	100%	0.92
NON-AG (91)	22/24	60/67	91.7%	89.6%	0.83
NON-BS (88)	23/26	56/62	88.5%	90.3%	0.84
IN VIVO EVIDENCE (66)	18/21	42/45	85.7%	93.3%	0.86
SPLICEAI (107)	3/31	75/76	9.7%	98.7%	0.17
MMSPLICE (107)	0/31	76/76	0%	100%	0

*Predicted/experiment (Leman et al..[30])

c

Prediction and performance of the 5′ predictive models

	SIGNIFICANT*	NO EFFECT*	SENSITIVITY	SPECIFICITY	F-MEASURE
SPLICEAPP (314)	17/17	275/297	100%	92.6%	0.61
EXCL 5′SS (304)	13/13	274/291	100%	94.2%	0.60
SPLICEAI (314)	12/17	281/297	70.1%	94.6%	0.53
MMSPLICE (314) trained by this dataset	14/17	287/297	81.4%	96.6%	0.68

*Predicted/experiment (Adamson et al.[27])

Fig. 5 | Validation of predictive models. a, ROC of each model applied to the external data. **b**, Prediction for the 3′-end mt with the recommended settings of each tool. ADD AG, mutations that create novel AG; NON AG, mutations that do not create novel AG; NON BS, mutations that are not at reported BS; *IN VIVO* EVIDENCE, experiments with in vivo evidence, not by minigenes. **c**, Prediction for the 5′-end mt with the recommended settings for each tool. EXCL 5′SS, mutations that are not at the 5′ss GT; SPLICEAI, results with the default setting of SpliceAI; MMSPLICE, results with the default cutoff |2| of delta_logit_psi. Numbers of tested mutations are given in parentheses.

splicing enhancers heterogeneous nuclear ribonucleoprotein L-like (hnRNPLL) and serine and arginine rich splicing factor 7 (SRSF7) in novel exons promoted novel 3′ss usage. We calculated an overall area under the curve (AUC) = 0.881 from the receiver operating characteristic (ROC) curve predicting one-third of experiments excluded from the training set (Fig. 4c). Because the outcome of our model indicated that the splicing response of novel AG variants arises from 3′ss competition, we examined splice-site choice when the novel 3′ss and canonical 3′ss were swapped. We found that the stronger 3′ss was chosen for splicing (Extended Data Fig. 5f), supporting the hypothesis that the use of novel intronic 3′ss results from 3′ss competition.

For non-AG variants, the model identified 'G' variants as the strongest factor promoting splicing defects. Strong WT splicing efficiency and having the same predicted branchpoint in WT and mutated (alt) introns suppressed splicing defects (Fig. 4d). Other factors contributing to splicing defects included the level of conservation of the variant position (promoting), polypyrimidine tract strength (suppressing) and the distance between the variant and the 3′ss (suppressing). We calculated AUC = 0.836 from the ROC for the test data (Fig. 4e).

The model for the 5′-end splicing variants is relatively simple in that the majority of splicing signals lie in the splice site. Therefore, destruction of 5′ss is the major factor contributing to splicing change.

https://doi.org/10.1038/s41594-022-00844-1

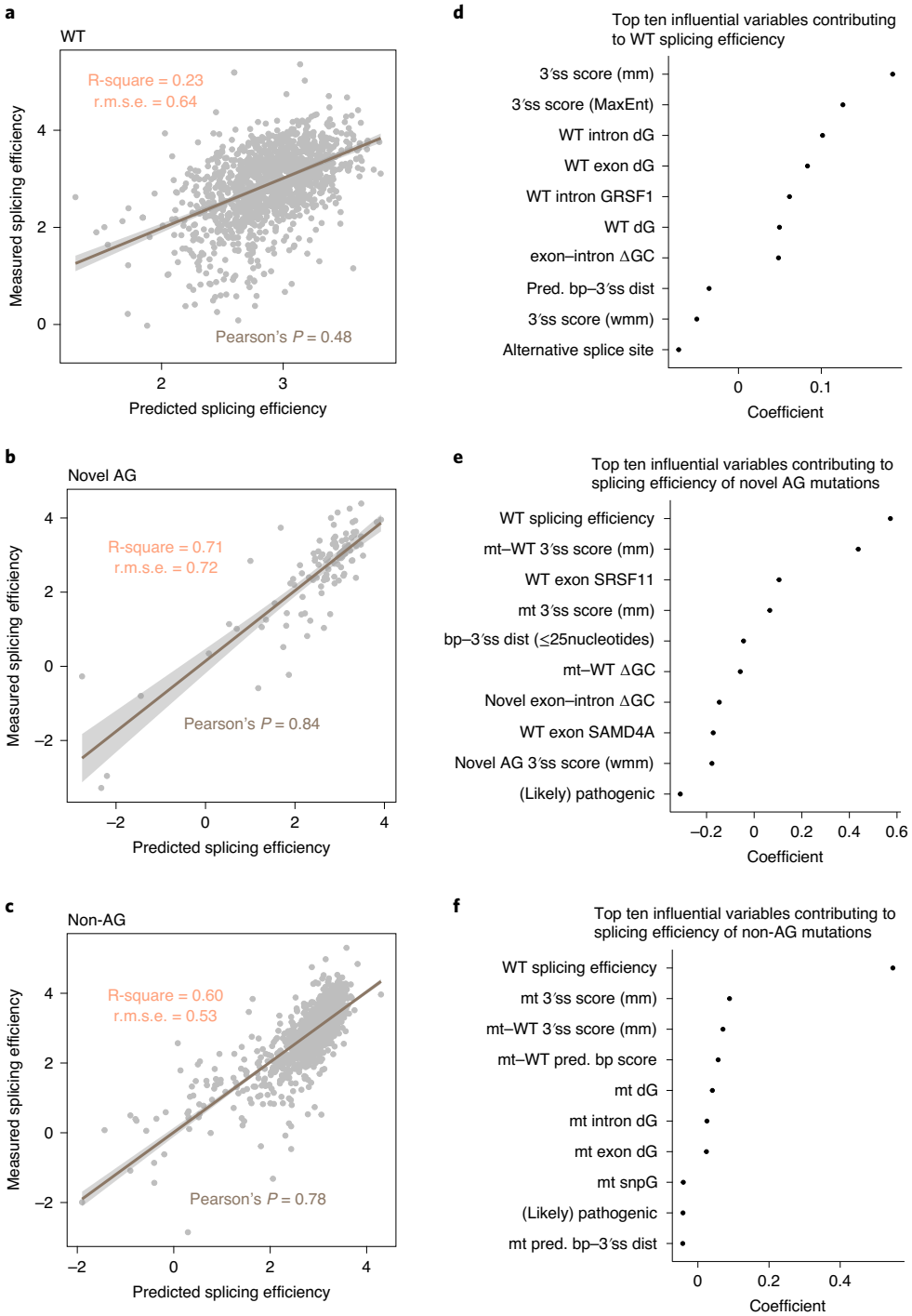


Fig. 6 | Generalized linear model to explain splicing efficiency. **a**–**c**, Splicing efficiency models of WT (**a**), novel AG mutations (**b**) and non-AG mutations (**c**). The explanatory power of each model on the test dataset was estimated by Pearson's correlation (two-tailed), R-square and r.m.s.e. The gray area displays the 95% confidence interval for predictions from the linear model. **d**–**f**, The contributory factors to the WT model (**d**), novel AG mutations model (**e**) and non-AG mutations model (**f**).

In line with the exon definition model, long introns and high exonic folding energy suppress the splicing change (Fig. 4f). We calculated AUC = 0.962 from the ROC for the test data (Fig. 4g). Contributory factors for splicing defects and their coefficients are listed in Supplementary Table 2, and a summary of the models is given in Fig. 4h–j.

To generalize the applicability of our model, we generated models independent of WT splicing efficiency (Extended Data Fig. 7a–d and Supplementary Table 2) and then examined their performance by predicting the splicing outcomes of 107 near-exon 3′ intronic variants (−44 to −18 to the 3′ss) and 314 near-exon 5′ (−3 to +30 to the 5′ss) variants, whose splicing defects have been assayed experimentally (Fig. 5)[27,30]. These models predicted almost perfectly the effect of 5′ variants and 3′ add-AG variants, and the effect of the non-AG variants was predicted with >91% sensitivity (Fig. 5b). Remarkably, the model reliably predicted both BS and nonBS variants (Fig. 5b), as well as 5′ variants outside the splice site (Fig. 5c). Finally, to test model performance in terms of

predicting in vivo splicing outcome in a genomic context, we examined the effects of sixty-six 3′ intronic variants for which in vivo splicing has been assessed by sequencing or RT–PCR. The results show that in vivo splicing effects are also well predicted by our model (Fig. 5b). Finally, compared with the deep-learning-based predictor, our model outcompetes the sensitivity of SpliceAI[8] and MMSplice[15] without sacrificing much specificity (Fig. 5). Together, these findings demonstrate that our model illustrates the features underlying splicing defects and predicts the splicing defects for near-exon intronic variants in minigene and genomic contexts.

Splicing efficiency models complement splicing defect models

Our experimental design (Fig. 1a) provided a unique opportunity to assess the effect of a variable 3′ss region against fixed 5′ss regions and exons. Again, we used LASSO regression to model the features contributing to splicing efficiency. Our model identified 3′ss score as the most important factor explaining splicing efficiency. Intriguingly, the model also revealed folding energy of both exons and introns to be a profound contributor promoting splicing efficiency. Moreover, if the exon represents an alternatively spliced region, it generally displayed low splicing efficiency (Fig. 6a). The model reflects experimental data on splicing efficiency with a Pearson's correlation of 0.48. We adopted a similar approach to model the splicing efficiency of add-AG and non-AG variants. Both models identified WT splicing efficiency and the difference between mt and WT 3′ss score as dominant features for predicting mt splicing efficiency (Fig. 6b,c and Supplementary Table 2). In agreement with the prediction of splicing defects (Fig. 4b,d), distal branchpoints and pathogenicity, as reported in databases, were correlated with weaker mt splicing efficiency (Fig. 6b,c). Overall, these models adequately explain the splicing efficiency of add-AG variants (Pearson's correlation of 0.84) and that of non-AG variants (Pearson's correlation of 0.78). Models for add-AG and non-AG variants, but excluding WT splicing efficiency further emphasized the contribution of mt 3′ss strength, differential exon-to-intron GC content and mt folding energy (Extended Data Fig. 7e–h and Supplementary Table 2). Together, these models indicate that both splice-site strength and structural cues determine splicing efficiency.

Intrinsic property of the BS corresponds to splicing outcome

Our modeling results indicate that dependency on the BS for splicing is sensitive to its location relative to the 3′ss, nucleotide composition and base-pairing energy with U2 snRNA. Disruption of a strong and evolutionarily conserved BS likely impairs splicing (Figs. 4 and 6). Therefore, we further analyzed these traits in a genomic context. We first analyzed the BS distribution relative to the 3′ss and contingent on the branchpoint sequence[5]. The canonical A and C branchpoints peaked at the −25 nucleotide position upstream of the 3′ss, whereas T branchpoints peaked at the −28 nucleotide position, likely because of skipping of reverse transcriptase to the −2 position (Extended Data Fig. 8a). Overall, the branchpoints are located between −18 and −40 nucleotides upstream of the 3′ss. Because A and C are the most frequently reported functional branchpoints, the similar distribution of HC significant variants (Extended Data Fig. 5a; median, −25 nucleotides) suggests disruption of functional branchpoints. Branchpoints located at positions between −20 and −30 nucleotides from the 3′ss are most likely to support constitutive splicing (Extended Data Fig. 8b). Additionally, branchpoints between the −18 and −26 nucleotide position base pair most strongly with the U2 BS recognition sequence (Extended Data Fig. 8c) and are the most evolutionarily conserved (Extended Data Fig. 8d), indicating an optimal functional capacity for BS in the region between the −20 and −30 nucleotide positions upstream of the 3′ss and further implying nonuniform evolutionary pressure for BS relative to the 3′ss. This genome-wide BS profile supports functional dependency of the BS on intrinsic properties and position.

Functional impact of intronic splicing variants in TWB data

To examine whether the splicing defect of BS variants influences population health, we investigated the disease association and biochemical index of individuals with BS variants in the TWB[31]. We used a collection of 68,978 community samples from the TWB. We found that 16 (15 imputed) of the experimentally defined and 328 SpliceAPP-predicted significant splicing variants were probed in the TWB and displayed allele frequencies >0.01. To test whether the splicing variants associate with abnormal physiological presentation, we employed a quantile–quantile plot to compare the *P* value distribution of the significant splice-altering intronic SNPs associated with skewed biochemical indices against theoretical *P* values (Fig. 7a). Inflation of *P* values toward the splicing variants suggests that TWB subjects harboring splicing variants are more likely to display abnormal biochemical phenotypes. About half of the splicing variants were associated with self-reported diseases and/or a skewed biochemical index relative to the major allele (Supplementary Table 3 and Fig. 7b,c). These results indicate that splicing defects derived from the intronic variants potentially alter physiological fitness.

To further examine the underlying etiological mechanism, we examined one of the disease-associated branchpoint mutations from the TWB. We found that branchpoint mutation rs72835097 of *Calmodulin-binding transcription activator 2* (*CAMTA2*) was associated with hypertension in human males (Fig. 7b) and increased platelet counts in aged individuals (Fig. 7c). This outcome supports the findings of a previous study showing that CAMTA2 is a transcriptional coactivator involved in cardiac growth and is responsible for expression of natriuretic peptide A (NPPA)[32]. Reduced NPPA levels result in salt-sensitive hypertension in NPPA-disrupted mouse models[33,34]. In addition, meta-analyses have indicated that *CAMTA2* polymorphisms are associated with platelet properties[35,36]. These congruent findings prompted us to further characterize rs72835097 BS variants.

The rs72835097 mutation, representing an A-to-G transition at position −41 in intron 14 of *CAMTA2*, was one of the predicted branchpoints displaying a moderate conservation score (Fig. 7d) and this has been verified by sequencing. Mutation of −41A resulted in partial loss of the normal spliced form in a minigene assay (Fig. 7e), conducted similarly to our library design (Fig. 1a). Using Lariat PCR, we discovered two probable branchpoints in *CAMTA2* intron 14 at positions −35A and −41A (Fig. 7f). Splicing results for the full-length *CAMTA2* exons 14–16 genomic fragment indicated that the −41A and −35A branchpoints are partially redundant, but that −41A is the predominant branchpoint (Fig. 7g). *CAMTA2* exon 15 skipping causes a frameshift error and premature translation termination. The resulting truncated protein lacks the nuclear localization signal and ankyrin-repeat domain, the latter being necessary for histone deacetylase (HDAC5) binding to repress CAMTA2 transactivation activity[32]. A previous study has shown that CAMTA2 is recruited to the *NPPA* promoter to induce cardiac growth by associating with the cardiac homeodomain protein Nkx2-5 (ref. [32]). Accordingly, we tested the transactivation activity of WT and truncated CAMTA2 by means of a promoter luciferase assay with a Nkx-binding element (NKE). The truncated CAMTA2 isoform failed to activate the NKE promoter and interact with HDAC5 (Fig. 7h). Finally, we tested the requirement of CAMTA2 activity on expressing cell adhesion molecule CD62P (P-selectin) associated with activated endothelial cells and platelets[37]. The results showed that WT CAMTA2 activity induced *CD62P* expression, whereas truncated CAMTA2 did not (Fig. 7i). Overall the evidence suggested that *CAMTA2* branchpoint mutation abolished its normal function in transcription transactivation relevant to human health. Collectively, we have demonstrated that BS variants are associated with disease and abnormal biochemical indices in the Taiwanese human population, and that diagnosis of splicing variants lays the foundation for precision medicine.

Discussion

Given exponential increases in reports on genetic variation in this postgenomic era, precise annotation and interpretation of each variant is urgently needed. A significant proportion of genetic variation alters splicing, thereby leading to disease. However, because of the ill-defined intronic splicing regulatory elements, it is difficult to identify

https://doi.org/10.1038/s41594-022-00844-1

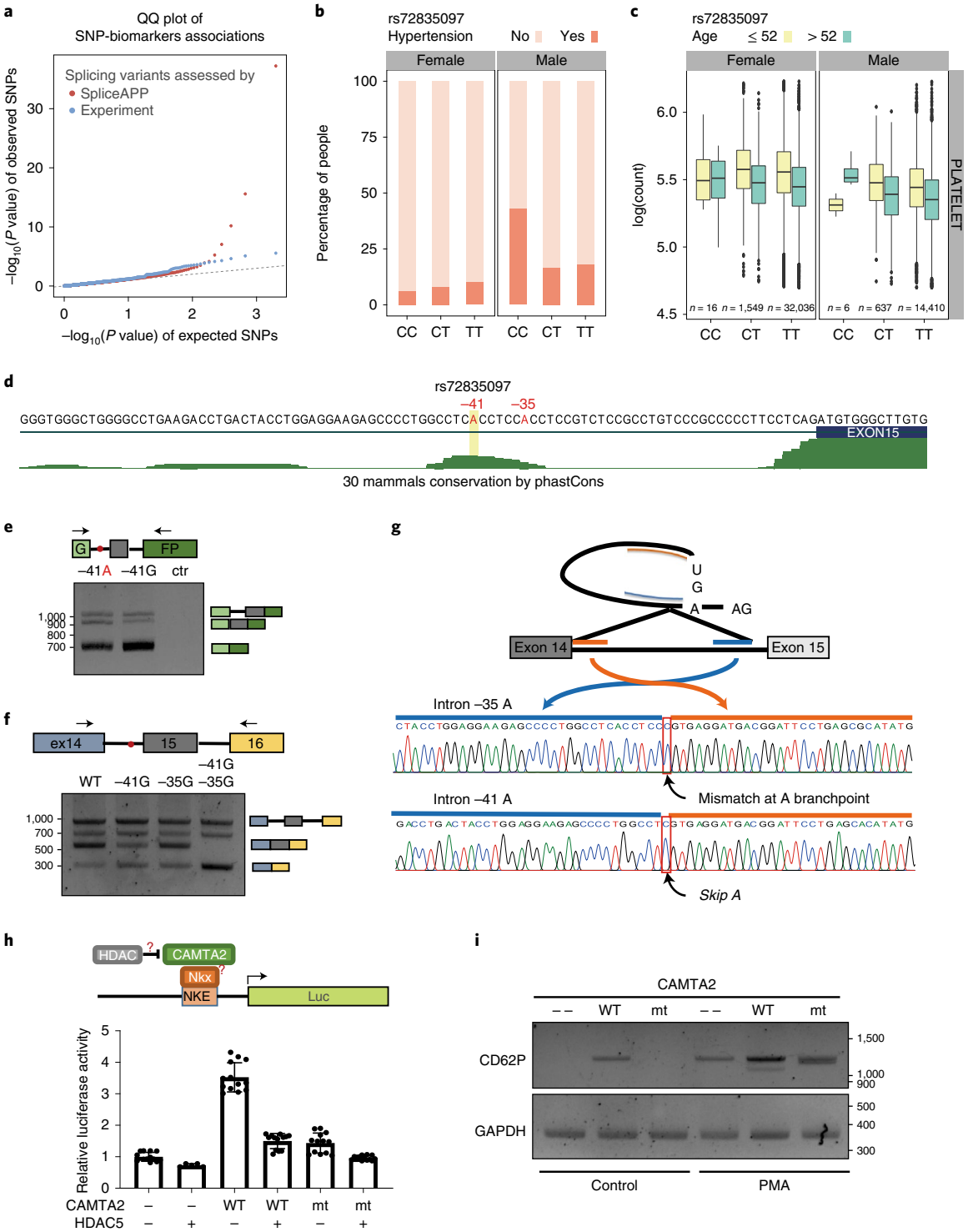


Fig. 7 | Functional branchpoint mutations in the TWB. a, Quantile–quantile plot of the nominal *P* value distribution of SNPs associated with biochemical indices in the TWB, determined by logistic/linear regression. The dashed line represents a hypothetical distribution equal to the theoretical *P* value. **b**, Significantly higher risk of hypertension arising from the *CAMTA2* branchpoint SNP rs72835097 minor allele in homozygous male carriers. **c**, Significantly skewed platelet count arising from the rs72835097 minor allele among homozygous carriers in the TWB. The boxes in box plots represent medians (central line) and interquartile ranges (25th to 75th percentile). Whiskers indicate ±1.5× the interquartile range from the box or the last data point within that, and the dots show outliers. **d**, Higher conservation level around two predicted branchpoints of *CAMTA2* intron 14, especially at position −41A (rs72835097). **e**, Minigene assay of partial *CAMTA2* intron 14 and exon 15 ligated to GFP exons, as shown in Fig. 2a. **f**, Lariat PCR traversing the lariat junction identified two branchpoints (positions −35 and −41) in *CAMTA2* intron 14. **g**, Minigene assay of full-length *CAMTA2* exons 14–16. The −35 and −41 branchpoints were mutated into G, individually or in concert. **h**, Transactivation activity of WT or truncated CAMTA2 on the NPPA promoter, as revealed by luciferase reporter assays. NKE binds to Nkx2-5 and CAMTA2. The CAMTA2 interactor HDAC5 was coexpressed to test the antagonizing effects of CAMTA2 and HDAC5. $n = 13$ or 5 biologically independent samples. Data are presented as mean ± s.d. **i**, CAMTA2-mediated CD62P expression in HEK293T cells with or without PMA treatment. A representative experiment of three repeats is presented (**e**,**f**,**i**).

intronic splicing variants. It is particularly challenging to interpret 3′-end intronic variations for two major reasons. First, unlike 5′ss in which the splice-site motif is strictly located to six nucleotides in the intron, 3′ss recognition requires a three-nucleotide 'YAG' splice site, a polypyrimidine tract and a BS, which can extend up to 60 nucleotides into the intron, with some exceptionally distal BS[2,5]. Second, the human BS motif is not well-defined and it remains unclear whether there is redundancy among multiple BS. Hence, to identify the determinants of splicing defects arising from variations in the near-exon intronic region, we collected data on disease-relevant and rare BS variants and examined their effect on splicing by means of MaPSy.

In searching for features corresponding to the defective splicing phenotype, we found that the canonical 3′ss strength, the variant position relative to the splice site and the conservation level of the variant position are negatively correlated with the splicing defect (Fig. 4 and Extended Data Fig. 5). Baeza-Centurion et al.[38] proposed that splicing variants are concentrated around alternatively spliced exons with intermediate inclusion levels, suggesting a protective role for constitutive splicing against sequence variations. Furthermore, exonic structural features, such as highly differential exon-to-intron GC content and high exonic folding energy also preserve the splicing reaction against intronic variation (Fig. 4 and Extended Data Fig. 5). This scenario supports the notion that exonic mutations are more likely to interrupt splicing in diseases frequently caused by splicing mutations[20,39]. However, the BS with higher sequence conservation (stronger base pairing with U2 snRNA) and the branchpoint position (bp 0) are more sensitive to mutations (see models in Fig. 4). Thus, well-defined exons and robust splice sites are rather immune to intronic sequence variation, whereas well-conserved BS are relatively indispensable for splicing.

To better explain the influence of variants on splicing, we divided the variants into 5′ and 3′, which include potential novel 3′ss (add AG) and other (non-AG) mutations, and then trained three models by LASSO regression, before testing the models on datasets of near-exon intronic mutations. Although the 3′ models were trained on BS mutations, they performed surprisingly well even on the nonBS variants, potentially because the models encompass general features of the variants, such as location and conservation level. We found that the exonic features in the intron, such as GC content, SRSF binding sites, are particularly important in activating the novel 3′ss, suggesting that the 3′ss selection depends on exon definition. Interestingly, although it seems that high splicing efficiency and strong splice sites can preserve splicing in general, in our 3′ add-AG model, splicing efficiency is not a predictive feature. In other words, the splice-site competition model depends on the relative strength of the two competing 3′ss, but not on the absolute strength of the canonical 3′ss. For the 3′ model, we did not expect to find that presence/absence or number of additional 'As' between −18 and −40 nucleotides (the optimal branchpoint location) hardly affected the splicing outcome, because it has been hypothesized that additional As near BS mutations could be used as a cryptic BS to rescue splicing[40]. Also, in all our models, we found that local structural 'openness' had a limited impact on splicing prediction, although it has been proposed previously that there is a pronounced preference for or against splice sites with secondary structures in different settings[20,41]. Finally, we did not detect more severe splicing defects from indels compared with SNPs, which was also unexpected considering the greater sequence alteration caused by indels.

By contrast to the 3′-end, a novel 5′ss is not a major splice-altering mechanism of the 5′ intronic variants. Although exonic structure does protect the splicing from the sequence variation, the destruction of canonical 5′ss is the main cause of 5′ss dysfunction. Furthermore, although the PhyloP score, a per-base conservation score, contributes to both 3′ and 5′ prediction, the phastCons score that measures conservation of a sequence window did not inform 3′ prediction at all. Together with the fact that 3′-end introns display lower phastCons scores, this outcome suggests that the 3′-end region we analyzed was degenerate with little conservation beyond a 'motif' level, whereas the 5′ intronic splicing signals were more structured and consolidated in sequence blocks. This observation supports the notion that 3′ss selection involves more complex splicing regulation.

Finally, we compared our predictive power with SpliceAI developed by Illumina[8] and MMSplice[15], and found that our models out-performs both, especially in terms of its sensitivity for detecting intronic splicing variants (Fig. 5) when using the recommended cutoff (△ score 0.5 of SpliceAI, delta_logit_psi |2| of MMSplice). Indeed, we achieved greater sensitivity even when compared with a high-recall cutoff (△ score 0.2) specified by SpliceAI (3′: 90.3% versus 54.8%; 5′: 100% versus 82.4%), as well as for our in vivo splicing data (3′: 85.7% versus 9.5%). Importantly, because our principal goal was to identify splicing variants among variants of unknown significance, we feel it is justifiable to maintain a high-sensitivity cutoff. These results suggest that model training against specific classes of variants, such as intronic mutations, is necessary to accurately annotate the variants' effects, especially relative to generic deep learning on mock splice sites.

In conclusion, our study addresses the need to explain and classify intronic variants that affect splicing and, conceivably, provides a molecular foundation for interpreting intronic mutations in the context of genetic diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41594-022-00844-1.

References

1. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**, 437–451 (2017).
2. Wilkinson, M. E., Charenton, C. & Nagai, K. RNA splicing by the spliceosome. *Annu. Rev. Biochem.* **89**, 359–388 (2020).
3. Gooding, C. et al. A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones.*Genome Biol.* **7**, R1 (2006).
4. Mercer, T. R. et al. Genome-wide discovery of human splicing branchpoints. *Genome Res.* **25**, 290–303 (2015).
5. Taggart, A. J. et al. Large-scale analysis of branchpoint usage across species and cell lines. *Genome Res.* **27**, 639–649 (2017).
6. Pineda, J. M. B. & Bradley, R. K. Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes Dev.* **32**, 577–591 (2018).
7. Gao, K. P., Masuda, A., Matsuura, T. & Ohno, K. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res.* **36**, 2257–2267 (2008).
8. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535 (2019).
9. Lim, K. H., Ferraris, L., Filloux, M. E., Raphael, B. J. & Fairbrother, W. G. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl Acad. Sci. USA* **108**, 11093–11098 (2011).
10. da Costa, P. J., Menezes, J. & Romao, L. The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease. *Int. J. Biochem. Cell Biol.* **91**, 168–175 (2017).
11. Group, P. T. C. et al. Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).
12. Gupta, A. K. et al. Degenerate minigene library analysis enables identification of altered branch point utilization by mutant splicing factor 3B1 (SF3B1). *Nucleic Acids Res.* **47**, 970–980 (2019).
13. Cheung, R. et al. A multiplexed assay for exon recognition reveals that an unappreciated fraction of rare genetic variants cause large-effect splicing disruptions. *Mol. Cell* **73**, 183 (2019).

14. Xiong, H. Y. et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
15. Cheng, J. et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* **20**, 48 (2019).
16. Pertea, M., Lin, X. & Salzberg, S. L. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* **29**, 1185–1190 (2001).
17. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).
18. Jian, X., Boerwinkle, E. & Liu, X. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet. Med.* **16**, 497–503 (2014).
19. Riepe, T. V., Khan, M., Roosing, S., Cremers, F. P. M. & 't Hoen, P. A. C. Benchmarking deep learning splice prediction tools using functional splice assays. *Hum. Mutat.* **42**, 799–810 (2021).
20. Soemedi, R. et al. Pathogenic variants that alter protein code often disrupt splicing. *Nat. Genet.* **49**, 848–855 (2017).
21. Lin, H. et al. RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants.*Genome Biol.* **20**, 254 (2019).
22. Jagadeesh, K. A. et al. S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat. Genet.* **51**, 755 (2019).
23. Stenson, P. D. et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
24. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).
25. Sherry, S. T., Ward, M. H. & Sirotkin, K. dbSNP – Database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677–679 (1999).
26. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
27. Adamson, S. I., Zhan, L. & Graveley, B. R. Vex-seq: high-throughput identification of the impact of genetic variation on pre-mRNA splicing efficiency. *Genome Biol.* **19**, 71 (2018).
28. Amit, M. et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep.* **1**, 543–556 (2012).
29. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).
30. Leman, R. et al. Assessment of branch point prediction tools to predict physiological branch points and their alteration by variants. *BMC Genomics* **21**, 86 (2020).
31. Lin, J. C., Fan, C. T., Liao, C. C. & Chen, Y. S. Taiwan Biobank: making cross-database convergence possible in the Big Data era. *Gigascience* **7**, 1–4 (2018).
32. Song, K. et al. The transcriptional coactivator CAMTA2 stimulates cardiac growth by opposing class II histone deacetylases. *Cell* **125**, 453–466 (2006).
33. John, S. W. M. et al. Genetic decreases in atrial-natriuretic-peptide and salt-sensitive hypertension. *Science* **267**, 679–681 (1995).
34. Chan, J. C. Y. et al. Hypertension in mice lacking the proatrial natriuretic peptide convertase corin. *Proc. Natl Acad. Sci. USA* **102**, 785–790 (2005).
35. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
36. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415 (2016).
37. Massaguer, A. et al. Characterization of platelet and soluble-porcine P-selectin (CD62P).*Vet. Immunol. Immunopathol.* **96**, 169–181 (2003).
38. Baeza-Centurion, P., Minana, B., Valcarcel, J. & Lehner, B. Mutations primarily alter the inclusion of alternatively spliced exons.*eLife* **9**, e59959 (2020).
39. Braun, S. et al. Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nat. Commun.* **9**, 3315 (2018).
40. Chiang, H. L., Wu, J. Y. & Chen, Y. T. Identification of functional single nucleotide polymorphisms in the branchpoint site. *Hum. Genomics* **11**, 27 (2017).
41. Mikl, M., Hamburg, A., Pilpel, Y. & Segal, E. Dissecting splicing decisions and cell-to-cell variability with designed sequence libraries. *Nat. Commun.* **10**, 4572 (2019).
42. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
43. Corvelo, A., Hallegger, M., Smith, C. W. J. & Eyras, E. Genome-wide association between branch point properties and alternative splicing.*PLoS Comput. Biol.* **6**, e1001016 (2010).

Methods

Construction of MaPSy library DNA templates

MaPSy library DNA templates were assembled by overlap extension PCR with Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific). Initially, the oligonucleotide library sequences (CustomArray) were amplified into double strands by PCR and cleaned up using columns (Qiagen). The intron-containing enhanced GFP backbones were created according to a previous study[44], followed by insertion of *CAMTA2* exon 15 and its flanking introns between the BamHI and SalI sites (pGint-*CAMTA2* exon 15). The library amplicons replaced the 3′ss of *CAMTA2* exon 15 through sequential overlapping PCR. The final product contained a CMV promoter and three exons in which the first 3′ss is the oligonucleotide library. Details of the procedure can be found in Extended Data Fig. 2. Finally, the assembled full-length DNA templates were again subjected to PCR clean-up. Before the following transfection experiment, we conducted next-generation DNA sequencing to check sequence completeness, which revealed that 93.6% of the MaPSy library pairs had been successfully reconstituted.

Overexpression, RNA extraction and RT–PCR

HEK293T cells cultured in six-well plates were transfected with 1 µg of constructs and harvested 24-h posttransfection. For the phorbol 12-myristate 13-acetate (PMA) assay, cells were treated with 10 nM PMA for 24 h at 37 °C. Total RNA was extracted according to the Direct-zol RNA MiniPrep Plus kit (Zymo Research) instructions for RNA extraction. cDNA was prepared from 2 µg of total RNA using SuperScript IV reverse transcriptase (Thermo Fisher Scientific) with a random hexamer, following the manufacturer's protocol. All RT–PCR products were sequenced to confirm the spliced outcome (normal splicing, intron inclusion and exon skipping). See Supplementary Table 4 for primer sequences.

Amplicon sequencing reads alignment

The PCR amplicons were subjected to Illumina NextSeq 150 single-end sequencing. Between 0 and 3 random nucleotides were attached at the end of the amplicons to ensure balanced fluorescence detection on the NextSeq platform. Single-end reads were aligned to our synthetic 'reference genome' by hisat2 (v.2.2.1). High-quality (q60) reads selected by SAMtools[45] (v.1.7) crossing barcode position with GT–AG splice junctions annotation by RegTools[46] (v.0.5.2) were preserved as spliced reads. Reads without junctions spanning the most used splice-site position were preserved as unspliced reads.

Criteria for selecting splice-altering mutation candidates

WT and mt were separated into three categories: spliced versus unspliced, canonical versus noncanonical in all spliced reads and canonical versus noncanonical plus unspliced reads. In case the minigene assay could not reflect an unanimous splicing outcome in the genomic context, the canonical definition we used is the most used splice site in WT across four repeats and derived mt canonical splice site by WT result. To assess the effects of the variant in splicing efficiency and accuracy, a two-sided Fisher's exact test was applied followed by false discovery rate correction to identify splicing variants altering the ratios of spliced to unspliced, canonical to noncanonical and/or canonical to both unspliced and noncanonical isoforms. Both WT/mt pairs exceeding 100 read counts with a q-value <0.05 across four repeats are classified as significant. Candidates with a twofold odds ratio change with either WT or mt having >5% unspliced and noncanonical reads are recognized as HC splicing variants.

Minigene splicing for validation

DNA oligos (Integrated DNA Technologies) of selected MaPSy candidate sequences were amplified into double strands by PCR with Phusion High-Fidelity DNA Polymerase. The PCR products were then cloned into pGint-*CAMTA2* exon 15 via the BbsI and SmaI sites. The resulting constructs were expressed and recovered from HEK293T as described above. Splicing isoforms were amplified by primers targeting the first two exons of the minigene (Lib0F and Lib0Rl), resolved by electrophoresis and visualized using a Bio-Rad Gel Doc EZ System. The intensity of each splicing isoform was quantified using Image J (National Institutes of Health).

Origins of the factor input for model building

Exon conservation. We calculate the average score retrieved from the University of California Santa Cruz Genome Browser (https://genome.ucsc.edu/) PhyloP basewise conservation score derived from Multiz alignment of 7 vertebrates, 20 mammals, 30 primates and 100 vertebrates.

Splice-site strength. MaxEntScan[42] was used to calculate 3′ss strength by 23-mers (20 intronic and 3 exonic sequences) and 5′ss strength by 9-mers (3 exonic and 6 intronic sequences) using different models, including the maximum entropy model, first-order Markov model, weight matrix model and maximum dependence decomposition model (5′ only). We analyzed other 3′ss essential elements; BS and polypyrimidine tract strength were analyzed using a mammalian U2 branchpoint prediction tool SVM-BPfinder[43] to score the branchpoint sequence using the Support Vector Machine (SVM) learning algorithm classifier and predicted polypyrimidine tract length and score.

RNA secondary structure and U2 interaction prediction. We used the Vienna RNA package (v.2.4.8)[47] to calculate the minimum free-energy structures and base pair probabilities in our minigene library, and to predict the dimer-forming secondary structures of the BS sequences using the U2 snRNP by RNAcofold function.

Motif enrichment. Sequences are analyzed by Weblogo 3 (ref. [48]) to identify motifs enriched in significant or nonsignificant candidate library segments.

Splicing enhancer and silencer. ESE, ESS, ISE and ISS were collected from published literature[49–51]. They were mapped to exonic and intronic sequences in each WT or mt minigene construct, and the novel exon sequences annotated by the mutations creating novel AG as an alternative 3′ss.

RBP prediction. RBPs were collected from the ATtRACT database[52] to scan for specific motifs on sequences by position weight matrix. A 95% position weight matrix match as defined by R package Biostrings (v.2.62.0) (https://bioconductor.org/packages/Biostrings) is the minimum for counting a match.

For statistical confidence, only pairs in which both WT and mt have more than 100 aligned reads in all four experiments and for which the canonical splice sites are the most used splice sites in the WT minigene construct were input for statistical model building.

Statistical models built by linear regression

We build statistical models with LASSO regularization to predict splicing effects and splicing efficiency in the R package glmnet[53,54] (v.4.1-4). For the splicing effect models, the outcome Y is distributed in a binary fashion; that is, a mutation either affects or does not affect (1 or 0) the splicing decision, therefore:

$$\log\frac{P(Y_i=1)}{1-P(Y_i=1)}=\beta_0+\sum_{j=1}^{p}\beta_j X_{ji}$$

where $Y_i=1$ is the ith mt sequence that significantly affects splicing, and X_{ji} is the feature j in the ith sequence. The predictive features were scaled according to the standard deviation for comparable coefficients. Intronic variants were categorized as 3′ novel AG-creating variants, 3′ non-AG-creating variants and 5′ variants to model their contribution to splicing decisions in our independently built models. LASSO

regularization minimizes $L_{\log} + \lambda \sum_{j=1}^{p} |\beta_j|$, where the negative log likelihood is:

$$L_{\log} = -\sum_{i=1}^{n}\left[-\ln\left(1+e^{\left(\beta_0+\sum_{j=1}^{p}\beta_j X_{ji}\right)}\right)+y_i\left(\beta_0+\sum_{j=1}^{p}\beta_j X_{ji}\right)\right]$$

We utilized tenfold cross-validation to choose the tuning parameter λ. Ultimately, the value of λ that gives the minimum mean cross-validated error was selected to build each model. In our splicing efficiency models, logit transformation is applied to the continuous outcome *Y*, representing WT or mt splicing efficiency (percentage of the canonically spliced read count of each sequence). Two-thirds of the samples were randomly selected for our training models, and the remainder were used to test model performance. The classification power of the splicing effect models is measured by ROC, and the performance of splicing efficiency models are measured by R-square and root mean square error (r.m.s.e.).

SpliceAI splicing-altering variants prediction

We used SpliceAI v.1.3.1 to predict whether variants affect splicing outcome. For near 3′ss variants, any delta score in the columns 'DS_AG Delta score (acceptor gain)' or 'DS_AL Delta score (acceptor loss)' higher than 0.2 (high recall) or 0.5 (recommended) are characterized as splicing-altering variants. Because of our 3′ss variants collection, parameter -D, the maximum distance between the variant and gained/lost splice site, was set as 200. For near 5′ss variants, any delta score in the columns 'DS_DG Delta score (donor gain)' or 'DS_DL Delta score (donor loss)' higher than the 0.2 (high recall) or 0.5 (recommended) are characterized as splicing-altering variants.

MMSplice splicing-altering variants prediction

MMSplice was downloaded from https://github.com/gagneurlab/MMSplice_MTSplice, installed and performed in Python v.3.7.11. The required inputs for MMSplice, the reference genome (FASTA) file and the genome annotation (GTF) file, were downloaded from the Ensembl database (GRCh38.105/GRCh37.87)[55].

ROC and AUC calculation

The ROC curve was constructed by plotting sensitivity versus specificity for various thresholds to evaluate the diagnostic performance of each prediction model. AUC was estimated using the 'auc' function in the R package precrec[56] (v.0.12.9).

The Taiwan Biobank data

Genotyping and phenotypic data of 68,978 Taiwanese people were obtained from the TWB[31,57] (https://www.biobank.org.tw/) to determine the association of candidate SNPs and the recorded phenotypes, including various diseases and biochemical indices. Disease information was self-reported and collected using questionnaires. Each participant was genotyped by a specifically designed chip—an Affymetrix Axiom genome-wide TWB 2.0 array plate with a total of 752,921 SNPs. The study was approved by the Institutional Review Board of Academia Sinica.

Imputation and quality control of genotyping data

Imputation. Before retrieving selected SNPs from the genotyping data, imputation was performed using SHAPEIT2 (v.2.r790) and IMPUTE2 (v.2.3.1) with whole genome sequencing data of 973 individuals from TWB and 504 East Asian individuals from the 1,000 genome project as a reference panel. The following quality control steps were conducted based on filtered imputed data with an information score higher than 0.3.

Quality control. To ensure the reliability of the genotyping data, a series of quality control procedures was employed to remove chips with low quality and problematic individuals using PLINK v.1.9 and PLINK v.2.0 (refs. [58–60]).

Per sample quality control. First, we checked the sex discrepancy between the gender recorded in the questionnaire and the gender based on their X chromosome heterozygosity/homozygosity rates. Second, we excluded individuals with a missing call rate (the proportion of missing SNPs for each participants) >0.05 or heterozygosity rate of ±3 s.d. from the population mean. Third, related individuals were determined by kinship coefficient and removed for pairs having a kinship value >0.0884, excluding second-degree relations[61]. Lastly, using autosomal chromosome SNP genotyping data of 1,397 people from the 1,000 genome project (phase 3) followed by principal component analysis, we excluded individuals whose genetic information is far from the Chinese Han population.

Per SNP quality control. SNPs with a low genotyping call rate (<0.05) or low minor allele frequency (<0.01) were excluded from subsequent analyses. We also tested for SNPs that significantly violated Hardy–Weinberg equilibrium at $P < 1 \times 10^{-7}$.

Association analyses

We performed statistical analyses to evaluate the association of 16 selected markers with 23 self-reported traits and 24 biochemistry indices levels. Linear regression was used to evaluate the association of each SNP with a continuous biochemical index, and logistic regression was implemented to test the association of each SNP with a trait that is dichotomous. For biochemical indices having more than two levels, multinomial logistic regression models were fitted using the R package nnet[62] (v.7.3-17) and likelihood ratio tests were used to determine *P* values. All SNPs were tested using codominant genetic models. Square root of age, gender, dwelling place and batch of array were included in the regression models to adjust for confounding effects. To control for population stratification, the first ten principal components were also included as covariates in the models.

Luciferase assay

Twenty-four hours after cotransfection with 200 ng of pGL3-Basic *NPPA* promoter constructs with 200 ng of different pCMV-Tag 2A *CAMTA2* isoforms and 20 ng of pRL-TK vector, with or without 200 ng of GFP-HDAC5 (Addgene Plasmid no. 32211), HEK293T cells cultured in 12-well plates were harvested and processed according to the instructions for the Dual-Glo Luciferase Assay System (Promega). Afterwards, 50 µl of clear supernatant was transferred to 96-well black plates for the measurement of firefly luminescence using an EnSpire Multimode Plate Reader (PerkinElmer). Subsequently, 50 µl of Dual-Glo Stop & Glo Reagent was added into each well. After incubation at room temperature for 10 min, *Renilla* luminescence was measured as above. All experiments were performed in triplicate and repeated at least three times.

Reporting summary

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

Data availability

RNA Sequencing datasets generated during this study are available at the NCBI GEO: GSE179892. Other databases used in the study: UCSC PhyloP: https://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=1351580935_14MOQtNDW7V78RaXEDp3Yy4m4PTb&c=chr2&hgTracksConfigPage=configure&hgtgroup_compGeno_close=0#compGenoGroup ATtRACT: https://attract.cnic.es/download; Ensembl: https://asia.ensembl.org/info/data/ftp/index.html. Source data are provided with this paper. Further information and requests for resources should be directed to and will be fulfilled by the corresponding author.

Code availability

Custom codes and the training features used in the study are available at https://github.com/chienlinglin/modeling-intron-variants/.

References

44. Bonano, V. I., Oltean, S. & Garcia-Blanco, M. A. A protocol for imaging alternative splicing regulation in vivo using fluorescence reporters in transgenic mice. *Nat. Protoc.* **2**, 2166–2181 (2007).
45. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. Cotto, K. C. et al. RegTools: Integrative analysis of genomic and transcriptomic data to identify splice altering mutations across 35 cancer types.*Cancer Res.* **80**(16 Suppl), 2136 (2020).
47. Lorenz, R. et al. ViennaRNA Package 2.0.*Algorithm Mol. Biol.* **6**, 26 (2011).
48. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: A sequence logo generator. *Genome Res* **14**, 1188–1190 (2004).
49. Ke, S. et al. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* **21**, 1360–1374 (2011).
50. Culler, S. J., Hoff, K. G., Voelker, R. B., Berglund, J. A. & Smolke, C. D. Functional selection and systematic analysis of intronic splicing elements identify active sequence motifs and associated splicing factors. *Nucleic Acids Res.* **38**, 5152–5165 (2010).
51. Wang, Y., Ma, M., Xiao, X. & Wang, Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat. Struct. Mol. Biol.* **19**, 1044–1052 (2012).
52. Giudice, G., Sanchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT – a database of RNA-binding proteins and associated motifs. *Database (Oxford)* **2016**, baw035 (2016).
53. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
54. Tibshirani, R. et al. Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Series B Stat. Methodol.* **74**, 245–266 (2012).
55. Howe, K. L. et al. Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
56. Saito, T. & Rehmsmeier, M. Precrec: fast and accurate precision-recall and ROC curve calculations in R. *Bioinformatics* **33**, 145–147 (2017).
57. Lin, J. C., Hsiao, W. W. W. & Fan, C. T. Transformation of the Taiwan Biobank 3.0: vertical and horizontal integration. *J. Transl. Med* **18**, 304 (2020).
58. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
59. Shaun Purcell, C. C. PLINK. v.1.9 edn; www.cog-genomics.org/plink/1.9/ (2019).
60. Shaun Purcell, C. C. PLINK. v.2.0 edn; www.cog-genomics.org/plink/2.0/ (2019).
61. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
62. Ripley, B., Venables, W. & Ripley, M. B. Package 'nnet'. R. package v.7, 3–12 (2016).

Acknowledgements

We thank M.-C. Tsai, Senior Scientific Editor at *Cell*, for constructive advice and editing the manuscript. We thank the Genomics Core of Institute of Molecular Biology (IMB), Academia Sinica, for performing the amplicon sequencing. We thank all members of IMB, particularly H.-J. Cheng, J.-Y. Leu, S.-C. Cheng and S.-H. Chen, for tremendous help and support. This work was supported by Career Development Award and Multidisciplinary Health Cloud Research Program of Academia Sinica (AS-CDA-108-M03 and AS-PH-109-01-3), Career Development Award of National Health Research Institute, Taiwan (NHRI-EX109-10908BC) and Excellent Young Scholar Research Grants and Ta-You Wu Memorial Award of Ministry of Science and Technology, Taiwan (MOST 109-2628-B-001-014-MY1 and 108-2118-M-001-013-MY5).

Author contributions

H.-L.C., Y.-T.C., J.-Y.S., Y.-L.W. and C.-L.L. carried out experiments and analysis. H.-N.L., C.-H.A.Y., Y.-J.H., Y.-T.C. and C.-L.L. established the web server tool. Y.-T.H. supervised statistical analysis. H.-L.C., Y.-T.C., J.-Y.S. and C.-L.L. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

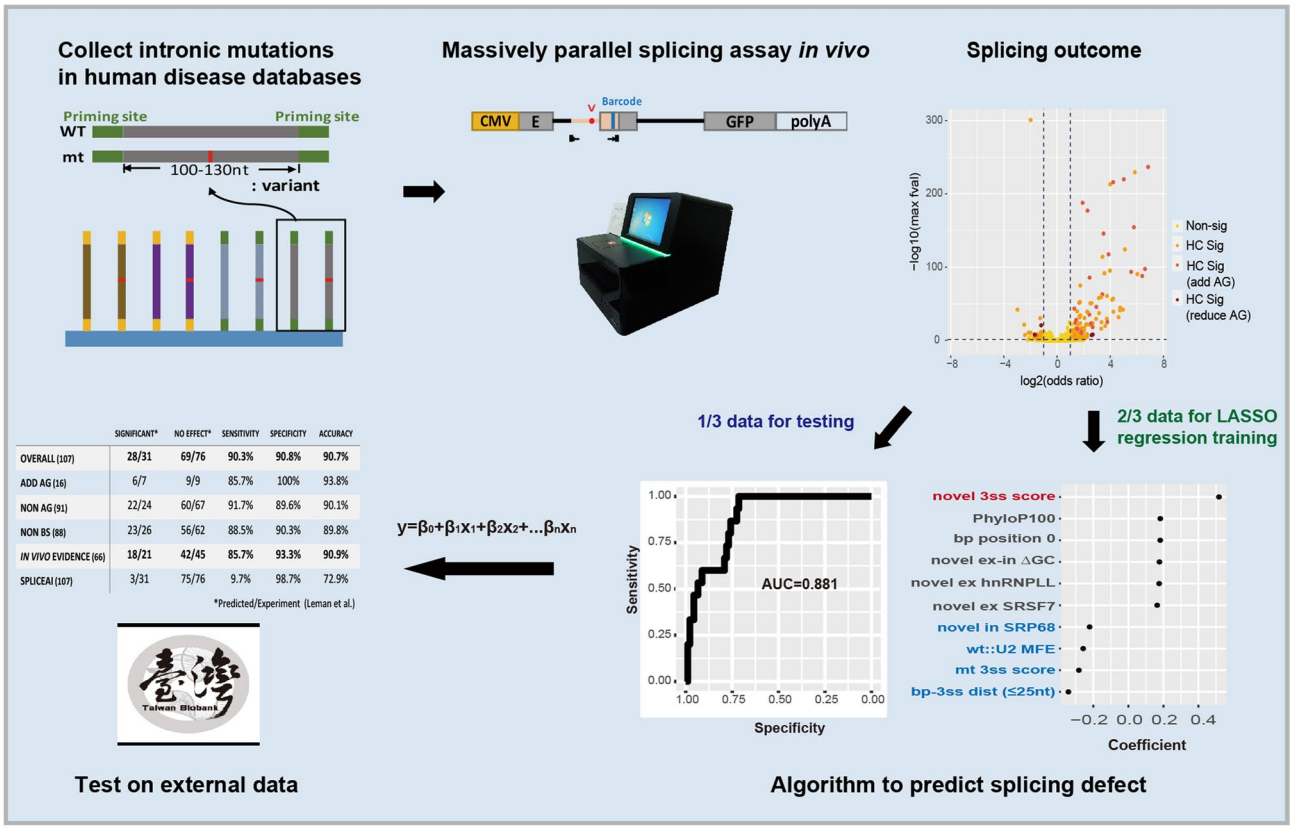
Extended data is available for this paper at https://doi.org/10.1038/s41594-022-00844-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41594-022-00844-1.

Correspondence and requests for materials should be addressed to Chien-Ling Lin.

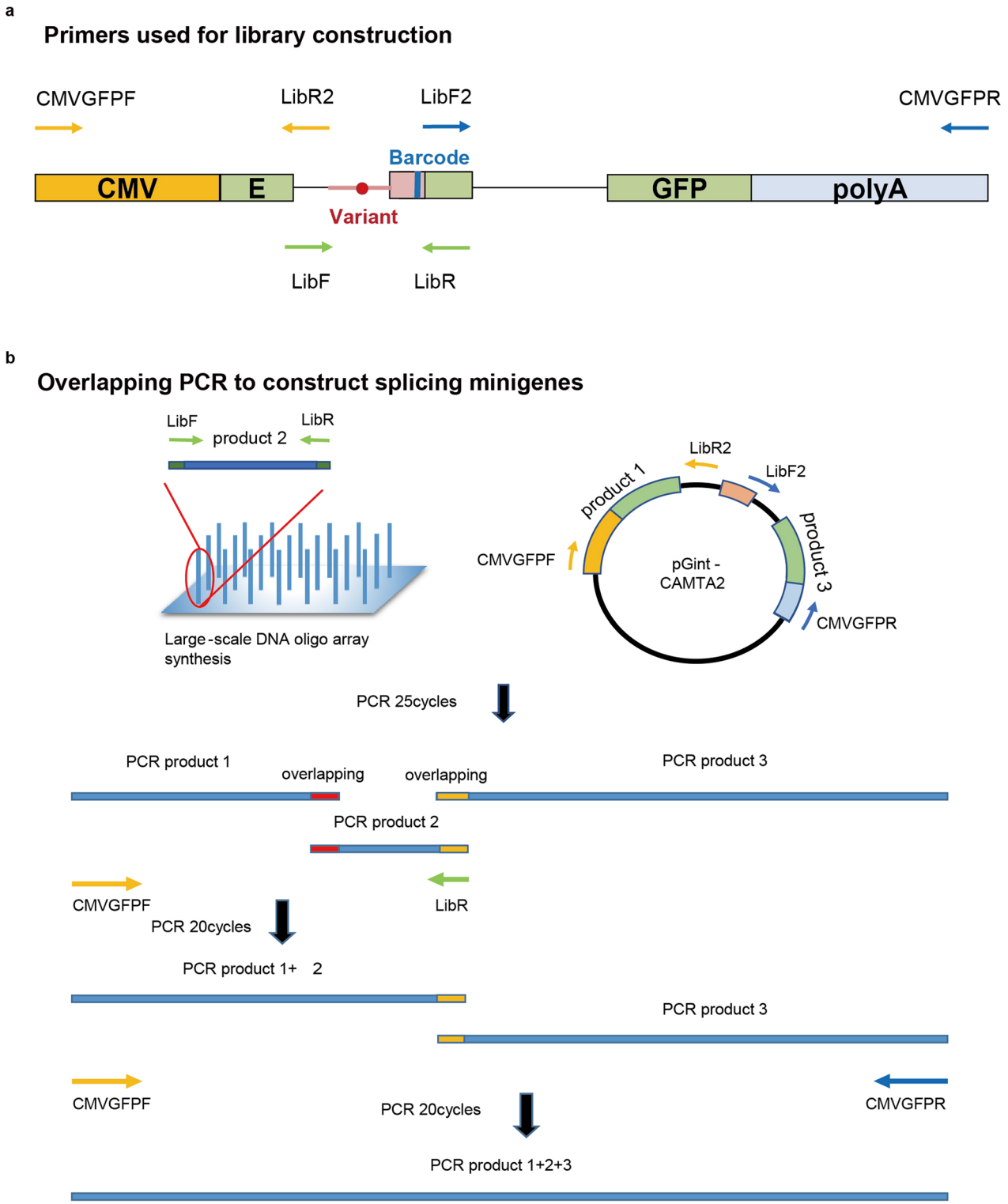
Peer review information *Nature Structural & Molecular Biology* thanks Ana Fiszbein and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Sara Osman was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

https://doi.org/10.1038/s41594-022-00844-1



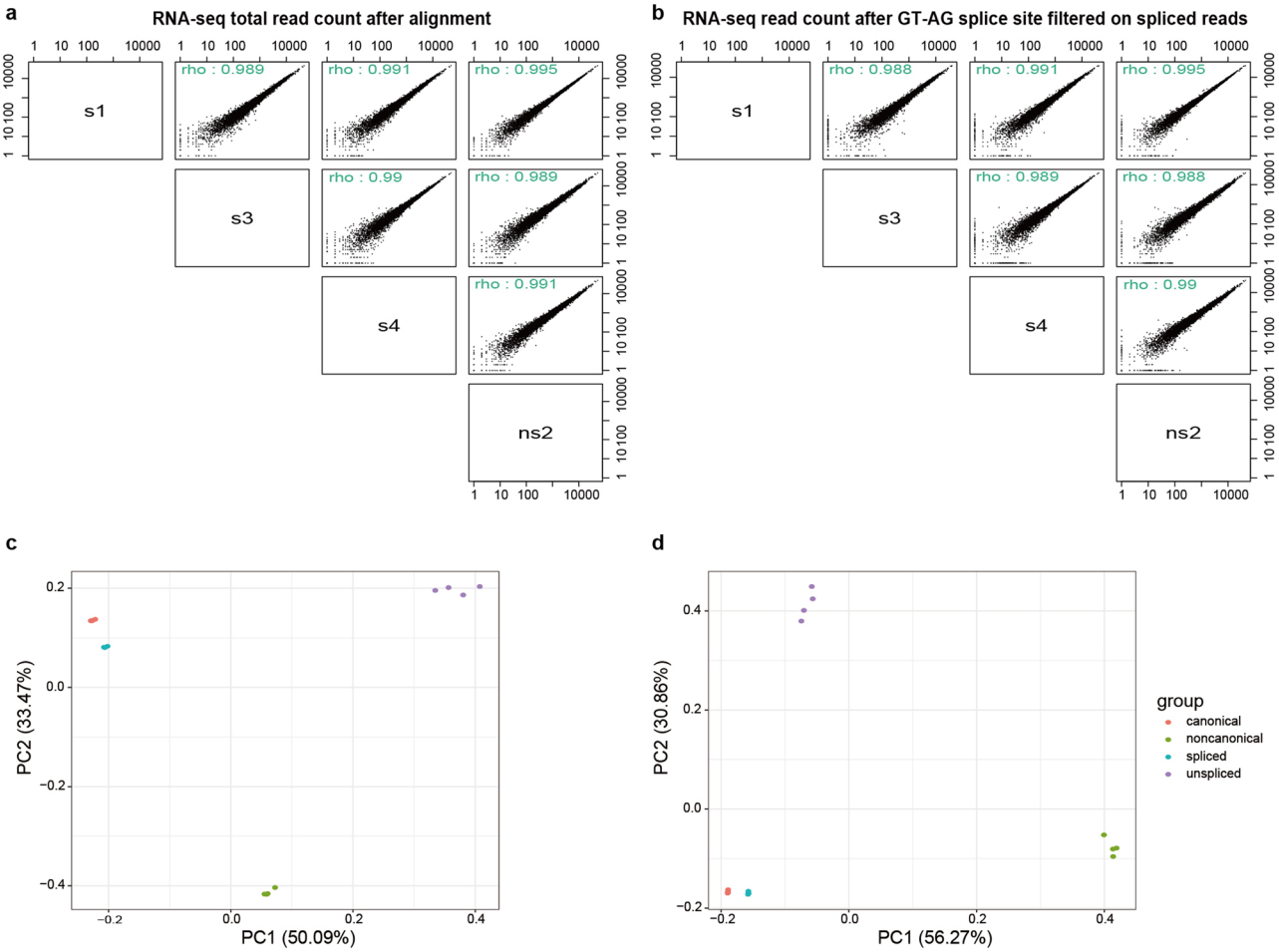
Extended Data Fig. 1 | Analysis flow of the study.

https://doi.org/10.1038/s41594-022-00844-1



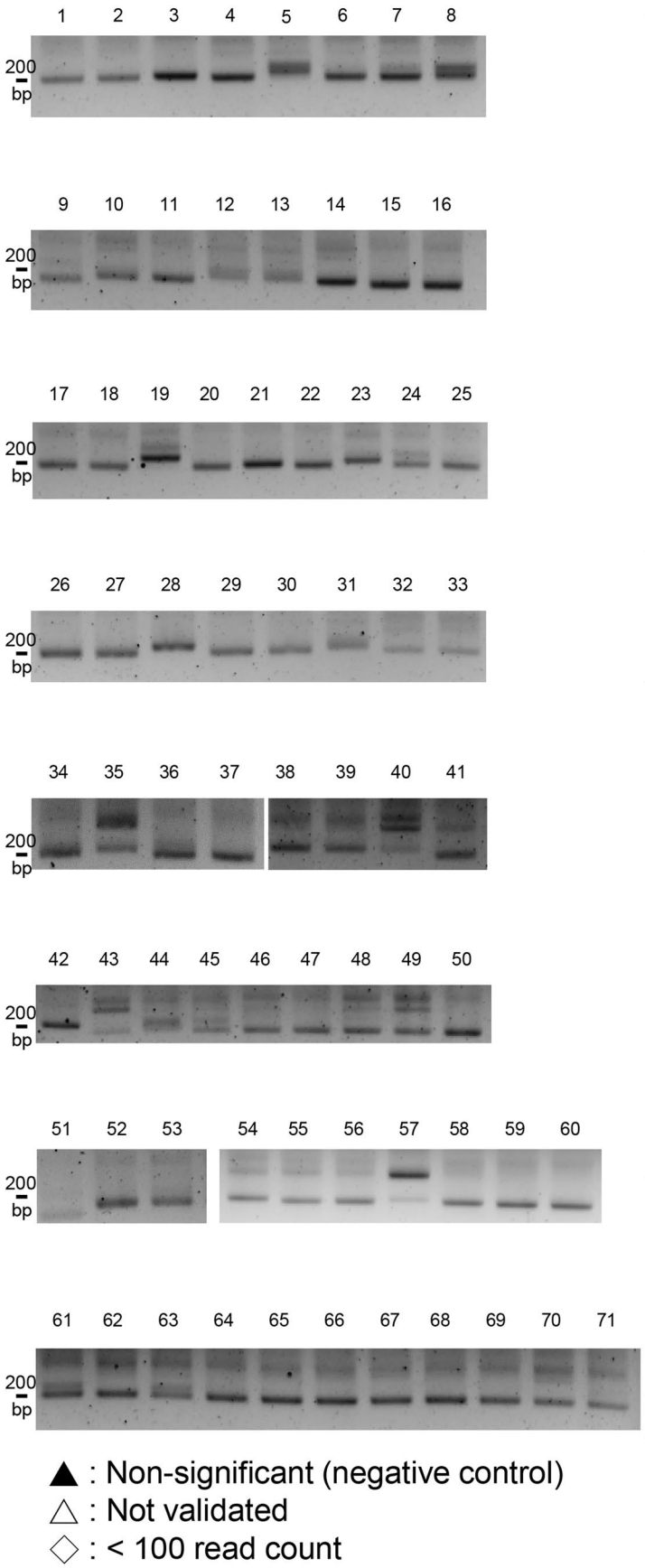
Extended Data Fig. 2 | Workflow of the library construction. (a) Primers used in the overlapping PCR. **(b)** The procedure of overlapping PCR. In brief, oligo pools and the other parts of the splicing minigenes were amplified by 25 PCR cycles. The fragment containing the promoter and the first exon (PCR product 1) was stitched to the oligo pool (PCR product 2) by overlapping PCR using 20 amplification cycles. Then, the stitched product (PCR product 1+2) was further stitched with the fragment containing the 3rd exon and polyadenylation signal (PCR product 3) using 20 amplification cycles to obtain the final construct (PCR product 1+2+3).

https://doi.org/10.1038/s41594-022-00844-1



Extended Data Fig. 3 | Massively parallel splicing assay (MaPSy) showing high consistency between repeats. (a) Spearman's correlation between 4 RNA-seq total read counts after alignment. **(b)** Spearman's correlation between 4 RNA-seq read count with spliced reads using GT-AG as splice sites. **(c)** Principle analysis for the spliced outcome of total RNA-seq read count. **(d)** Principle analysis for the spliced outcome of RNA-seq read count after GT-AG splice site filtration on spliced reads.

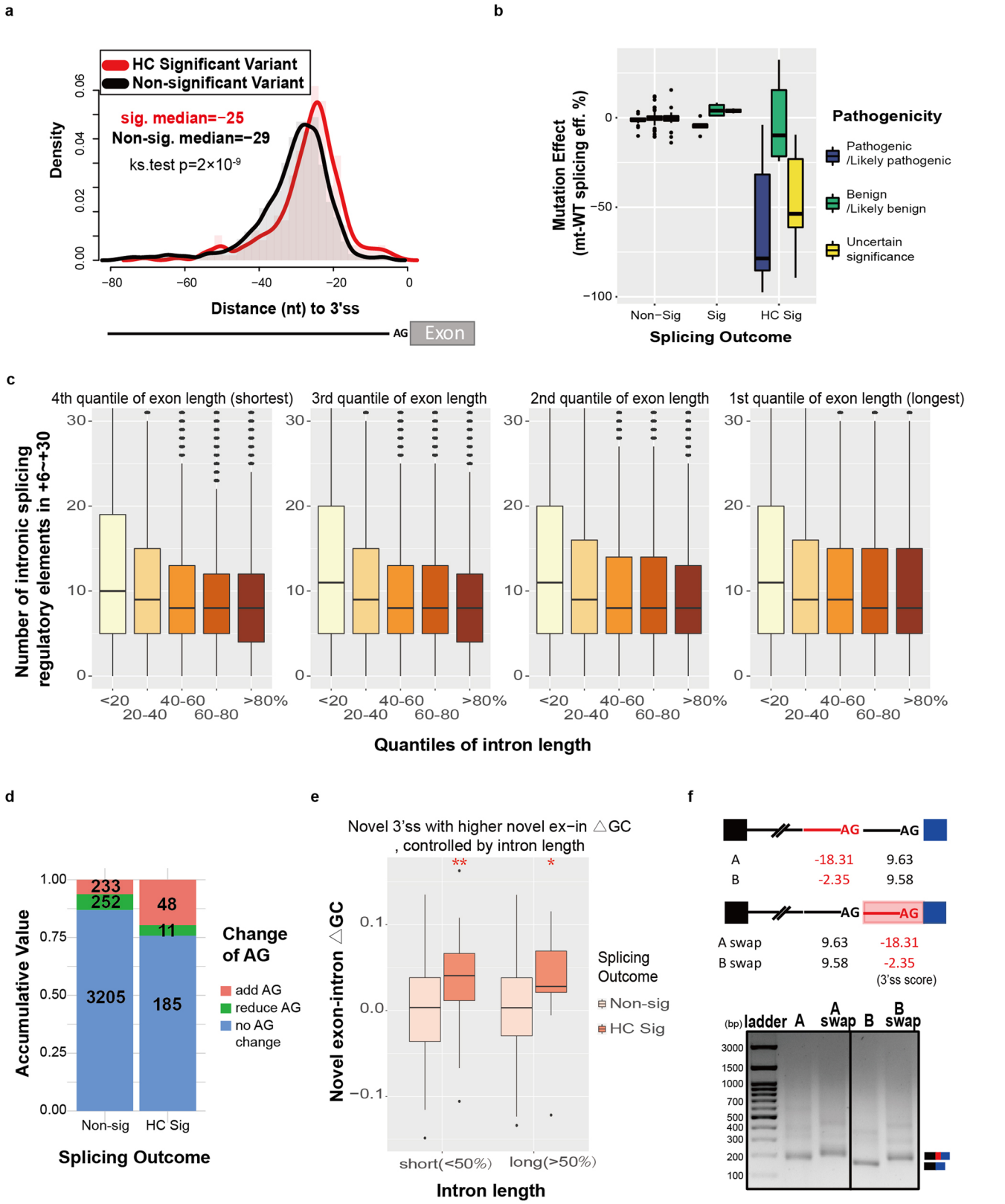
https://doi.org/10.1038/s41594-022-00844-1



▲ 1 : chr16:3026788-3026901:+_-21CtoG NM_024339
2 : chr16:3026788-3026901:+wt THOC6 intron8
3 : chr5:87333188-87333301:+_-9GtoA NM_002890
4 : chr5:87333188-87333301:+wt RASA1 intron3
5 : chr5:87349135-87349248:+_-19AtoT NM_002890
6 : chr5:87349135-87349248:+wt RASA1 intron7
7 : chr2:102732374-102732487:-_-24CtoT NM_032718
8 : chr2:102732374-102732487:-wt MFSD9 intron1
9 : chr4:521929-522042:+_-13CTTGTTtoC ENST00000506898
10 : chr4:521929-522042:+wt PIGG intron3
11 : chr7:142953921-142954034:-_-21GtoT NM_000420
12 : chr7:142953921-142954034:-_-25CTtoT KEL intron8
13 : chr7:142953921-142954034:-wt
14 : chr14:77285597-77285710:-_-14GtoA NM_013382
15 : chr14:77285597-77285710:-_-14GtoT POMT2 intron12
16 : chr14:77285597-77285710:-wt
▲ 17 : chr19:44865192-44865305:+_-33CtoG NM_002856
18 : chr19:44865192-44865305:+wt NECTIN2 intron1
19 : chr2:189007421-189007534:+_-43TtoG NM_000090
20 : chr2:189007421-189007534:+wt COL3A1 intron44
21 : chr3:33053514-33053627:-_-10CtoG NM_001079811
22 : chr3:33053514-33053627:-wt GLB1 intron6
23 : chr2:241740954-241741067:+_-23AtoG NM_152783
24 : chr2:241740954-241741067:+_-23AtoT D2HGDH intron2
25 : chr2:241740954-241741067:+wt
▲ 26 : chr2:68395601-68395714:+_-33GtoA NM_002664
27 : chr2:68395601-68395714:+wt PLEK intron8
28 : chr10:49473552-49473665:-_-26AtoG NM_001346440
29 : chr10:49473552-49473665:-wt ERCC6 intron13
30 : chr12:6577882-6577995:-_-16AtoT NM_001363606
31 : chr12:6577882-6577995:-wt CHD4 intron36
▲ 32 : chr19:38609990-38610103:-_-23CtoT NM_007181
33 : chr19:38609990-38610103:-wt MAP4K1 intron11
34 : chr1:19153904-19154017:-_-28TtoA NM_020765
35 : chr1:19153904-19154017:-wt UBR4 intron44
36 : chr3:48575867-48575980:-_-19AtoG NM_000094
37 : chr3:48575867-48575980:-wt COL7A1 intron71
38 : chr2:241224018-241224131:+_-33CtoT NM_001370694
39 : chr2:241224018-241224131:+wt ANO7 intron24
40 : chr9:12698372-12698485:+wt NM_000550
◇ 41 : chr9:12698372-12698485:+_-32CAGtoC TYRP1 intron3
◇ 42 : chrX:48512247-48512360:+_-46TtoA NM_203474
43 : chrX:48512247-48512360:+wt PORCN intron4
44 : chr1:27357841-27357954:-_-29CtoT NM_004672
45 : chr1:27357841-27357954:-wt MAP3K6 intron21
46 : chrX:101346625-101346738:-_-23AtoC NM_004085
47 : chrX:101346625-101346738:-wt TIMM8A intron1
48 : chrX:101354659-101354772:-_-23AtoC NM_001287345
49 : chrX:101354659-101354772:-_-23AtoG BTK intron13
50 : chrX:101354659-101354772:-wt
51 : chr4:102669014-102669127:-_-37AtoG NM_005908
▲ 52 : chr4:102669014-102669127:-_-40CtoA MANBA intron9
53 : chr4:102669014-102669127:-wt
54 : chr17:50196332-50196445:-_-9GtoA NM_000088
▲ 55 : chr17:50196332-50196445:-_-9GtoT COL1A1 intron13
56 : chr17:50196332-50196445:-wt
57 : chr15:43037848-43037961:-_-6TtoG NM_174916
58 : chr15:43037848-43037961:-wt UBR1 intron16
▲ 59 : chr2:58048748-58048861:+_-30AtoC NM_001288838
60 : chr2:58048748-58048861:+wt VRK2 intron3
61 : chrX:108601806-108601919:+_-18AtoG NM_033380
62 : chrX:108601806-108601919:+wt COL4A5 intron26
63 : chr16:87730609-87730722:-_-17TtoA NM_001184854
64 : chr16:87730609-87730722:-wt KLHDC4 intron3
△ 65 : chr3:39408527-39408640:+_-32CtoT NM_002295
66 : chr3:39408527-39408640:+wt RPSA intron2
▲ 67 : chr19:39383420-39383533:+_-39GtoC NM_001384577
68 : chr19:39383420-39383533:+wt SAMD4B intron12
▲ 69 : chr5:132486338-132486451:-_-64CtoT NM_002198
▲ 70 : chr5:132486338-132486451:-_-68AGtoG IRF1 intron6
71 : chr5:132486338-132486451:-wt

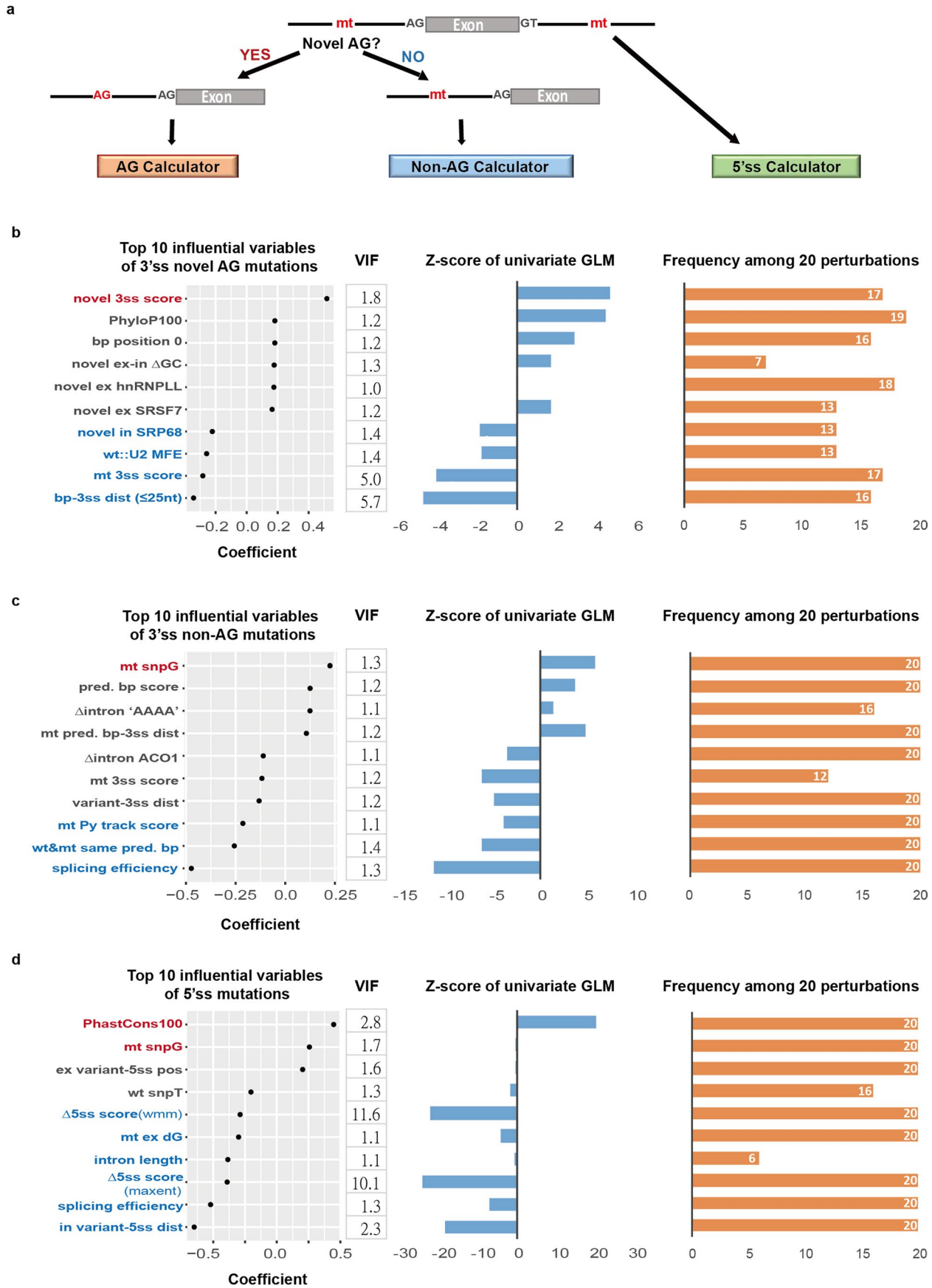
Extended Data Fig. 4 | Individual WT/mt pair validation of the Massively parallel splicing assay (MaPSy). Single minigene WT was transfected in HEK293T cells for splicing. The splicing outcome was examined by RT-PCR. Genomic coordinates, transcript ID, gene name and the corresponding introns were labeled accordingly. A representative experiment of three repeats is presented.

https://doi.org/10.1038/s41594-022-00844-1



Extended Data Fig. 5 | Characters of intronic 3′-end splicing variants. (a) Significant splicing variants are significantly closer to the 3′ss. Identity of two distributions was examined by two-sided Kolmogorov-Smirnov test. **(b)** Mutation effect on splicing efficiency of variants of various pathogenic levels. n = 247 variants. **(c)** Genome-wide association between the number of intronic splicing regulatory elements near the 5′ss and the intron length, stratified by exon length, related to Fig. 2f. n = 704,953 introns. **(d)** Enrichment of add-AG variants in the significant splicing variants. **(e)** Differential exon-intron GC content for non-significant and significant add-AG variants, stratified by intron length, related to Fig. 3f. n = 281 add AG variants. **: P-value of two-sided Wilcoxon test between the non-significant and HC significant variants 9×10^{-4}; * 0.01. The boxes in box plots represent medians (central line) and interquartile ranges (IQR; 25th to 75th percentile). The whiskers indicate ±1.5 × IQR from the box or the last data point within that and the dots show the outliers (b,c,e). **(f)** Preference of 3′ss with various 3′ss strengths was examined by 3′ss swapping assay with splicing minigenes. A representative experiment of three repeats is presented.

https://doi.org/10.1038/s41594-022-00844-1

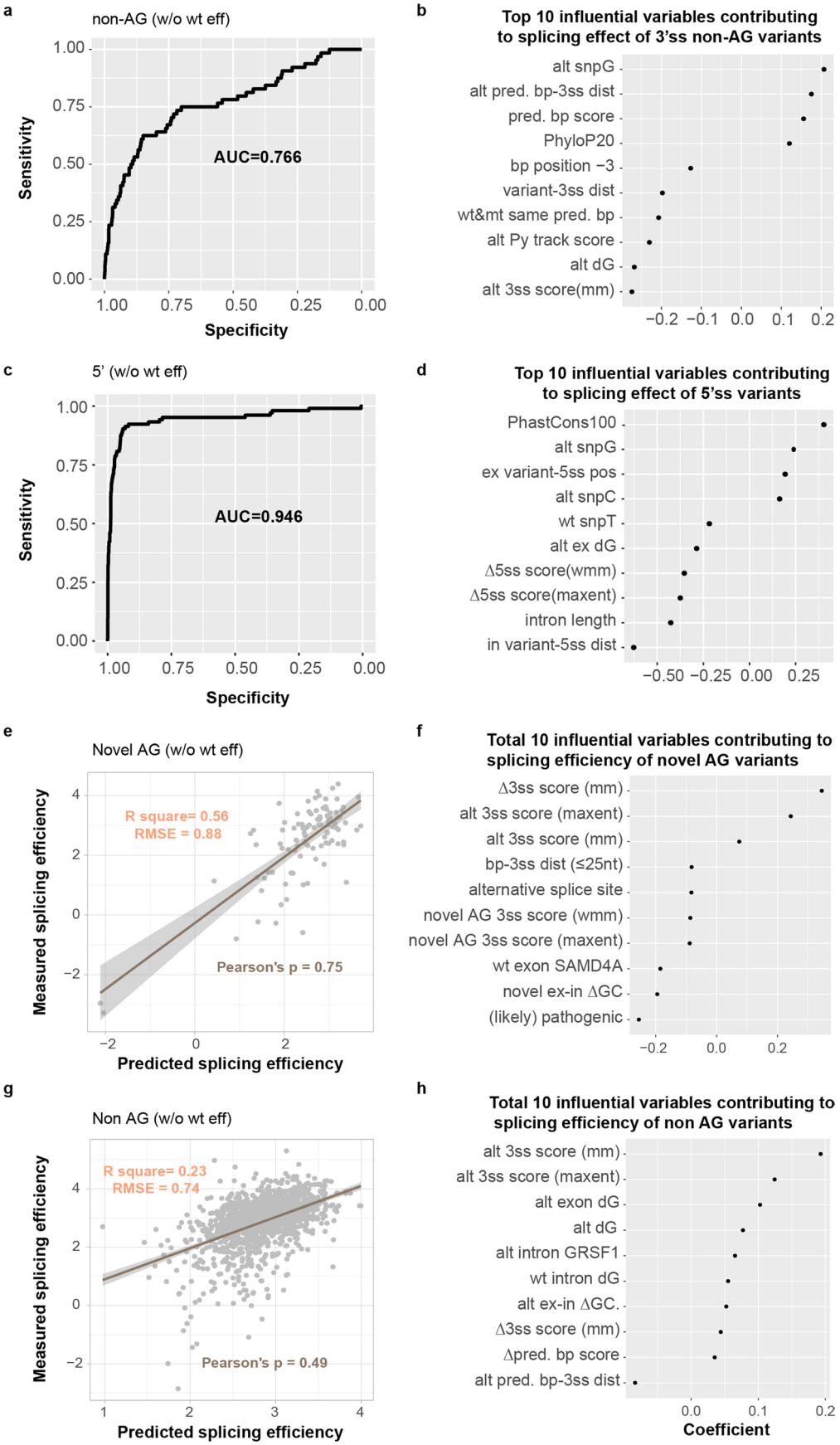


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41594-022-00844-1

Extended Data Fig. 6 | Sensitivity analysis of the generalized linear model that predicts splice-altering intronic mutations, related to Fig. 4. (a) Segregation of intronic mutations into two models based on intronic location and AG addition, same as Fig. 4a. **(b-d)** The left-most presents the top 10 contributory factors predicting mutations in each category that affect splicing, same as Fig. 4b, d and f. Right next to the factors is the 'variance inflation factor (VIF)' that examines the collinearity of variables. VIF smaller than 5 is an indication of independence of variables without a collinear effect. Z-score of univariate GLM in the middle column shows the size of the marginal influence of each variable (without other variables in the model). The right-most figure shows the consistency of variable selection with 20 different random selections of training data.

https://doi.org/10.1038/s41594-022-00844-1

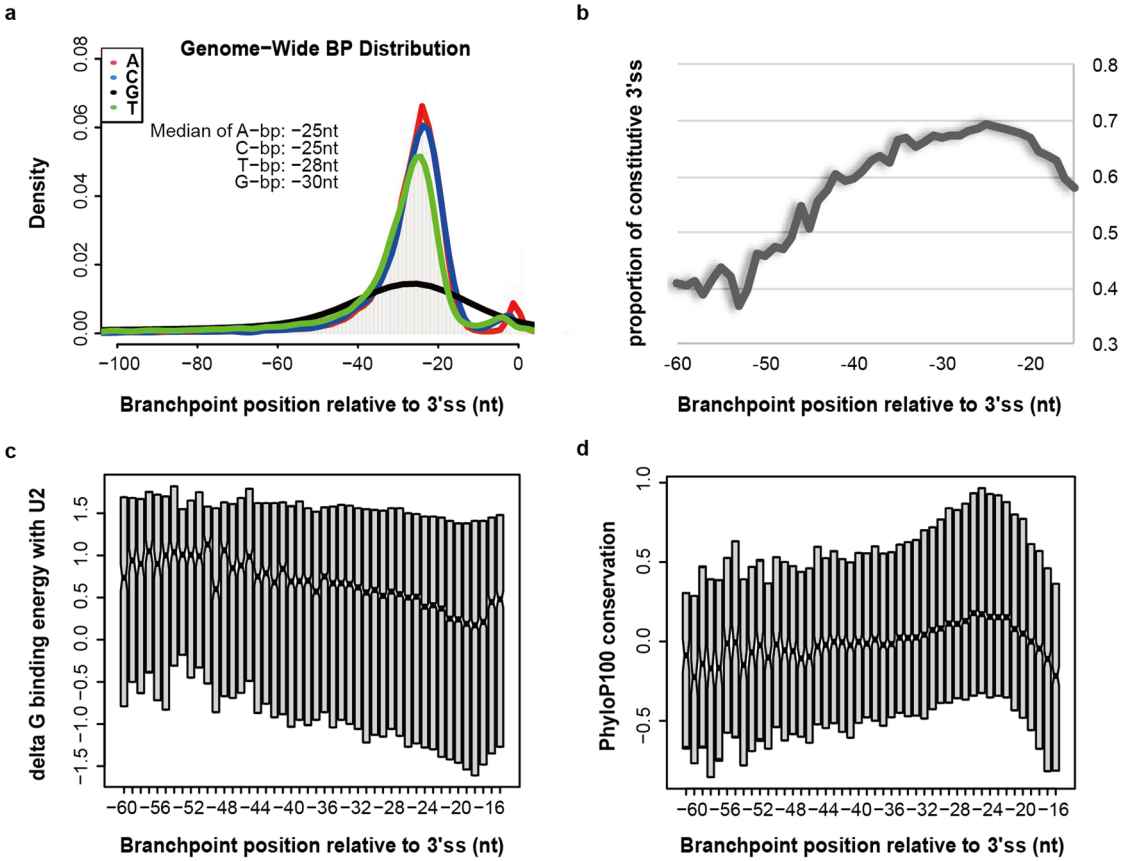


Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s41594-022-00844-1

Extended Data Fig. 7 | Generalized linear model to synthesize predictors of splicing altering non-AG-creating intronic mutations and splicing efficiency without WT splicing efficiency. (a) ROC curve of the non-AG model without WT splicing efficiency, similar to Fig. 4e. **(b)** Top 10 contributing factors to predict non-AG mutations affecting splicing without WT splicing efficiency, similar to Fig. 4d. **(c)** ROC curve of the 5′-end model without WT splicing efficiency, similar to Fig.4g. **(d)** Top 10 contributing factors to predict 5′-end mutations affecting splicing without WT splicing efficiency, similar to Fig. 4f. **(e,g)** Explanation power of each splicing efficiency model without WT splicing efficiency of (e) Novel AG mutations (g) non-AG mutations. The explanatory power of each model on the test dataset was estimated by Pearson's correlation (two-tailed), R square, and RMSE (root-mean-square error). The gray area displays the 95% confidence interval for predictions from the linear model. **(f,h)** All contributing factors to predict (f) novel AG mutations and (h) non-AG mutations affecting splicing without WT splicing efficiency. An 'alt' factor refers to a canonical property in the context of sequence variation. A '△ (delta)' factor refers to the difference of scores/motifs between the alt and WT sequence. A 'novel' factor refers to a new property associated with the novel 3′ss AG (f). More detailed descriptions of the factors can be found in Supplementary Table 2.

https://doi.org/10.1038/s41594-022-00844-1



Extended Data Fig. 8 | Intrinsic features of the BS and 3′ss regulate splice outcome. (a) Genome-wide distribution of A-, C-, T- and G-branchpoints (bp) relative to the 3′ss. **(b)** Proportion of branchpoints supporting constitutive splicing, sorted according to relative distance to the 3′ss. **(c)** Minimum free energy of BS pairing with the U2 BS recognition region, represented by boxplots for each position. **(d)** PhyloP100 conservation level of bp relative to the 3′ss.

nature portfolio

Corresponding author(s): Chien-Ling Lin

Last updated by author(s): 2022/05/16

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	no software was used.
Data analysis	software used in the manuscript: hisat2 (v 2.2.1), SAMtools (v 1.7), RegTools (v 0.5.2), MaxEntScan (10.1089/1066527041410418), SVM-BPfinder (10.1371/journal.pcbi.1001016), Vienna RNA package (v 2.4.8), Weblogo 3 (10.1101/gr.849004), SpliceAI (v 1.3.1), MMSplice (10.1101/438986), Python (3.7.11), SHAPEIT2 (v 2.r790), IMPUTE2 (v 2.3.1), Quality control of TWB data was performed using PLINK 1.9 and PLINK 2.0, R (v 4.1.1): {R package Biostrings (v 2.62.0), R package glmnet (v 4.1-4)} , R (v 4.2.0): {precrec (v 0.12.9), nnet (v 7.3-17)} codes are deposited at: https://github.com/chienlinglin/modeling-intron-variants/

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

Data

Policy information about availability of data

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

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

RNA Sequencing datasets generated during this study are available at the NCBI GEO: GSE179892.

Other databases used in the study:
UCSC PhyloP: https://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=1351580935_14MOQtNDW7V78RaXEDp3Yy4m4PTb&c=chr2&hgTracksConfigPage=configure&hgtgroup_compGeno_close=0#compGenoGroup
ATtRACT : https://attract.cnic.es/download
Ensembl: https://asia.ensembl.org/info/data/ftp/index.html

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	11,191 intronic mutations. First batch of 5,307 mutations were designed according to the limitation of batch synthesis. Another 5,884 mutations were retrieved from the published data (Cheung et al. 2019). With the estimation of 10% splicing altering variants among disease-relevant mutations, sample sizes to detect 90% signal (power=0.9) that determine splicing defects range between 885 to 3,050. Hence we reason that our sample sizes are sufficient to model the influential regulations.
Data exclusions	mutations below 100 reads were excluded
Replication	4 repeats. All attempts at replication were successful.
Randomization	training sets were randomly selected by the 'sample' function of R.
Blinding	calculation was automatically performed by the script. Therefore, blinding was not relevant to this study.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293T cells were purchased from ATCC (American Type Culture Collection).
Authentication	None of the cell line was authenticated.
Mycoplasma contamination	Cell lines were tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in the study.