QUALITY ASSESSMENT OF SPLICE SITE ANNOTATION BASED ON CONSERVATION ACROSS MULTIPLE SPECIES

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 even the best annotations of the human reference genome sometimes contradict one another. Hence, refinement of the human genome annotation is an important challenge. The use of evolutionarily conserved sequences provides a strategy for addressing this problem, and the rapidly growing number of genomes from other species increases the power of an evolution-driven approach. Using the latest large-scale whole genome alignment data, 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, from both protein-coding genes and lncRNAs. We trained a logistic regression classifier to distinguish between the conservation patterns exhibited by splice sites from MANE versus sites that were flanked by the standard GT-AG dinucleotides, but that were chosen randomly from a sequence not under selection. We found that up to 70% of splice sites from annotated protein-coding transcripts outside of MANE exhibit conservation patterns closer to random sequence as opposed to highly-conserved splice sites from MANE. Our study highlights potentially erroneous splice sites that might require further scrutiny.

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, its annotation remains inaccurate and incomplete [Amaral et al., 2023]. 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]. One consequence is that the leading gene annotation databases for the human reference genome often contain contradictory information, and they do not agree even on basic statistics such as the number of protein-coding genes [Pertea et al., 2018]. Thus, quality assessment and improvement of human genome annotation remains an important challenge in genomics.

One approach that has proven itself repeatedly is to use evolutionary conservation to identify genes and gene-related elements, because most of them are constrained by negative or purifying selection. Thus, sequences from the human genome (or any genome) that are conserved in multiple other species are likely to represent functional elements such as protein-coding genes. Splice sites flanking introns are of particular interest for evolutionary analysis due to their importance for the correct transcription of genes. Although multiple previous studies have suggested that conservation data can be used to spot non-functional elements (e.g., [Lareau et al., 2004]), no comprehensive method yet exists to spot annotation mistakes using large-scale, whole-genome alignment data.

In this study, we attempt to assess the quality of splice site annotation using conservation in a multiple-genome alignment containing 470 mammalian species. First, we observed that the canonical dinucleotides GT/AT that flank introns are very highly conserved in protein-coding genes in MANE [Morales et al., 2022], with most of them being intact in more than 400 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 non-conserved by our classifier have higher rates of single nucleotide polymorphisms (SNPs) in the human population, and they are less likely to appear in multiple isoforms of the same gene. Our findings suggest that some of the non-conserved sites are likely to be mis-annotated and require further scrutiny.

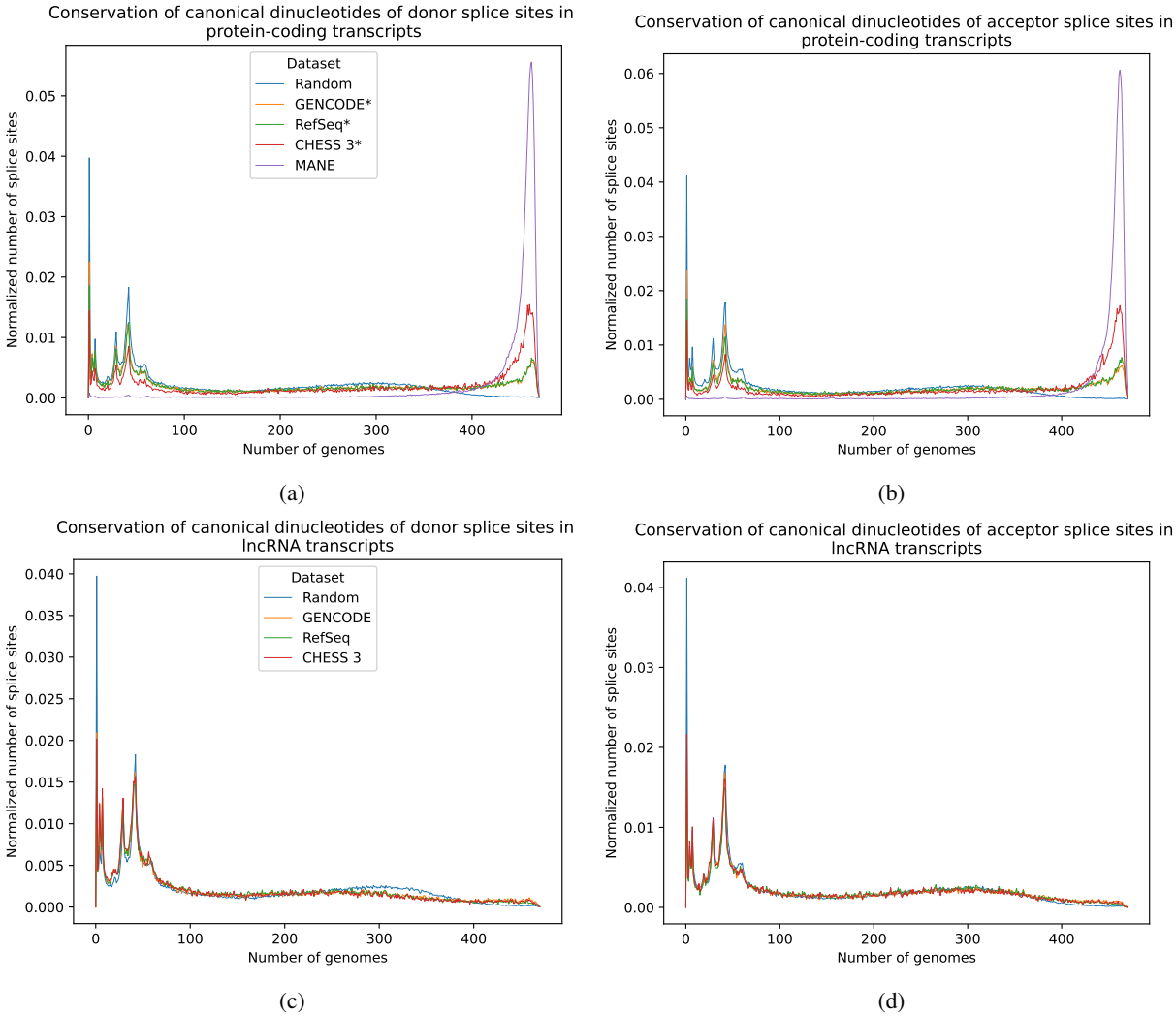


Figure 1: Distribution of the number of human splice sites with canonical dinucleotides (GT for donor and AG for acceptor sites) conserved in 470 mammals, computed for donor (a) and acceptor (b) sites of protein-coding genes, and donor (c) and acceptor (d) sites of lncRNAs. Each point shows a number 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.

2 Results

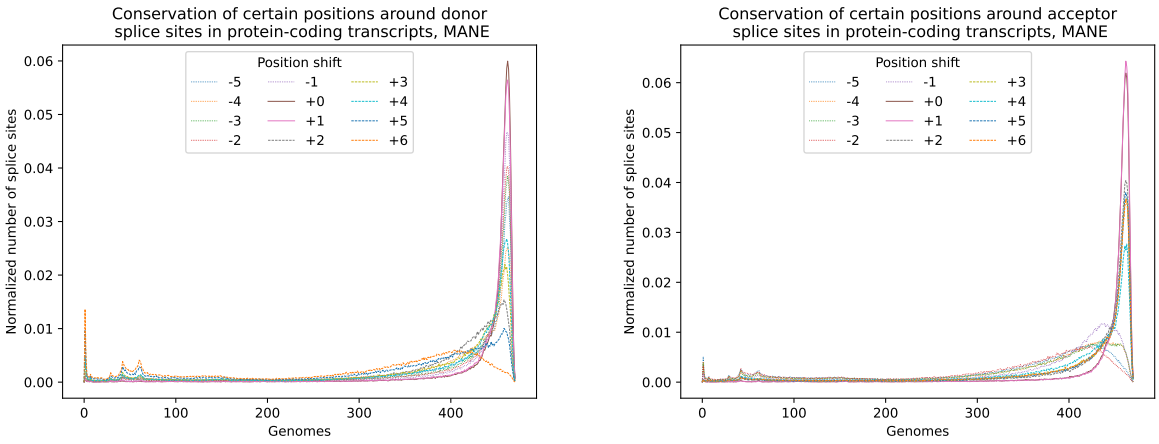
2.1 Exploratory data analysis of splice site conservation

First, we evaluated the evolutionary conversation of splice sites from four different human genome annotation databases: GENCODE [Frankish et al., 2021] version 38, RefSeq [O'Leary et al., 2016] release 110, CHESS 3 [Varabyou et al., 2023] v.3.0, and MANE [Morales et al., 2022] v1.0. 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.

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. 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.

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 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.



(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 sites		"Conserved" sites	
Dataset	# Donor	# Acceptor	# Donor	# Acceptor
Protein-Coding				
MANE	181,928	181,890	-	-
GENCODE*	62,294	53,772	17,604	16,925
RefSeq*	46,500	38,089	13,552	12,351
CHESS 3*	51,802	46,791	24,839	23,809
lncRNA				
GENCODE	47,312	48,232	5,142	6,107
RefSeq	45,172	45,294	4,082	5,246
CHESS 3	48,123	48,549	4,496	5,576
Synthetic data				
Random	180,000	180,000	-	-

Table 1: Number of unique donor and acceptor splice sites in each dataset. The second and the third columns represent the total number of sites, while last two columns show the number of splice classified as "conserved" by our model as per subsection 2.2.

We further explored the question of whether poorly conserved splice sites might be be human-specific. Specifically, we calculated the fraction of splice sites that have at least one single-nucleotide polymorphisms (SNP) that overlaps its canonical dinucleotides GT or AG. To determine the presence of SNPs in the human population, we used the gnomAD database version 4.0.0 [Karczewski et al., 2020], focusing on loci that have at least one homozygous sample since a homozygous SNP at a splice site is very likely to cause incorrect splicing. We then 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. Figure 5 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;

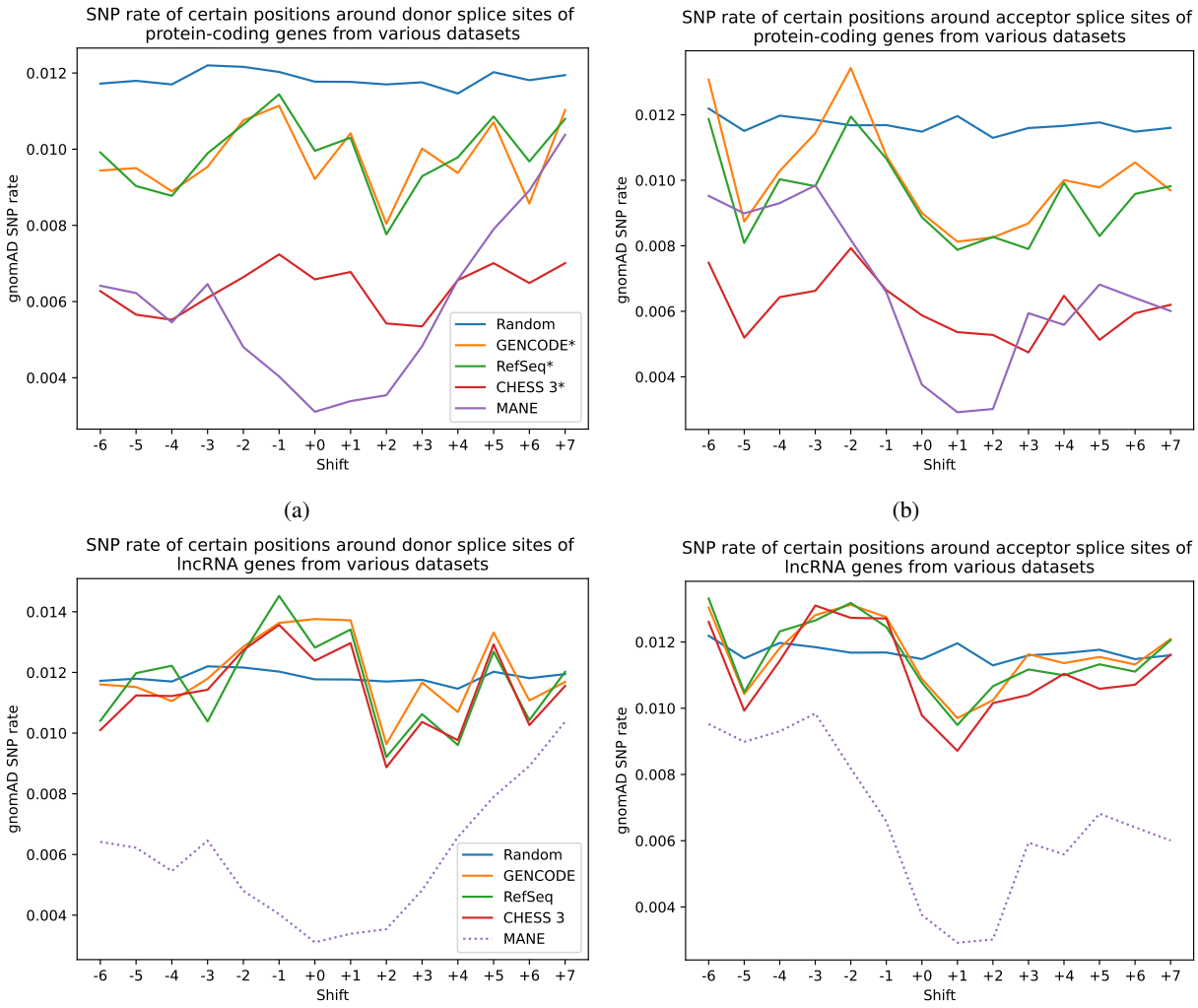


Figure 3: Rate of SNPs at positions near splice sites. Each point represents a proportion of splice sites from a certain dataset that have an SNP from the gnomAD dataset at a position either down- or upstream of the "canonical" dinuncleotides. For example, for donor splice sites +0 is usually "G", +1 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.

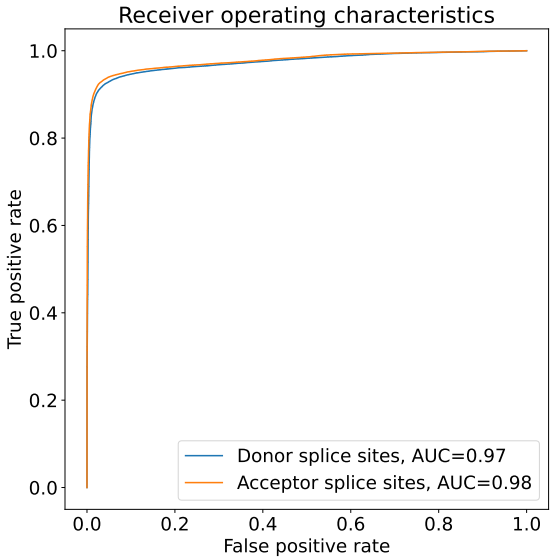


Figure 4: Receiving operating characteristic (ROC) curve for logistic regression models trained to distinguish between GT/AG sites chosen at random from intronic sequences and splice sites from MANE.

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 that might represent non-functional and/or erroneous annotations; otherwise, their average SNP rates should have been more similar to what we observed in splice sites from MANE.

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 cluster 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. 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. The model shows high accuracy on the test data with area under the ROC curve (AUROC) measuring 0.97 for donor and 0.98 for acceptor sites. 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.

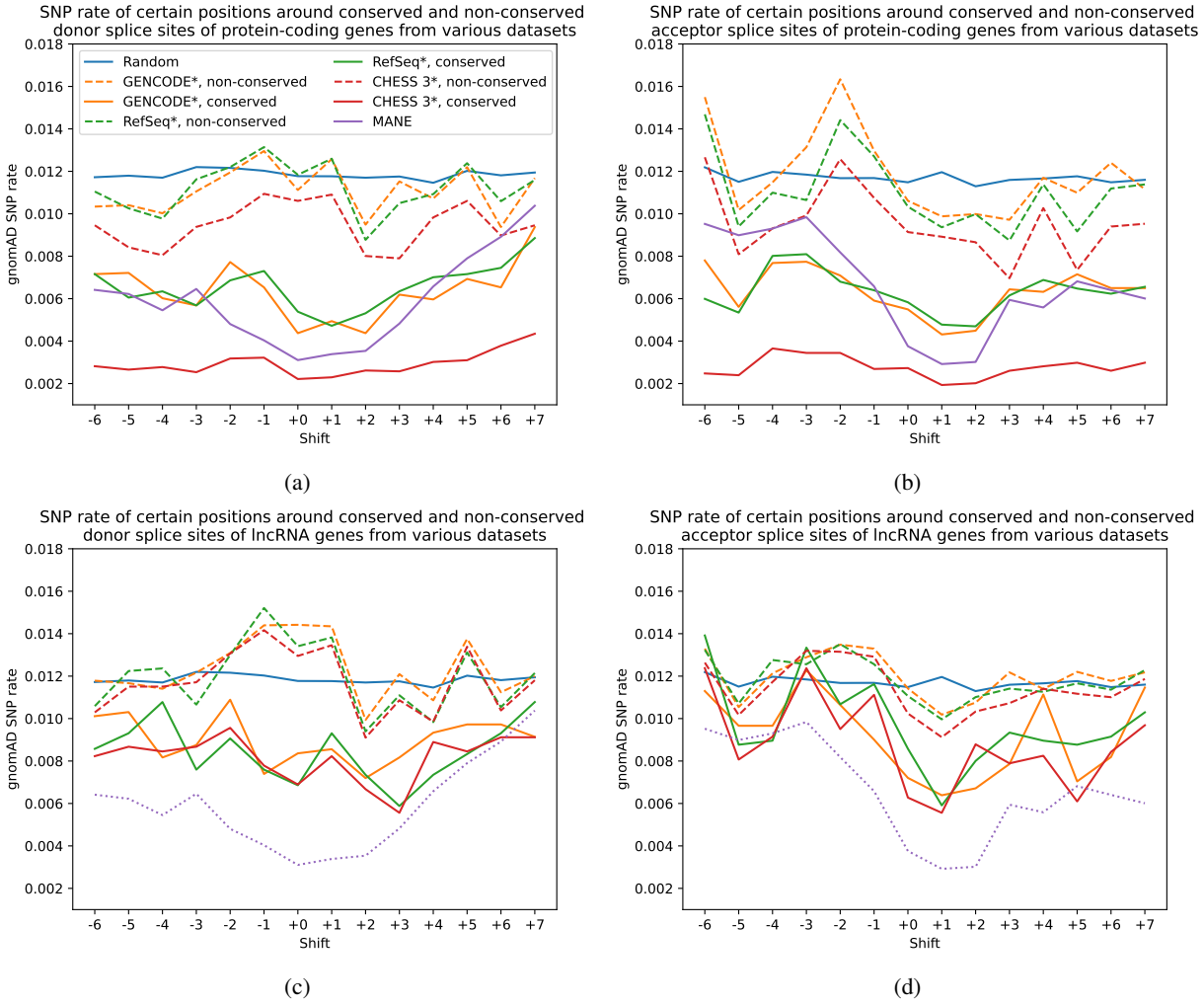
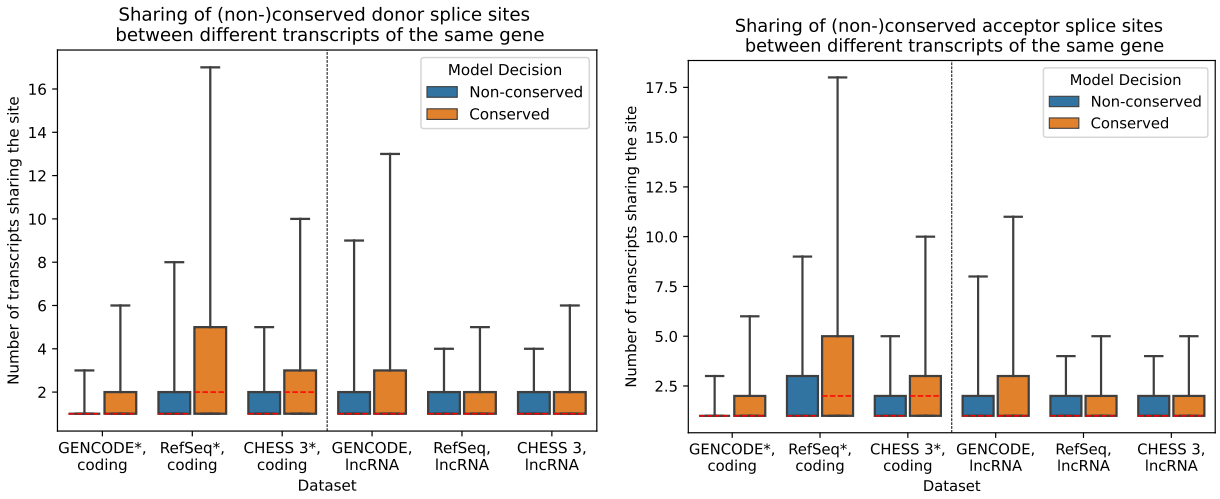


Figure 5: Rate of homozygous SNPs at positions near splice sites. Each point represents a proportion of splice from a certain dataset that have an SNP at a position either down- or upstream of the "canonical" dinuncleotides. For example, for donor splice sites +0 is usually "G", +1 is "T", and -1 is 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.



(a) (b)

Figure 6: 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.

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 28–32% of splices sites were conserved according to the model, while for CHESS 3*, the proportion was substantially higher, at 48% 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, suggesting that many of these are either non-functional or alternatively that they have a far higher likelihood of being species- or clade-specific.

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.

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 6 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. 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 similar to MANE. We also found that non-conserved splice sites are less likely to be shared by different isoforms of the same gene. These findings suggest that at least some of these splice sites are likely to be mis-annotated or to belong to non-functional isoforms.

Previous studies of splice site evolution [Denisov et al., 2014, Kurmangaliyev et al., 2013, Shimada et al., 2010] have largely assumed that the underlying genome annotation was correct. However, gene annotation is far from perfect, and evolutionary data has great potential for highlighting potential errors. Our study is the first one to illustrate how inconsistencies in conservation patterns of splice sites across a very large number of vertebrate species and be used to identify possibly erroneous annotations. We note that previous studies also found a lack of conservation of some splice sites using whole genome alignments [Sharma et al., 2016, 2017], but those studies assumed that these patterns arose due to errors in alignment rather than the annotation itself. 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. This finding suggests that at least a subset of splice sites lacking conservation is likely to be mis-annotated as opposed to being poorly aligned.

This study has several limitations. First, 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, although we characterized a set of splice sites whose pattern of conservation resembles neutrally evolving sequences, we cannot resolve precisely which splice sites are mis-annotated. The fact that this group has a higher rate of SNPs overlapping their canonical dinucleotides suggests that the group contains erroneous splice sites. However, pinpointing exactly which ones are errors would require further study incorporating extra data. Third, 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. 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. We also highlighted a subset of splice sites in the leading human annotation catalogues that might require further scrutiny and analysis. 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. 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 several 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 $g_{1,o(a_r)+1} = G$.

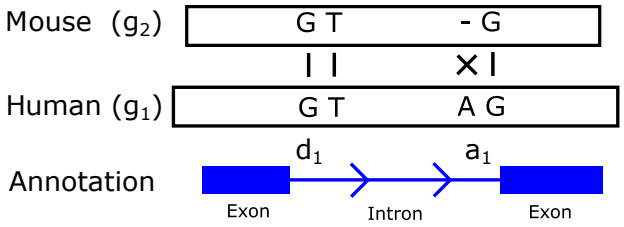


Figure 7: An example of mapping splice sites from a reference genome to another species using a whole-genome alignment. The donor site d_1 of the human reference genome g_1 has both its canonical dinucleotides intact in the target mouse genome, g_2. However, this is not true for the acceptor site a_1 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 7 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 15 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:

$$\log(\frac{p(a_r)}{1-p(a_r)}) = \alpha_0 + \alpha_1 \sum_{1<t\leq m} I[C(a_r, 0, t) = 1 \wedge C(a_r, 1, t) = 1] + \sum_{-15<\ell\leq 16,\, \ell\neq 0,1} \alpha_\ell \sum_{1<t\leq m} C(a_i, \ell, t)$$

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.

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 [Kent et al., 2002].

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.

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.

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.

Liana F Lareau, Richard E Green, Rajiv S Bhatnagar, and Steven E Brenner. The evolving roles of alternative splicing. *Current opinion in structural biology*, 14(3):273–282, 2004.

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.

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.

Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

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.

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.

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. doi:10.1093/nar/gkw210. URL `https://doi.org/10.1093/nar/gkw210`.

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. doi:10.1093/bioinformatics/btx527. URL `https://doi.org/10.1093/bioinformatics/btx527`.

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.