CONSERVATION ASSESSMENT OF HUMAN SPLICE SITE ANNOTATION BASED ON A 470-GENOME ALIGNMENT

A PREPRINT

Ilia Minkin
Department of Biomedical Engineering
Center for Computational Biology
Johns Hopkins University, Baltimore, MD 21211, USA
ivminkin@gmail.com

Steven L. Salzberg
Department of Biomedical Engineering
Center for Computational Biology
Department of Computer Science
Department of Biostatistics
Johns Hopkins University, Baltimore, MD 21211, USA
salzberg@jhu.edu

ABSTRACT

Despite many improvements over the years, the annotation of the human genome remains imperfect, and different annotations of the human reference genome sometimes contradict one another. The use of evolutionarily conserved sequences provides a strategy for selecting a high-confidence subset of the annotation that is more likely to be related to biological functions, and the rapidly growing number of genomes from other species increases its power. Using the latest whole genome alignment, we found that splice sites from protein-coding genes in the high-quality MANE annotation are consistently conserved across more than 400 species. We also studied splice sites from the RefSeq, GENCODE, and CHESS databases that are not present in MANE. We trained a logistic regression classifier to distinguish between the conservation exhibited by sites from MANE versus sites chosen randomly from neutrally evolving sequence. We found that splice sites classified by our model as conserved have lower SNP rates and better transcriptomic support. We then computed a subset of transcripts only using either "conserved" splice sites or ones from MANE. This subset is enriched in high-confidence transcripts of the major gene catalogs that appear to be under purifying selection and are more likely to be correct and functionally relevant.

1 Introduction

The annotation of the human genome is a fundamental resource for a broad range of biomedical research and clinical applications. However, more than two decades after the initial publication of the genome itself, the scientific community have not reached a point where a consensus genome annotation is available [Amaral et al., 2023]. For example, one consequence is that the leading gene annotation databases for the human reference genome often disagree even on basic statistics such as the number of protein-coding genes [Pertea et al., 2018]. This is due to a variety of reasons, including the imperfect technologies used to assemble RNA transcripts and the noise inherent in the transcription process itself [Raj et al., 2006, Struhl, 2007, Cavallaro et al., 2021].

One of the challenging aspects of constructing a genome annotation is correctly determining the positions of introns inside the genes. The existence of introns and the mechanism of alternative splicing, first proposed by Gilbert [Gilbert, 1978], are critical for the functioning of cells. At the same time, the evolutionary origin of introns has been subject of a scientific debate for decades [Doolittle, 1978, Lynch and Richardson, 2002, Rogozin et al., 2012, Irimia and Roy, 2014].

A recent effort to address the challenge of the discrepancy between different human annotations resulted in the creation of a limited, high-quality gene annotation database called MANE [Morales et al., 2022]. This annotation was intended to include a single representative transcript for each protein-coding gene that has identical exon and intron structures in both RefSeq and GENCODE, two of the leading human annotation databases. The transcripts in MANE are chosen based on criteria that include expression levels and evolutionary conservation, which is a strong predictor of biological function. This approach yielded a subset of the human transcriptome that is under strong purifying selection. Although some studies have performed annotation-wide analysis of evolutionary consistency for human annotation [Fong et al., 2013], the conservation status of the rest of the annotations is less clear.

In this study, we address this question by studying the conservation of splice sites in major gene catalogs, both across multiple species and population level. First, we observed that the canonical dinucleotides GT/AG that flank introns are very highly conserved in protein-coding genes in MANE, with most of them being intact in more than 400 species in a multiple-genome alignment containing 470 mammalian species. We then investigated the patterns of conservation among splice sites that are not in MANE but that are present in one or more of the leading gene catalogs RefSeq, GENCODE, and CHESS. We found that while many of those splice sites closely follow the pattern of conservation found in MANE, others resemble randomly generated sites from neutrally evolving sequences.

To compare the properties of these two groups of splice sites, we developed a logistic regression model that classifies splice sites as either conserved or non-conserved. The model relies on a comparison of conservation patterns of splice sites from MANE to neutrally evolving sequences. As we detail below, we found that sites predicted as conserved by our classifier have lower rates of single nucleotide polymorphisms (SNPs) in the human population, are enriched in clinically relevant polymorphisms, and have better transcriptomic support. We then obtained a subset of transcripts from each major gene catalog for which all splice sites were either classified as conserved by our model or included

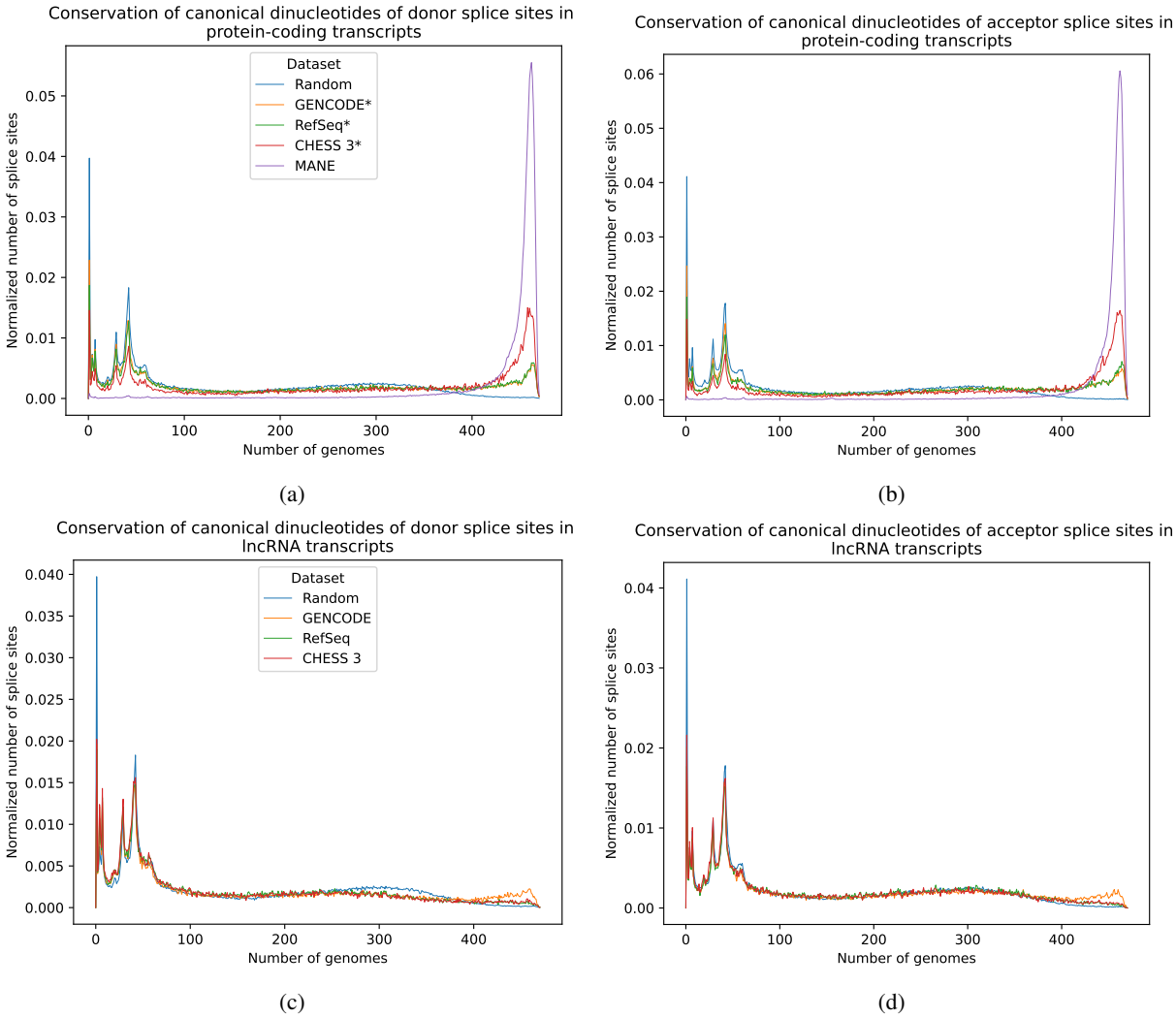


Figure 1: Distribution of the number of human splice sites with canonical dinucleotides (GT for donor and AG for acceptor sites) conserved in 470 mammals, computed for donor (a) and acceptor (b) sites of protein-coding genes, and donor (c) and acceptor (d) sites of lncRNAs. Each point shows a number of splice sites conserved (y-axis) in a given number of species (x-axis). Numbers are normalized by the total number of sites in the corresponding dataset in each category. The figure shows this statistic for annotations from GENCODE, RefSeq, CHESS 3 and MANE, as well as artificial splice sites ("Random") generated from internal sequences of introns which are assumed to evolve neutrally. For protein-coding genes, we created subsets GENCODE, RefSeq and CHESS 3 from which we removed MANE annotations, because each of these datasets are supersets of MANE; the resulting datasets are designated as GENCODE*, RefSeq* and CHESS 3* correspondingly.

in a transcript from MANE. These transcripts appear to be under strong purifying selection and are more likely to be functional and clinically relevant.

2 Results

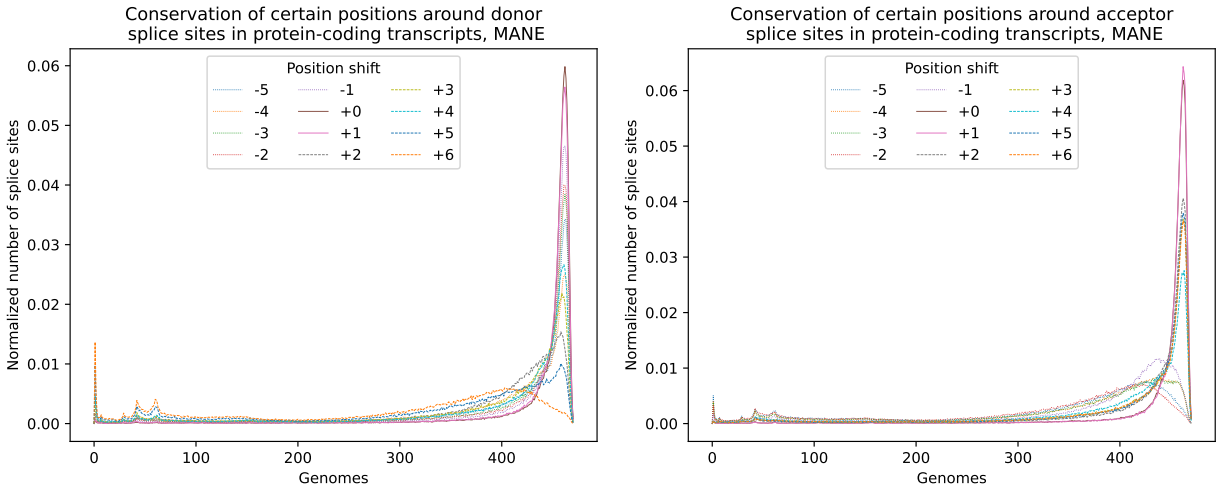
2.1 Exploratory data analysis of splice site conservation

First, we evaluated the evolutionary conversation of splice sites from four different human genome annotation databases: GENCODE [Frankish et al., 2021] version 45, RefSeq [O'Leary et al., 2016] release 111, CHESS 3 [Varabyou et al., 2023] v.3.0.1, and MANE [Morales et al., 2022] v1.3. GENCODE, RefSeq, and CHESS all contain every gene and transcript in MANE, which was created by GENCODE and RefSeq scientists with the goal of providing a single high-confidence transcript for every human protein-coding gene. To take into account this confounding factor and observe the differences between annotations more clearly, we removed the MANE splice sites from each of the other catalogs and created reduced versions that we designate as GENCODE*, RefSeq*, and CHESS 3* respectively. This procedure only affected protein-coding genes, because MANE does not currently contain noncoding genes or other types of annotation. Table 1 shows the numbers of donor and acceptor sites in each dataset, as well as the total number of transcripts.

For every donor and acceptor splice site in the databases, we computed how many species preserve the consensus dinucleotides (GT and AG) that appear at the beginning and end of most introns. We used a 470-species alignment available at the UCSC Genome Browse website [Kent et al., 2002] generated using MultiZ whole-genome aligner [Blanchette et al., 2004] to assess this conservation. In addition, we created a set of "false" splice sites intended to capture a baseline of neutrally evolving sequences. This annotation consists of 180,000 GT and AG dinucleotide sites randomly located within human intronic regions outside of splice site consensus motifs.

Figure 1 shows the pattern of conservation across species for each of these sets of donor and acceptor sites. First, we note that splice sites from protein-coding genes in MANE yield a plot that is clearly distinct from the other gene catalogs: most of the sites from MANE are conserved in >400 species. Second, protein-coding splice sites from the other datasets (after removing the MANE splice sites) seem to fall into two distinct categories: (1) MANE-like, and (2) neutral-like conservation. In contrast, lncRNAs from all datasets have very similar distributions and all of them closely follow the conservation pattern of random GT/AG sites. Both donor and acceptor splice show similar patterns of conservation. We note that randomly generated sites along with lncRNAs and some sites from coding genes exhibit several peaks in conservation in fewer than 50 species. We also calculated the most common species in which these splice sites are conserved, represented in Table A1. These species mostly constitute primates, which suggests that their conservation is merely a result of having a relatively recent common ancestor with humans. These splice sites may be clade-specific, or alternatively they might represent erroneous annotations. We also calculated the same statistic for splice sites that have non-canonical dinuclotides on the introns flanks. Figure A1 shows these numbers; they follow the same pattern as splice with the canonical dinucleotides.

Given the striking pattern of conservation of the canonical dinucleotides of splice sites from MANE, we investigated conservation of different positions around splice sites. Figure 2 shows the pattern of conservation of bases as a function



(a) (b)

Figure 2: Distribution of the number of donor (a) and acceptor (b) splice sites with a certain position around splice site motif conserved in a given number of species, for protein-coding genes from the MANE dataset. Each line represents conservation of a position either down- or upstream of the "canonical" dinuncleotides. For example, for donor splice sites +0 is usually "G", +1 is "T", and -1 is the first nucleotide upstream of the splice site. Numbers are normalized by the total number of sites in the corresponding dataset in each category.

	All splice sites		"Conserved" splice sites		All transcripts	"Conserved" transcripts
Protein-Coding						
MANE	183,012	182,973	-	-	18,264	-
GENCODE*	62,988	54,521	14,644	15,920	149,982	77,150
RefSeq*	45,720	37,408	11,377	11,308	137,189	46,687
CHESS 3*	50,850	45,863	23,593	23,046	85,584	40,110
lncRNA						
MANE	-	-	-	-	-	-
GENCODE	56,513	58,070	7,570	8,906	54,406	1,239
RefSeq	45,771	45,953	3,321	4,569	32,203	250
CHESS 3	48,134	48,539	3,577	4,688	35,716	358
Synthetic data						
Random	180,000	180,000	-	-	-	-

Table 1: Summary statistics of splice site conservation analysis. The second and the third columns represent the total number of sites in each datasets and the third and fourht columns show the number of splice classified as "conserved" by our model as per subsection 2.2. Last two columns indicate the total number of transcripts in each dataset and the number of transcripts that have all splice either from MANE dataset or classified as "conserved."

of their distance from the GT/AG splice site. As expected, the canonical dinucleotides (GT for donor sites and AG for acceptor sites) are the most conserved. On the other hand, upstream positions for donor sites and downstream positions for acceptor sites show similar patterns of conservation. However, downstream positions for donor sites and upstream ones for acceptor sites are much less conserved, which is expected because these positions are intronic.

We further explored the question of how well are splice sites conserved at the level of human population. Specifically, we calculated the fraction of splice sites having an SNP at a certain position, similarly to the cross-species conservation of different positions shown in Figure 2. To determine the presence of SNPs in the human population, we used the gnomAD database version 4.0.0 [Chen et al., 2023], focusing on loci that have at least one homozygous sample since a

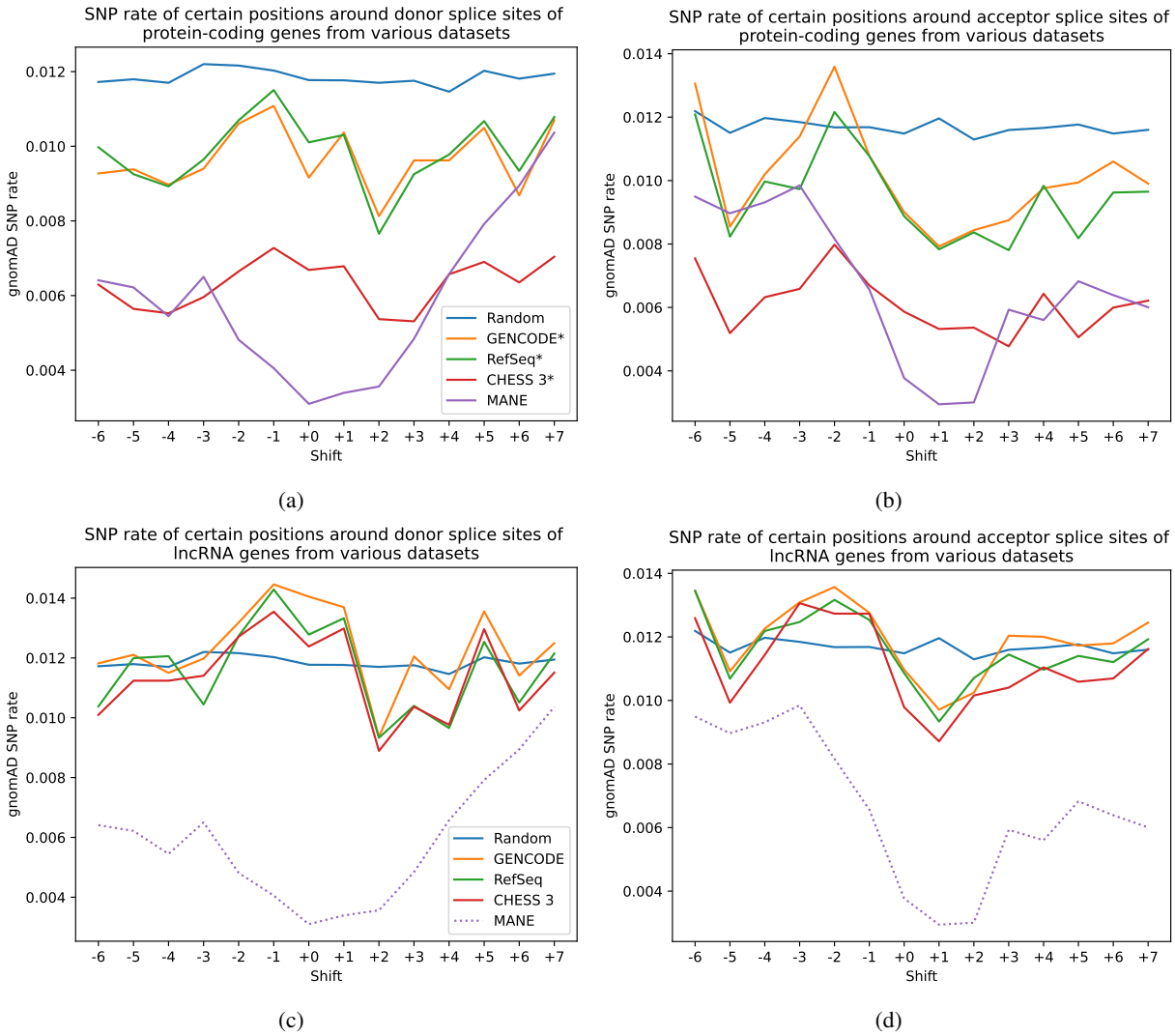


Figure 3: Rate of SNPs at positions near splice sites. Each point represents a proportion of splice sites from a certain dataset that have an SNP from the gnomAD dataset at a position either down- or upstream of the "canonical" dinuncleotides. For example, for donor splice sites +0 is usually "G", +1 is "T", and -1 is the first nucleotide upstream of the splice site. We only considered SNPs that have at least one homozygous sample. Panels (a, b) show donor and acceptor sites of protein-coding genes, while (c, d) show values for donor and acceptor sites of lncRNAs. For lncRNAs, we included MANE sites from protein coding genes as a baseline for splice sites under strong selection as MANE does not contain lncRNAs yet.

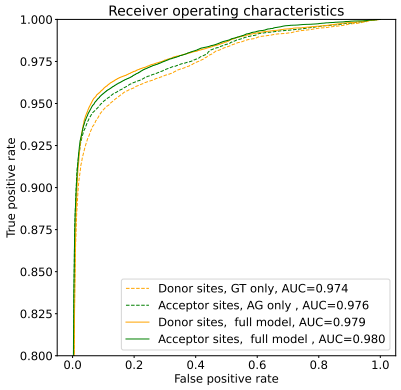


Figure 4: Receiving operating characteristic (ROC) curve for logistic regression models trained to distinguish between GT/AG sites chosen at random from intronic sequences and splice sites from MANE. Dashed lines correspond to the models using only the conservation of the canonical dinucleotides themselves, while solid ones show the performance of the model taking into account the conservation of the positions in a larger window.

homozygous SNP at a splice site is very likely to cause incorrect splicing. Figure 3 shows these fractions, which we call "SNP rates," calculated for each of the different gene catalogs. As expected, for protein-coding genes and their donor and acceptor splice sites, MANE has a much lower fraction of SNP sites at the canonical dinucleotides compared to random GT/AT positions, 0.4% versus 1.2%. On the other hand, splice sites in GENCODE* and RefSeq* have only slightly lower SNP rates than randomly evolving sequences; CHESS 3*'s rate is closer to MANE, but still somewhat higher. For lncRNAs from all of the catalogs, we observed that the SNP rates are relatively close to those of neutrally evolving sequence.

Our analysis suggests that some protein-coding splice sites and many more lncRNA splice sites include a subset of sites that are not under strong purifying selection. Otherwise, their average SNP rates should have been more similar to what we observed in splice sites from MANE. These sites might potentially represent non-functional and/or erroneous annotations.

2.2 Classifying splice sites based on their conservation

Above we showed that splice sites from the major gene catalogs exhibit two clearly distinct patterns of conservation: MANE-like and random-like. For brevity, we refer to the former as "conserved" and the latter as "non-conserved." We next decided to classify splice sites based on their conservation across species, and to compare their properties to see whether non-conserved sites might be misannotated. To do so, we trained a binary classifier based on logistic regression that uses the number of species in which a certain position around a splice site is conserved; we trained models for donor and acceptor sites separately. We used the randomly generated sites as negative examples and the whole MANE database as positive ones, with 20% of the data set aside for testing; the Methods section contains a detailed description of the model.

Model	Recall	Precision	F-score
Canonical dinucleotides only			
Donor sites	0.950	0.919	0.934
Acceptor sites	0.943	0.909	0.926
Full model			
Donor sites	0.946	0.937	0.942
Acceptor sites	0.950	0.942	0.946

Table 2: Precision and recall of models classifying splice sites as either conserved or non-conserved on the test dataset containing. The first two rows show the performance of the models taking into account the canonical dinucleotides GT/AG only, while the next two rows correspond to the model considering a larger window. The classification threshold for probability is 0.5. Donor (acceptor) test set consisted of 36,000 false examples, randomly generated from intronic sequences, and 36,000 positive examples from MANE.

Figure 4 shows the receiving operating characteristic (ROC) curve illustrating the tradeoff between true positive and false positive rates for these models on the test data. We evaluated two types of models: one using only the conservation of the canonical dinuclotides GT/AG themselves, and the other one using the conservation of the other positions in the splicing motif (see the Methods section for more details). Both models show high accuracy on the test data with with area under the ROC curve (AUROC) measuring 0.975-0.98 for donor and acceptor sites. However, the full model has a slightly lower false positive rate for a given classification threshold, hence we chose it for further analyses. For classification, we used a threshold of 0.5 for the probability predicted by the regression model to classify sites as conserved and non-conserved. Given this threshold, the full models for donor and acceptor sites have F-scores of 0.942 and 0.946 correspondingly, see Table 2.

In addition, we compared the probability output by the regression model to PhastCons [Siepel et al., 2005] scores indicating whether a particular position in genome is under negative selection. To do so, we used the PhastCons scores track available at the UCSC Genome Browser that were computed using the same 470-species alignment. For each site, we took a minimum of two PhastCons scores of the positions of their canonical dinucleotides (usually GT/AG) and computed Pearson correlation between this quantity and the probability predicted by the regression model; this resulted in 0.772 Pearson correlation value across all sites from the datasets under consideration.

We then applied the model to each dataset under consideration to label sites as conserved or non-conserved. Table 1 (columns 4 and 5) contains the number of donor and acceptor splice sites in each of the annotation databases classified by the model as conserved. For protein-coding genes, we observed that in GENCODE* and RefSeq*, only 23–24% of donor 29% of acceptor splices sites were conserved according to the model, while for CHESS 3*, the proportion was substantially higher, at 46% for donor sites and 50% for acceptor sites, suggesting the CHESS 3 has somewhat more reliable annotations of protein-coding transcripts. For lncRNAs, no more than 12% splice sites were classified as conserved across all datasets. Columns 6 and 7 of Table 1 also show the number total number of transcripts in each annotation and the number of "conserved transcripts" where each splice site is either shared with a transcript from MANE, or classified as "conserved" by our model. We observed that for protein-coding genes, GENCODE, RefSeq and CHESS 3 51%, 33% and 47% of transcripts have fully conserved splice sites The number of conserved lncRNAs is

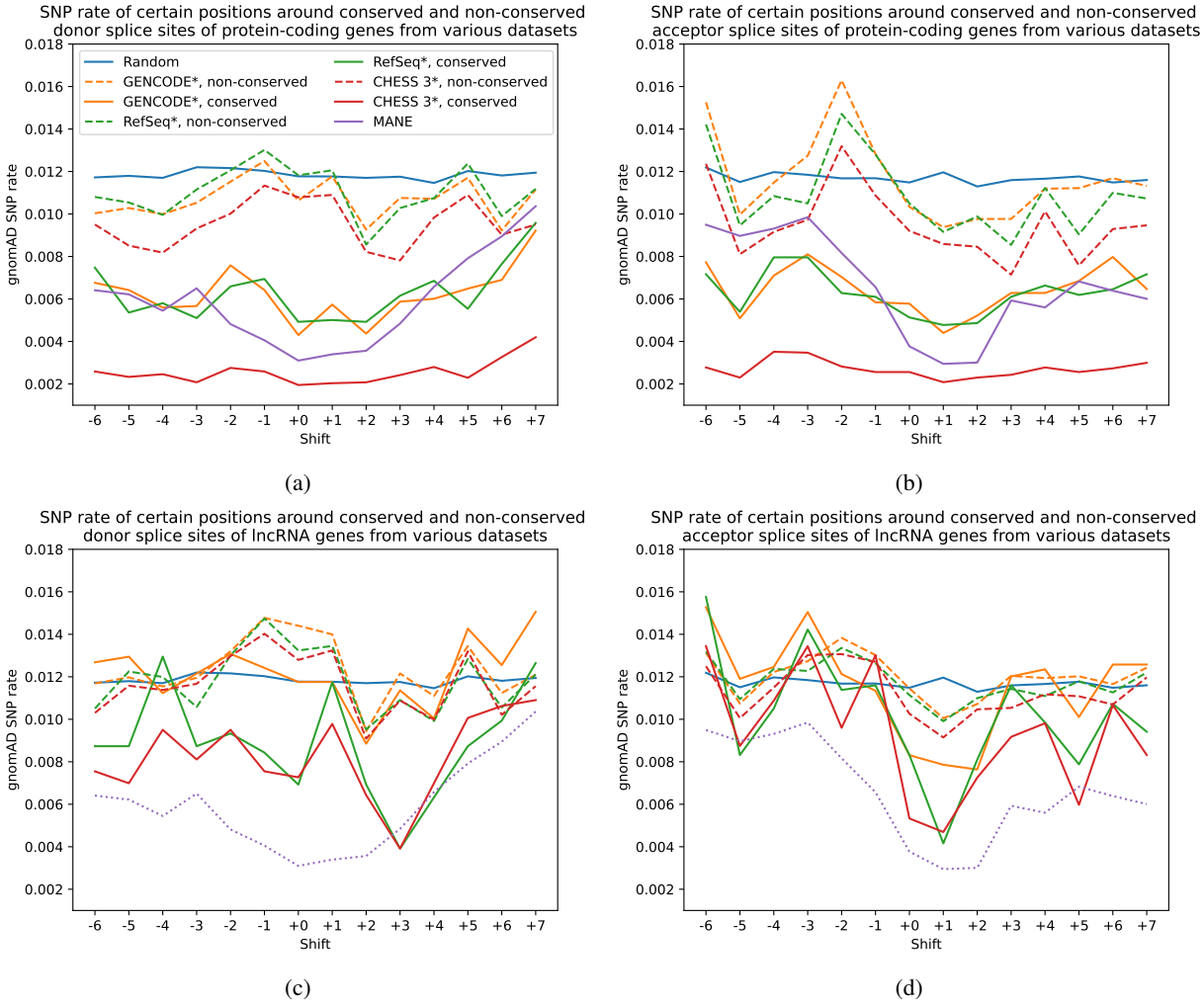


Figure 5: Rate of homozygous SNPs at positions near splice sites. Each point represents a proportion of splice from a certain dataset that have an SNP at a position either down- or upstream of the "canonical" dinuncleotides. For example, for donor splice sites +0 is usually "G", +1 is "T", and -1 is the first nucleotide upstream of the splice site. We only considered SNPs from the gnomAD database that have at least one homozygous sample. Panels (a, b) show donor and acceptor sites of protein-coding genes, while (c, d) show values for donor and acceptor sites of lncRNAs. Solid lines represent subsets classified as "conserved" by our model, while dashed ones correspond to "non-conserved" splice sites.

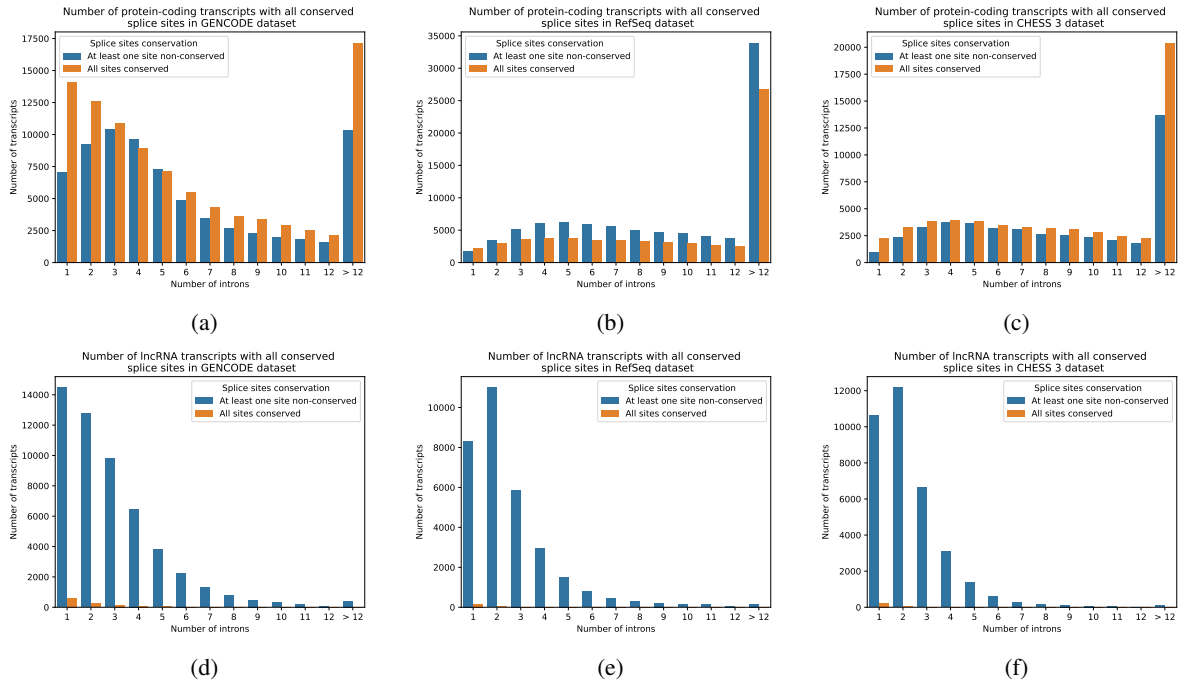


Figure 6: The number of protein-coding and lncRNA transcripts in GENCODE, RefSeq and CHESS 3 with a certain number of introns, split on whethere they all sites are conserved (as classified by our model or shared with transcript from MANE). Panels (a)-(c) show transcripts from protein-coding genes, while (d)-(f) represent lncRNAs.

much lower in each dataset: 0.8-1.0% for RefSeq and CHESS 3 and 2% for GENCODE. Figure 6 shows the number of conserved transcripts in each gene type and dataset split by the number of introns in a transcript.

We further compared SNP rates in the human population for conserved and non-conserved splice sites, again using the gnomAD human variation database and focusing on sites where at least one individual had a homozygous SNP. Figure 5 shows these rates for different datasets. For protein-coding genes, we observed that SNP rates for the canonical dinucleotides (positions +0 and +1) were 2–3 times lower for the conserved subset (as predicted by our classifier) compared to its non-conserved counterpart. We also note that the curves corresponding to non-conserved sites are closer to the Random (neutrally evolving) sites, while conserved sites in all three databases have SNP rates similar to MANE. However, for lncRNAs (Figure 5c-d) the separation is a little less clear: although the non-conserved sites have SNP rate pattern close to the Random ones, the conserved sites have only 30-50% smaller SNP rates at the canonical dinucleotides, and these rates are also much higher than the rates of protein-coding sites from MANE.

Apart from calculating the SNP rates, we compared the frequencies of homozygous SNPs overlapping the canonical dinucleotides of the splice sites classified by the model as either conserved or non-conserved. These frequencies for different datasets are displayed on the Figure 7. For donor and acceptor sites from protein-coding genes, median frequencies of conserved sites are 2-3 times smaller than for non-conserved ones. In addition, they 3-5 times smaller interquantile range. This is also true for donor sites of lncRNAs, but not their acceptor sites where frequencies distribution are closer.

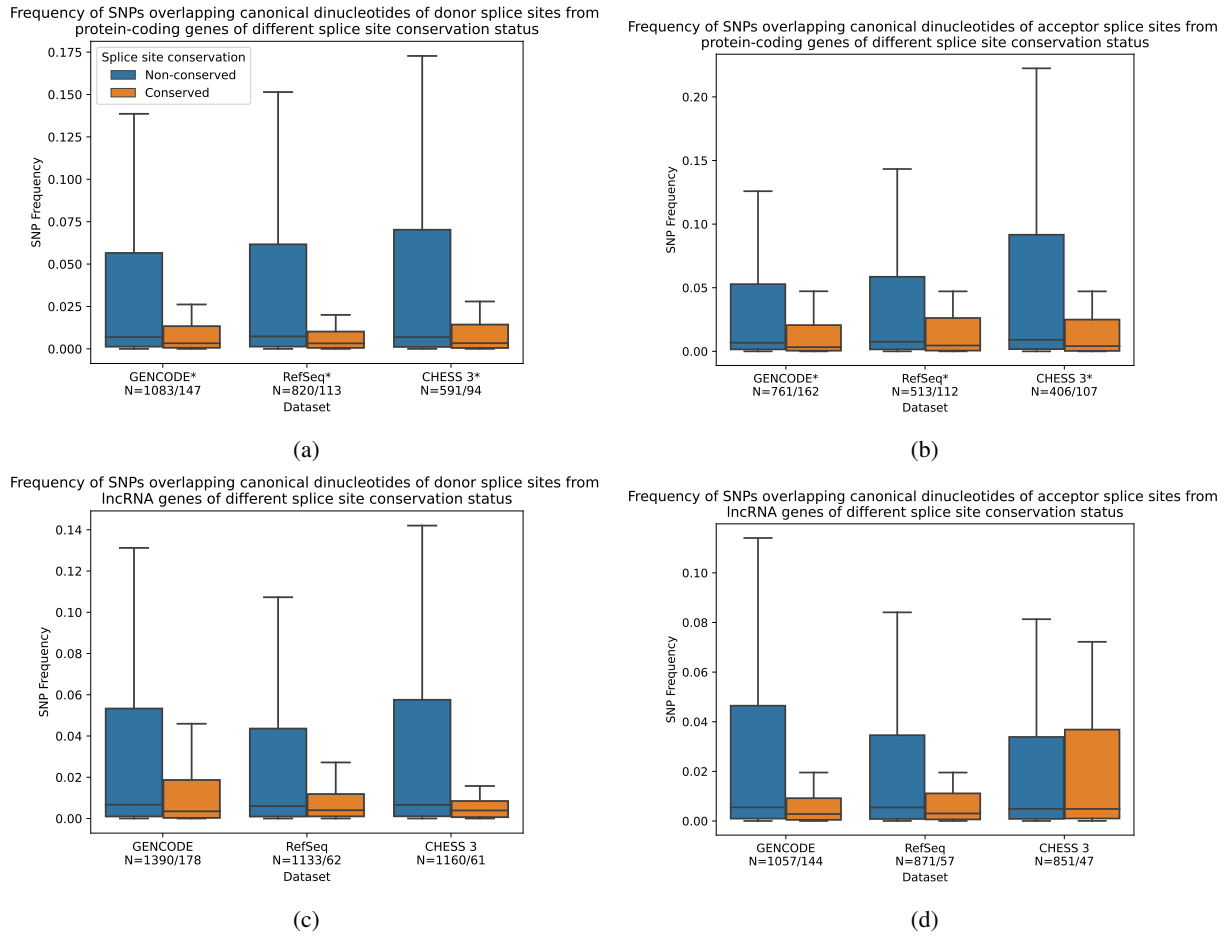


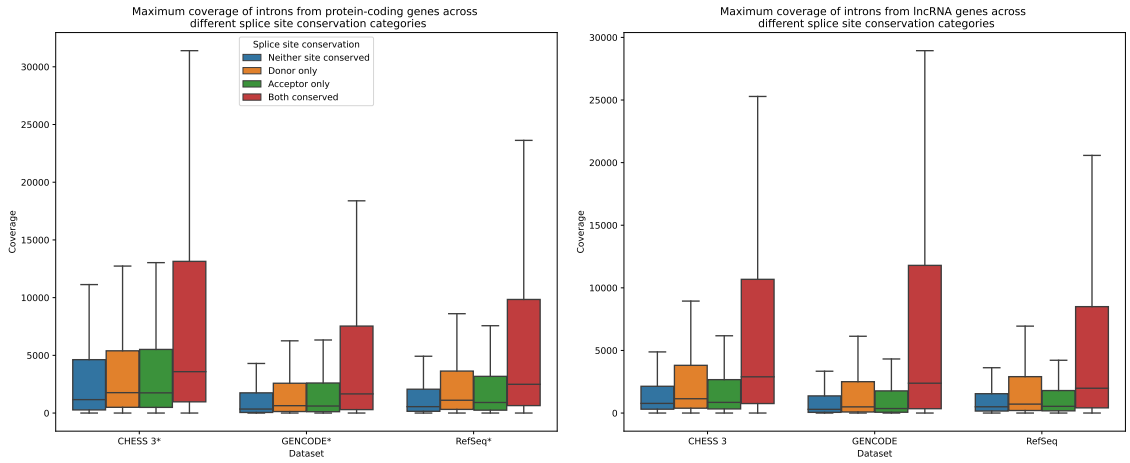
Figure 7: Box plots showing the distribution of SNP frequencies from the gnomAD dataset overlapping the canonical dinucleotides GT/AG of splice sites of different conservation status as determined by our model. Panel (a) and (b) show donor and acceptor sites from protein-coding genes, while (c) and (d) illustrate donor and acceptor sites from lncRNAs. Each pair of boxplots shows the distribution of frequencies of splice sites from a particular dataset and conservation categories; numbers before the dataset title show the number of SNPs overlapping splice sites in non-conserved/conserved categories. Each box plot shows the median (dashed red line), the interquartile range (solid top and bottom borders of the box), and maximum values within 95% percentile (whiskers), outliers are not shown.

We also examined how many splice sites have SNPs overlapping their canonical dinucleotides are associated with diseases. Table 3 shows the number and and the fraction of splice sites relative to their total number in the respective dataset that have at least one SNP from the ClinVar database [Landrum et al., 2014] classified as "pathogenic" or "likely pathogenic." Conserved splice sites of protein-coding genes have 2-4 higher ratio of potentially pathogenic SNPs than non-conserved ones across all datasets. Conserved sites of lncRNAs have consistently low ratio of pathogenic SNPs, while non-conserved ones from lncRNAs possess very few such variants.

We also examined the connection between multiple species conservation and gene expression using RNA-seq data. Our model classifies each acceptor and donor splice sites as either conserved or not. This way, each intron consisting of a pair of a donor and an acceptor splice sites belongs to either one of 4 categories: (1) neither site is conserved (2) only donor site is conserved (3) only acceptor side is conserved (4) boths sites are conserved. We employed data

	'Conserved" sites		"Non-conserved" sites	
Dataset	# Donor	# Acceptor	# Donor	# Acceptor
Protein-Coding				
GENCODE*	113 (0.77%)	129 (0.81%)	95 (0.20%)	61 (0.16%)
RefSeq*	69 (0.61%)	78 (0.69%)	58 (0.17%)	37 (0.14%)
CHESS 3*	66 (0.28%)	64 (0.28%)	47 (0.17%)	33 (0.14%)
lncRNA				
GENCODE	11 (0.15%)	10 (0.11%)	13 (0.03%)	3 (0.01%)
RefSeq	6 (0.18%)	6 (0.13%)	8 (0.02%)	4 (0.01%)
CHESS 3	8 (0.22%)	10 (0.21%)	8 (0.02%)	4 (0.01%)

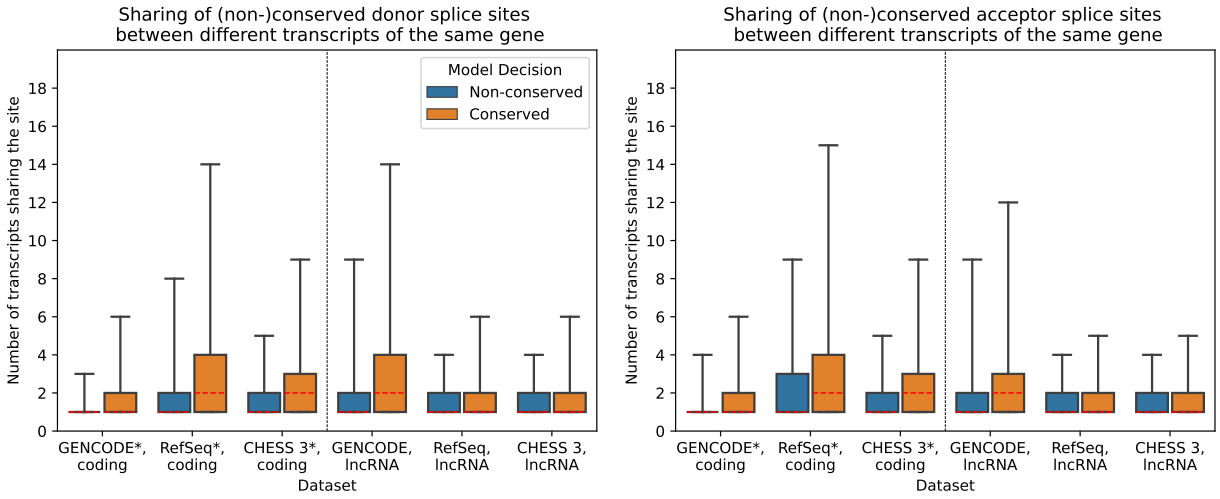
Table 3: Number of splice sites in each conservation category having an SNP classified as "pathogenic" and "likely pathogenic" from ClinVar database overlapping its canonical dinucleotides. Sites are classified as "conserved" or "non-conserved" by our model as per subsection 2.2. Numbers in parentheses show the percentage of the sites relative to the total number of sites in that category.



(a) (b)

Figure 8: Box plots showing the maximum coverage across all tissues of introns GENCODE*, RefSeq*, and CHESS 3* datasets. Figure (a) shows data for protein-coding genes, while the Figure(b) represents lncRNAs. Each box plot shows the median (dashed red line), the interquartile range (solid top and bottom borders of the box), and minimum and maximum values within ±1.5 of the interquartile range (whiskers), outliers are not shown.

from the GTEx project [Consortium et al., 2015] that was assembled using StringTie2 [Kovaka et al., 2019] and was postprocessed by TieBrush [Varabyou et al., 2021] to obtain the junction coverage. This data was used to generate the CHESS 3 gene catalog, further technical details on the pipeline can be found in [Varabyou et al., 2023]. For each intron we have the number of reads supporting the particular donor-acceptor junction reflecting expression of isoforms using this intron. To integrate the data, we calculated maximum coverage for each intron across all tissues; breakdown for each individual tissues is available in Appendix, figures A2 and A3. Figure 8 shows distributions of read coverage between introns of different conservation categories, for both protein-coding in lncRNA genes. As the figure shows, introns with both sides that are conserved have median max coverage that is 2-3 times higher than the introns that have at least one non-conserved site, which can we observed for both protein-coding and lncRNA genes. Figures A2 and A3 show coverage distribution across individual tissues that show the same pattern.



(a) (b)

Figure 9: Box plots showing the distribution of the number of transcripts sharing a certain donor (a) and acceptor (b) splice site, in GENCODE*, RefSeq*, and CHESS 3* datasets. The left part of each panel shows splice sites from protein-coding genes, while the right represents lncRNAs. Each box plot shows the median (dashed red line), the interquartile range (solid top and bottom borders of the box), and maximum values within 95% percentile (whiskers), outliers are not shown.

In addition, we explored how many isoforms of the same gene use conserved and non-conserved splice sites. In other words, for each splice site we computed the number of isoforms that use that particular site. Figure 9 shows the distribution of these values in each gene catalog, for both donor and acceptor sites in protein-coding genes and lncRNAs. For protein-coding genes, conserved splice sites are more likely to be shared between multiple isoforms. Notably, 71% of the non-conserved donor and acceptor sites from GENCODE* (leftmost plots in Figure 6a and 6b) were only used in a single isoform. However, we did not observe a similar pattern in lncRNAs, except for donor sites from the GENCODE annotation that also showed a notable difference between conserved and non-conserved sites. We note that lncRNA genes have fewer isoforms overall, which might explain some of the disparity between protein-coding genes and lncRNAs.

3 Discussion

In this study, we found that the canonical dinucleotides from both donor and acceptor splice sites of the consensus MANE dataset exhibit a striking pattern of conservation: nearly all of them are conserved in more than 400 mammalian species. In contrast, splice sites from the leading gene catalogs – GENCODE, RefSeq, and CHESS 3 – that are not shared with MANE exhibit two different patterns of conservation. The first pattern resembles MANE, where the splice sites are conserved in >400 species, while the second one resembles neutrally evolving sequences, at both the micro- and macroevolutionary levels. To compare the properties of these two groups of splice sites, we trained a logistic regression model using the MANE dataset as the source of positive examples and using randomly-chosen GT/AG sites from within introns to represent (albeit imperfectly) neutrally evolving sequences. We then applied this model to the

rest of the GENCODE, RefSeq, and CHESS gene catalogs excluding MANE to classify splice sites as either conserved or non-conserved.

We found that 30-50% of splice sites from coding gens and less than 12% splice sites from lncRNA can be classified as conserved. Splice sites classified as non-conserved had SNP rates in the human population that are consistent with neutrally evolving sequences, while conserved ones had patterns of SNP rates and frequencies similar to MANE. In addition, we observed that introns where both splice sites are conserved have better transcriptomic support. We also found that non-conserved splice sites are less likely to be shared by different isoforms of the same gene. We calculated the number of transcripts whose splice sites either classified as conserved by our model, or a shared with transcript from MANE. For protein-coding genes, 41%-56% belong to this category, only 0.5-2% have all their sites conserved. These transcripts can be used as a high-confidence subset of the gene catalogs we studied.

Our findings are consistent with the previous studies of splice site evolution. For example, it was observed before that genes that are highly conserved have higher expression levels [Green et al., 1993, Blencowe, 2006] and conserved exons are more likely to be included in multiple transcripts [Modrek and Lee, 2003]. Other studies [Kurmangaliyev et al., 2013, Shimada et al., 2010] found that splice sites that are not conserved in other species are more likely to carry disruptive SNPs in their motifs, and transcripts enriched in such variants have lower expression levels [Denisov et al., 2014]; it was also observed that splice sites of lncRNAs are less conserved than ones of mRNA [Washietl et al., 2014].

Some previous studies also found a lack of conservation of some splice sites using whole genome alignments [Sharma et al., 2016, 2017], but they assumed that these patterns arose due to alignment errors. Using the large gnomAD collection of human variation, we were able to examine SNP rates in the human population and show that splice sites that are poorly conserved across species also have higher SNP rates and frequiencies. This finding suggests that the majority of splice sites lacking conservation across species are simply not under selection in the human population as opposed to being poorly aligned.

This study has several limitations. First, unlike previous studies of evolutionary dynamics of alternative splicing [Barbosa-Morais et al., 2012, Merkin et al., 2012], we rely purely on the conservation of DNA sequences without taking into account whether a conserved splice site in the other genome is actually functional (information that is usually not known). Unfortunately, the incomplete status of many other genome annotations prevented us from incorporating them into our analysis. Second, we realize that the training data we used for our model could introduce biases. For example, the MANE dataset was constructed by choosing one “best” isoform per protein-coding gene, and conservation was one of the criteria. This could potentially contribute to the stronger conservation signal we observed in that data. However, In addition, randomly chosen GT/AG sequences from the interior of introns might not be the ideal choice for neutrally evolving sequences. These biases might potentially make our model less accurate at distinguishing between conserved and non-conserved splice sites.

We hope this study will help improve human genome annotation by demonstrating the utility of using large-scale evolutionary conservation for functional annotation of splicing. According to our analysis, highly-conserved splice

sites from MANE alone constitute at least 75% of all the splice in protein-coding isoforms in all genome annotations, and together with their conserved counterparts from the complimentary subsets account for 80-90% of all splice sites (Table 1). Hence, we believe that conservation should be an important factor in constructing a genome annotation. However, only a few methods currently use this information directly either for annotation or splice site prediction [Rose et al., 2011]. A common data structure used in RNA assembly called splice graph was generalized to integrate sequence homology information between species [Zea et al., 2021], but it was used for finding clusters of orthologous exons and yet to be employed for RNA-seq assembly. As higher-quality genomes along with their alignments become available, conservation-based methods have the potential to be a powerful aid in constructing functional annotations.

We also highlighted a subset of splice sites and corresponding isoforms in the leading human annotation catalogues that appear to be under strong selection. This subset can be used as a high-confidence representation of the annotation. At the same time, non-conserved splice might require further scrutiny since splicing could be inherently error-prone or be parts of non-functional isoforms [Hurst, 2009, Hsu and Hertel, 2009, Pickrell et al., 2010, Gout et al., 2013]. This hypothesis is backed up by the fact poorly conserved splice sites have higher SNP rates and frequencies in human population consistent with randomly select sites, which suggests that they are not under as strong negative selection. In addition, a recent analysis using proteomics analysis found that only 1 in 6 alternative isoforms were predicted to be functional [Pozo et al., 2021]. However, pinpointing exactly which splice sites are errors would require further study incorporating extra data. Here we have focused on the human genome because it has the highest-quality annotation, but in the future we hope to extend our analysis to the annotations of other species.

4 Methods

Our method for classifying splice sites is based on a logistic regression model designed to predict the probability of a splice site having a MANE-like conservation pattern (conserved), or a conservation pattern similar to a neutrally evolving sequence (not conserved). One of the primary features used by the regression model is the number of species in which the canonical dinucleotides are conserved, computed from a large multiple genome alignment. In addition, it takes into account an array positions surrounding a splice site, as they appear to have similar conservation properties. The training data includes randomly chosen GT/AG sites from intronic sequences as negative examples and the whole MANE dataset as positive examples. Below, we give the necessary initial definitions and describe the model.

We are given a collection of genomes $G = \{g_1, \ldots, g_m\}$, where each genome is a string $g_i = b_{i,1} \ldots b_{i,|g_i|}$ over the *nucleotides* of the DNA alphabet $\{A, C, G, T\}$, where $|g_i|$ is the length of the i-th genome. The genome g_1 is called the *reference*, and any non-reference genome $g_t, t > 1$ is called a *target* genome. For the reference genome, we are also given a *splice site annotation* represented as two sets, donor sites $D = \{d_1, \ldots, d_{|D|}\}$, and acceptor sites $A = \{a_1, \ldots, a_{|A|}\}$. The *origin* of the site is the position of the nucleotide of the first of the canonical dinucleotides, which we designate as $o(s)$ where s is either a donor or splice site. Thus for most donor sites $g_{1,o(d_q)} = G$ and $g_{1,o(d_q)+1} = T$ and for most acceptor sites $g_{1,o(a_r)} = A$ and $g_{1,o(a_r)+1} = G$.

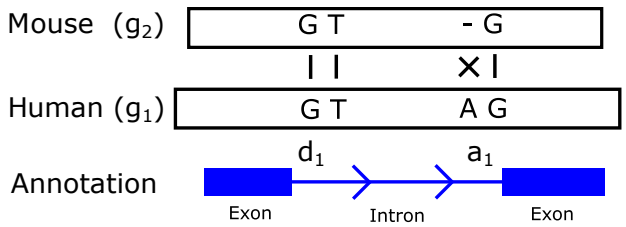


Figure 10: An example of mapping splice sites from a reference genome to another species using a whole-genome alignment. The donor site d_1 of the human reference genome g_1 has both its canonical dinucleotides intact in the target mouse genome, g_2. However, this is not true for the acceptor site a_1 which is mutated in mouse. In this example, the values of the conservation function for these two splice sites are $C(d_1, 0, 2) = C(d_1, 1, 2) = 1$, $C(a_1, 0, 2) = 0$, and $C(a_1, 1, 2) = 1$.

To find the corresponding sequence of each splice site of the reference in another species, we use a whole-genome alignment of m species. Formally, we define an alignment function $w(b_{1,i}, g_t)$ that maps each nucleotide $b_{1,i}$ of the reference genome to its homologous nucleotide $b_{t,j}$ of the target genome t included in the alignment if such nucleotide exists, otherwise, it maps $b_{1,i}$ to the character '-'. We also define the conservation function $C(s, \ell, t)$ as follows: it takes the value of 1 if the nucleotide with the shift ℓ of splice site s matches its homologous nucleotide in the genome t and 0 otherwise: $C(s, \ell, t) = I[b_{1,o(s)+\ell} = w(b_{1,o(s)+\ell}, t)]$. Figure 10 shows an example of mapping splice site sequence using whole-genome alignment and computation of the alignment and conservation function.

Our model consists of two types of variables to classify a splice site as conserved or non-conserved: (1) number of species in which the canonical dinucleotides are conserved *jointly* (2) number of species in which each nucleotide 30 position down- and up-stream of the canonical dinucleotide is conserved in, one variable per each position. This way, the log-odds of an acceptor site a_r being conserved are defined as:

$$\log(\frac{p(a_r)}{1 - p(a_r)}) = \alpha_0 + \alpha_1 \sum_{1<t\leq m} I[C(a_r, 0, t) = 1 \wedge C(a_r, 1, t) = 1] + \sum_{-30\leq\ell\leq 31,\, \ell\neq 0,1} \alpha_\ell \sum_{1<t\leq m} C(a_i, \ell, t)$$

Where α_0 is the interceptor term, α_1 corresponds to the conservation of canonical dinucleotides, and α_ℓ is the coefficient corresponding to the conservation of the position with the shift ℓ of the splice site. The log-odds of a donor splice being conserved are defined analogously. In addition, we evaluated a model taking into account the conservation of the canonical dinucleotides only to evaluate the contribution of the rest of the positions in the splicing motif. This way, we define the log-odds of probability p_0 of an acceptor splice site a_r being conserved as:

$$\log(\frac{p_0(a_r)}{1 - p_0(a_r)}) = \alpha_0 + \alpha_1 \sum_{1<t\leq m} I[C(a_r, 0, t) = 1 \wedge C(a_r, 1, t) = 1]$$

For fitting coefficients of the regression equations above, we used the logistic regression module from SciKit [Pedregosa et al., 2011]. To implement the alignment function $w(\cdot, \cdot)$, we utilized the AlignIO library from BioPython module [Cock et al., 2009] version 1.79 in conjunction with the whole-genome alignment of 470 mammal species available at the UCSC Genome Browser [Raney et al., 2024].

5 Data availability

The data and the code of the model is available via URL: ftp://ftp.ccb.jhu.edu/pub/iminkin2/splice-sites-pub-data.tar.gz.

6 Competing interests statement

The authors declare no competing interests.

7 Acknowledgements

We would like to thank Mihaela Pertea, Aleksey Zimin, Ales Varabyou, Beril Erdogdu, and Kuan-Hao Chao for useful discussions and suggestions. This work was supported in part by the U.S. National Institutes of Health under grants R01-HG006677 and R35-GM0130151.

References

Paulo Amaral, Silvia Carbonell-Sala, Francisco M De La Vega, Tiago Faial, Adam Frankish, Thomas Gingeras, Roderic Guigo, Jennifer L Harrow, Artemis G Hatzigeorgiou, Rory Johnson, et al. The status of the human gene catalogue. *Nature*, 622(7981):41–47, 2023.

Mihaela Pertea, Alaina Shumate, Geo Pertea, Ales Varabyou, Florian P Breitwieser, Yu-Chi Chang, Anil K Madugundu, Akhilesh Pandey, and Steven L Salzberg. Chess: a new human gene catalog curated from thousands of large-scale rna sequencing experiments reveals extensive transcriptional noise. *Genome biology*, 19(1):1–14, 2018.

Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mrna synthesis in mammalian cells. *PLoS biology*, 4(10):e309, 2006.

Kevin Struhl. Transcriptional noise and the fidelity of initiation by rna polymerase ii. *Nature structural & molecular biology*, 14(2):103–105, 2007.

Massimo Cavallaro, Mark D Walsh, Matt Jones, James Teahan, Simone Tiberi, Bärbel Finkenstädt, and Daniel Hebenstreit. 3'-5' crosstalk contributes to transcriptional bursting. *Genome biology*, 22:1–20, 2021.

W Gilbert. Why genes in pieces? *Nature*, 271(5645):501, 1978.

W Ford Doolittle. Genes in pieces: were they ever together? *Nature*, 272(5654):581–582, 1978.

Michael Lynch and Aaron O Richardson. The evolution of spliceosomal introns. *Current opinion in genetics & development*, 12(6):701–710, 2002.

Igor B Rogozin, Liran Carmel, Miklos Csuros, and Eugene V Koonin. Origin and evolution of spliceosomal introns. *Biology direct*, 7:1–28, 2012.

Manuel Irimia and Scott William Roy. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harbor perspectives in biology*, 6(6):a016071, 2014.

Joannella Morales, Shashikant Pujar, Jane E Loveland, Alex Astashyn, Ruth Bennett, Andrew Berry, Eric Cox, Claire Davidson, Olga Ermolaeva, Catherine M Farrell, et al. A joint ncbi and embl-ebi transcript set for clinical genomics and research. *Nature*, 604(7905):310–315, 2022.

Jessica H Fong, Terence D Murphy, and Kim D Pruitt. Comparison of refseq protein-coding regions in human and vertebrate genomes. *BMC genomics*, 14:1–16, 2013.

Adam Frankish, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Joel Armstrong, If Barnes, et al. Gencode 2021. *Nucleic acids research*, 49(D1):D916–D923, 2021.

Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

Ales Varabyou, Markus J Sommer, Beril Erdogdu, Ida Shinder, Ilia Minkin, Kuan-Hao Chao, Sukhwan Park, Jakob Heinz, Christopher Pockrandt, Alaina Shumate, et al. Chess 3: an improved, comprehensive catalog of human genes and transcripts based on large-scale expression data, phylogenetic analysis, and protein structure. *Genome biology*, 24(1):249, 2023.

W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arian FA Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708–715, 2004.

Siwei Chen, Laurent C Francioli, Julia K Goodrich, Ryan L Collins, Masahiro Kanai, Qingbo Wang, Jessica Alföldi, Nicholas A Watts, Christopher Vittal, Laura D Gauthier, et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*, pages 1–11, 2023.

Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, 2005.

Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(D1):D980–D985, 2014.

GTEx Consortium, Kristin G Ardlie, David S Deluca, Ayellet V Segrè, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

Sam Kovaka, Aleksey V Zimin, Geo M Pertea, Roham Razaghi, Steven L Salzberg, and Mihaela Pertea. Transcriptome assembly from long-read rna-seq alignments with stringtie2. *Genome biology*, 20:1–13, 2019.

Ales Varabyou, Geo Pertea, Christopher Pockrandt, and Mihaela Pertea. Tiebrush: an efficient method for aggregating and summarizing mapped reads across large datasets. *Bioinformatics*, 37(20):3650–3651, 2021.

Philip Green, David Lipman, LaDeana Hillier, Robert Waterston, David States, and Jean-Michel Claverie. Ancient conserved regions in new gene sequences and the protein databases. *Science*, 259(5102):1711–1716, 1993.

Benjamin J Blencowe. Alternative splicing: new insights from global analyses. *Cell*, 126(1):37–47, 2006.

Barmak Modrek and Christopher J Lee. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature genetics*, 34(2):177–180, 2003.

Yerbol Z Kurmangaliyev, Roman A Sutormin, Sergey A Naumenko, Georgii A Bazykin, and Mikhail S Gelfand. Functional implications of splicing polymorphisms in the human genome. *Human Molecular Genetics*, 22(17): 3449–3459, 2013.

Makoto K Shimada, Yosuke Hayakawa, Jun-ichi Takeda, Takashi Gojobori, and Tadashi Imanishi. A comprehensive survey of human polymorphisms at conserved splice dinucleotides and its evolutionary relationship with alternative splicing. *BMC Evolutionary Biology*, 10(1):1–12, 2010.

Stepan V Denisov, Georgii A Bazykin, Roman Sutormin, Alexander V Favorov, Andrey A Mironov, Mikhail S Gelfand, and Alexey S Kondrashov. Weak negative and positive selection and the drift load at splice sites. *Genome biology and evolution*, 6(6):1437–1447, 2014.

Stefan Washietl, Manolis Kellis, and Manuel Garber. Evolutionary dynamics and tissue specificity of human long noncoding rnas in six mammals. *Genome research*, 24(4):616–628, 2014.

Virag Sharma, Anas Elghafari, and Michael Hiller. Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. *Nucleic Acids Research*, 44(11):e103–e103, 03 2016. ISSN 0305-1048.

Virag Sharma, Peter Schwede, and Michael Hiller. CESAR 2.0 substantially improves speed and accuracy of comparative gene annotation. *Bioinformatics*, 33(24):3985–3987, 08 2017. ISSN 1367-4803.

Nuno L Barbosa-Morais, Manuel Irimia, Qun Pan, Hui Y Xiong, Serge Gueroussov, Leo J Lee, Valentina Slobodeniuc, Claudia Kutter, Stephen Watt, Recep Colak, et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593, 2012.

Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599, 2012.

Dominic Rose, Michael Hiller, Katharina Schutt, Jörg Hackermüller, Rolf Backofen, and Peter F Stadler. Computational discovery of human coding and non-coding transcripts with conserved splice sites. *Bioinformatics*, 27(14):1894–1900, 2011.

Diego Javier Zea, Sofya Laskina, Alexis Baudin, Hugues Richard, and Elodie Laine. Assessing conservation of alternative splicing with evolutionary splicing graphs. *Genome Research*, 31(8):1462–1473, 2021.

Laurence D Hurst. Evolutionary genomics and the reach of selection. *Journal of Biology*, 8:1–5, 2009.

Shu-Ning Hsu and Klemens J Hertel. Spliceosomes walk the line: splicing errors and their impact on cellular function. *RNA biology*, 6(5):526–530, 2009.

Joseph K Pickrell, Athma A Pai, Yoav Gilad, and Jonathan K Pritchard. Noisy splicing drives mrna isoform diversity in human cells. *PLoS genetics*, 6(12):e1001236, 2010.

Jean-François Gout, W Kelley Thomas, Zachary Smith, Kazufusa Okamoto, and Michael Lynch. Large-scale detection of in vivo transcription errors. *Proceedings of the National Academy of Sciences*, 110(46):18584–18589, 2013.

Fernando Pozo, Laura Martinez-Gomez, Thomas A Walsh, José Manuel Rodriguez, Tomas Di Domenico, Federico Abascal, Jesús Vazquez, and Michael L Tress. Assessing the functional relevance of splice isoforms. *NAR Genomics and Bioinformatics*, 3(2):lqab044, 2021.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

Brian J Raney, Galt P Barber, Anna Benet-Pagès, Jonathan Casper, Hiram Clawson, Melissa S Cline, Mark Diekhans, Clayton Fischer, Jairo Navarro Gonzalez, Glenn Hickey, et al. The ucsc genome browser database: 2024 update. *Nucleic Acids Research*, 52(D1):D1082–D1088, 2024.

Appendices

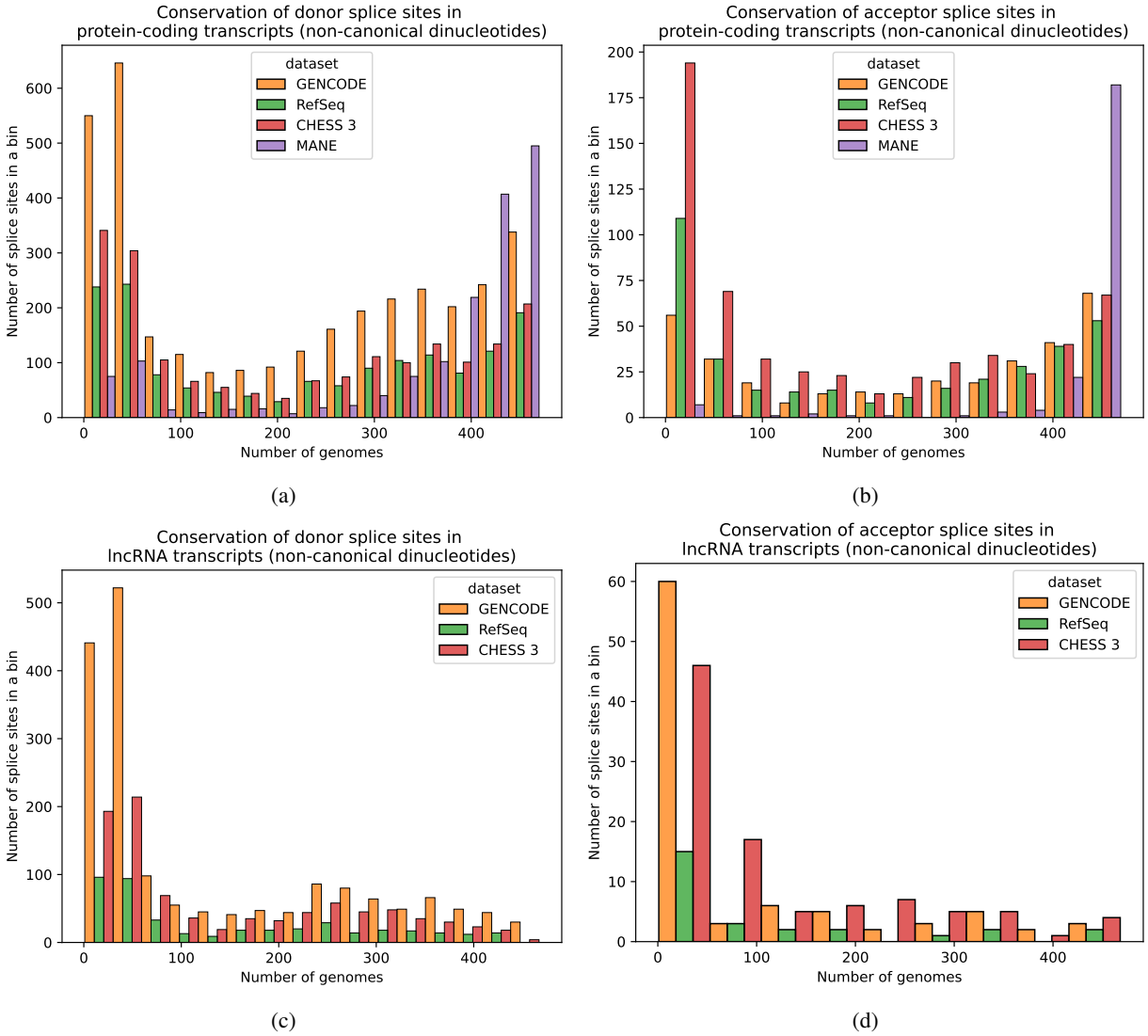


Figure A1: The number of human splice sites with non-canonical dinucleotides in place of GT/AG conserved in 470 mammals, computed for donor (a) and acceptor (b) sites of protein-coding genes, and donor (c) and acceptor (d) sites of lncRNAs. Each point shows a number of splice sites conserved (y-axis) in a given number of species (x-axis), data points are binned. For protein-coding genes, we created subsets GENCODE, RefSeq and CHESS 3 from which we removed MANE annotations, because each of these datasets are supersets of MANE; the resulting datasets are designated as GENCODE*, RefSeq* and CHESS 3* correspondingly.

	Donor sites				Acceptor sites			
Organism	GENCODE$^{(*)}$	RefSeq$^{(*)}$	CHESS 3$^{(*)}$	Random	GENCODE$^{(*)}$	RefSeq$^{(*)}$	CHESS 3$^{(*)}$	Random
Protein-Coding								
Chimpanzee	87.32%	88.24%	87.32%	85.41%	86.82%	87.70%	86.41%	85.14%
Pygmy chimpanzee	86.90%	87.62%	86.67%	85.39%	86.22%	86.87%	85.56%	85.15%
Western lowland gorilla	83.41%	84.58%	82.99%	82.51%	82.85%	83.82%	81.75%	82.33%
Sumatran orangutan	78.97%	80.13%	79.32%	78.71%	80.21%	80.57%	79.08%	78.47%
Northern white-cheeked gibbon	75.48%	76.34%	74.90%	75.14%	76.99%	77.25%	75.46%	74.76%
Silvery gibbon	75.18%	75.99%	74.22%	74.75%	76.55%	76.92%	75.13%	74.41%
Crab-eating macaque	66.99%	67.90%	67.43%	67.72%	69.75%	70.69%	70.28%	67.19%
Rhesus monkey	67.14%	68.27%	67.74%	67.51%	69.85%	70.98%	70.48%	67.05%
Gelada	67.06%	67.76%	67.41%	67.51˙	69.73%	70.06%	69.79%	67.12%
Japanese macaque	66.71%	67.47%	66.92%	67.04%	69.31%	70.27%	69.61%	66.57%
Golden snub-nosed monkey	66.76%	67.78%	67.68%	66.86%	69.22%	70.26%	69.85%	66.55%
Green monkey	66.87%	67.46%	67.25%	66.83%	68.83%	69.49%	69.09%	66.41%
Olive baboon	65.98%	66.76%	65.56%	66.70%	68.30%	69.05%	67.91%	66.36%
Sooty mangabey	66.46%	66.96%	66.39%	66.21%	68.68%	69.06%	68.27%	65.76%
Pig-tailed macaque	66.13%	66.76%	66.30%	66.26%	68.21%	69.10%	68.51%	65.89%
Mandrill	65.98%	66.76%	66.32%	66.19%	68.34%	69.02%	68.13%	65.73%
Ugandan red Colobus	65.72%	66.68%	66.02%	66.24%	67.82%	68.80%	67.93%	65.96%
Red shanked douc langur	65.83%	66.55%	66.09%	66.06%	68.07%	68.72%	68.09%	65.76%
Red guenon	65.84%	66.72%	65.80%	66.11%	68.18%	68.57%	67.43%	65.78%
Allen's swamp monkey	65.83%	66.29%	65.69%	66.05%	67.77%	68.42%	67.50%	65.67%
Francois's langur	65.86%	66.64%	65.99%	65.50%	68.34%	69.10%	68.42%	65.27%
Mona monkey	65.65%	66.20%	65.34%	65.81%	67.66%	67.96%	67.25%	65.39%
Black snub-nosed monkey	65.53%	66.19%	65.65%	65.27%	67.88%	68.47%	67.39%	64.98%
De Brazza's monkey	65.21%	65.47%	65.31%	65.54%	67.13%	67.55%	67.02%	65.00%
Hanuman langur	64.76%	65.46%	64.88%	65.15%	66.64%	67.36%	66.30%	64.82%
Drill	63.60%	63.87%	63.23%	64.90%	66.61%	67.32%	66.08%	64.56%
Angolan colobus	63.31%	64.09%	63.37%	64.34%	65.54%	66.68%	65.66%	64.11%
Proboscis monkey	49.14%	49.87%	48.99%	53.64%	51.32%	52.89%	50.48%	53.63%
White-faced saki	51.00%	51.41%	51.05%	51.46%	53.55%	54.33%	53.76%	51.25%
Black-handed spider monkey	50.68%	50.89%	50.90%	51.49%	53.11%	54.07%	53.11%	51.41%
lncRNA								
Chimpanzee	89.25%	89.38%	89.17%	-	88.75%	89.33%	88.75%	-
Pygmy chimpanzee	88.68%	88.80%	88.53%	-	88.47%	88.95%	88.28%	-
Western lowland gorilla	86.62%	86.31%	86.33%	-	86.44%	86.87%	86.31%	-
Sumatran orangutan	80.18%	80.57%	80.29%	-	82.23%	82.61%	82.34%	-
Northern white-cheeked gibbon	74.42%	74.80%	74.36%	-	76.29%	76.74%	75.97%	-
Silvery gibbon	74.09%	74.81%	73.88%	-	75.86%	76.69%	75.54%	-
Rhesus monkey	67.47%	67.33%	66.90%	-	70.59%	71.07%	69.82%	-
Crab-eating macaque	67.15%	67.08%	66.56%	-	70.21%	70.67%	69.53%	-
Japanese macaque	67.10%	66.90%	66.47%	-	70.09%	70.51%	69.29%	-
Gelada	66.89%	66.77%	66.37%	-	70.01%	70.49%	69.37%	-
Green monkey	66.77%	66.90%	66.26%	-	70.11%	70.60%	69.15%	-
Golden snub-nosed monkey	66.77%	66.86%	66.33%	-	69.65%	70.29%	69.07%	-
Olive baboon	66.42%	66.26%	65.63%	-	69.59%	69.90%	68.39%	-
Pig-tailed macaque	66.37%	66.29%	65.65%	-	69.41%	69.90%	68.49%	-
Sooty mangabey	66.23%	66.15%	65.67%	-	69.02%	69.62%	68.53%	-
Red shanked douc langur	66.09%	65.93%	65.52%	-	69.10%	69.28%	68.43%	-
Mandrill	66.21%	65.85%	65.32%	-	68.91%	69.63%	68.28%	-
Red guenon	65.73%	65.87%	65.35%	-	69.02%	69.57%	68.01%	-
Black snub-nosed monkey	65.78%	65.89%	65.38%	-	68.58%	69.30%	67.88%	-
Francois's langur	65.79%	65.58%	65.15%	-	68.75%	68.83%	67.67%	-
Ugandan red Colobus	65.28%	65.39%	64.70%	-	68.19%	68.66%	67.22%	-
Allen's swamp monkey	65.31%	65.35%	64.62%	-	68.12%	68.47%	67.35%	-
Mona monkey	65.10%	64.84%	64.13%	-	68.20%	68.55%	67.40%	-
Drill	64.54%	64.61%	64.02%	-	67.81%	68.39%	67.08%	-
De Brazza's monkey	64.62%	64.55%	64.02%	-	67.38%	67.99%	66.84%	-
Hanuman langur	64.83%	64.50%	63.92%	-	67.50%	67.61%	66.42%	-
Angolan colobus	64.11%	64.25%	63.64%	-	66.71%	67.69%	66.24%	-
Proboscis monkey	53.29%	53.07%	52.25%	-	55.39%	56.29%	54.27%	-
White-faced saki	49.85%	50.47%	50.05%	-	53.68%	54.54%	53.15%	-
Black-handed spider monkey	48.92%	49.37%	48.79%	-	53.34%	53.30%	52.25%	-

Table A1: Top 30 most common species in which donor and acceptor splice sites are conserved in different datasets, including the "Random" baseline dataset (for comparison). To compute these species, we only used splice sites from GENCODE, RefSeq, CHESS conserved in less than 50 species that appear to have a random-like conservation pattern. Each entry shows the proportion of splice sites that have their canonical dinucleotides conserved in certain species relative to the total number of sites in the respective category. The first half of the table shows sites from protein-coding genes, the second one shows sites from lncRNAs. For the second part, the "Random" column show dashes since the same set of generated sites were used as a baseline for comparing against both coding and non-coding genes.

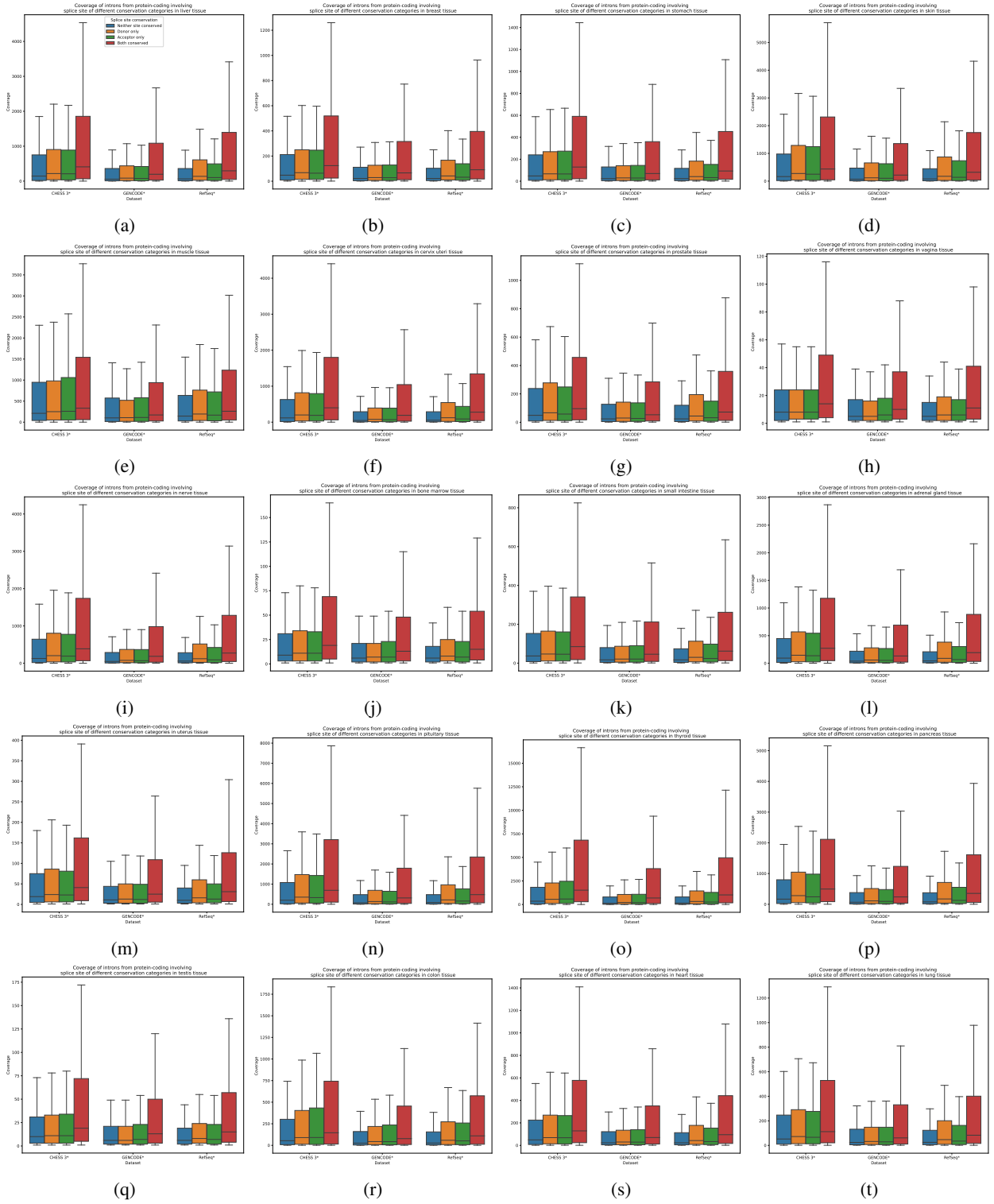


Figure A2: Box plots showing coverage of introns from protein-coding genes of GENCODE*, RefSeq*, from CHESS 3* across different tissues. Each box plot shows the median (dashed red line), the interquartile range (solid top and bottom borders of the box), and minimum and maximum values within ±1.5 of the interquartile range (whiskers), outliers are not shown.

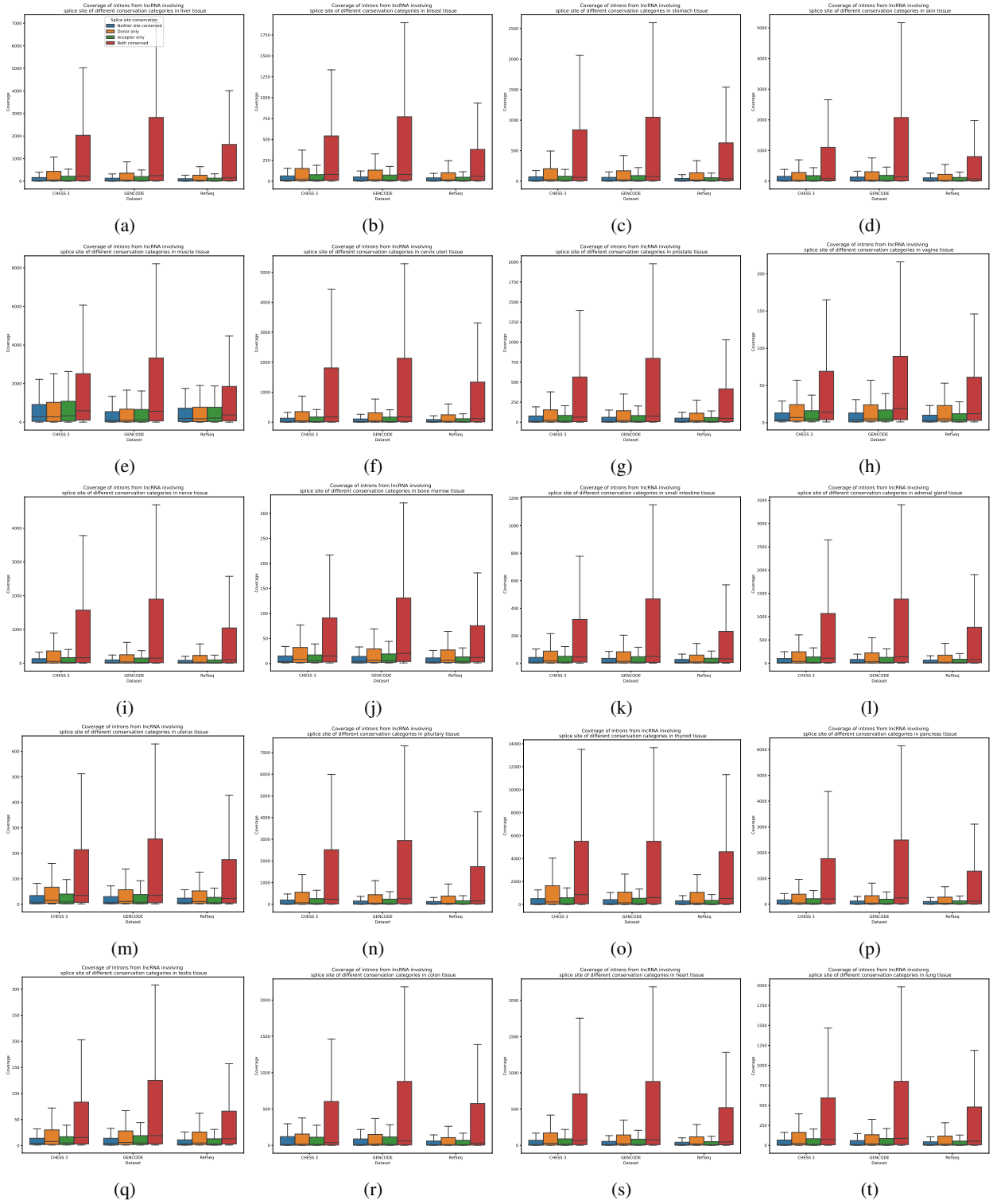


Figure A3: Box plots showing coverage of introns from lncRNAs of GENCODE*, RefSeq*, from CHESS 3* across different tissues. Each box plot shows the median (dashed red line), the interquartile range (solid top and bottom borders of the box), and minimum and maximum values within ±1.5 of the interquartile range (whiskers), outliers are not shown.