Augmenting transcriptome annotations through the lens of splicing evolution

Xiaofei Carl Zang, Ke Chen, Irtesam Mahmud Khan, et al.

Genome Res. published online May 29, 2026
Access the most recent version at doi:10.1101/gr.280661.125

Augmenting transcriptome annotations through the lens of splicing evolution

Xiaofei Carl Zang[1,2], Ke Chen[3], Irtesam Mahmud Khan[3], and Mingfu Shao[1,3,*]

[1]Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, Pennsylvania, 16802, USA
[2]Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, 19104, USA
[3]Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania, 16802, USA
[*]Correspondence: mxs2589@psu.edu

Abstract

Transcriptome annotations remain incomplete despite enormous efforts. Annotations are largely driven by experimental data, while little is understood from an evolutionary perspective. Here we present TENNIS, a model for isoform representation and inference. TENNIS models isoforms in a transcript group as nodes of a connected graph, where edges represent basic alternative splicing events, and predicts missing isoforms using a novel algorithm. Our analysis indicates that approximately 80% of the analyzed isoform groups satisfy our model, while the identified missing transcripts show high accuracy. TENNIS achieves these results without using additional sequencing data, offering insights into alternative splicing and a powerful tool for constructing annotations.

Introduction

Alternative splicing (AS) is a ubiquitous and prevalent mechanism in eukaryotes. It selectively splices in or splices out some exons from the same pre-mRNA (Sugnet et al., 2004). AS increases the diversity of transcript isoforms (Birzele et al., 2008; Wright et al., 2022) and happens more frequently and more independently than previously estimated (Zhang et al., 2017). It is estimated that over 90% of human genes are alternatively spliced (Pan et al., 2008; Wang et al., 2008). There are four basic types of AS events (Sugnet et al., 2004): (1) cassette exon (CE), also known as

exon skipping/inclusion, (2) alternative 3' splicing site (A3), (3) alternative 5' splice sites (A5), and (4) intron retention (IR). Some complex AS events, such as multiple exon skipping or mutually exclusive exons, may be considered as the synergy of two basic AS events.

The study of AS is extensive, ranging from the mechanisms of splicing regulation (Wang et al., 2014; Chen and Manley, 2009) to the functions of splicing isoforms and their associations with diseases (Tao et al., 2024; Hoyos and Abdel-Wahab, 2018; Scotti and Swanson, 2015). One important angle of studying AS is through evolution. It is known that AS is under rapid evolution and is elastically shaped by environments (Zhang et al., 2024). Elucidating the evolutionary relationship across splicing isoforms originating from the same pre-mRNA is crucial, as it is closely related to functional diversification of genes and offers a powerful tool to study splicing regulation (Kim et al., 2007a; Singh and Ahi, 2022). For example, AS might have originated through DNA mutations in the splicing sites, control sequences, and the evolution of splicing regulators (Keren et al., 2010; Ast, 2004). It was also reported that multi-intron genes may precede the emergence of AS, and in primate species, AS events combine independently with each other so that novel AS isoforms emerge (Ast, 2004; Zhang et al., 2017). Despite these biological advances, there remains a significant shortage of mathematical models that quantitatively characterize splicing evolution.

The catalog of all splicing isoforms of all genes, i.e. transcripts, for a species is called the transcriptome. These transcripts not only transcribe genetic information to encode proteins but also play important regulatory and functional roles (Statello et al., 2021; Mattick et al., 2023). Various biological and biomedical studies are heavily dependent on fine-grained transcriptome annotations, including the quantification of transcripts, the curation of a single-cell expression atlas, the identification of aberrant splicing in disease-related samples, and comparative transcriptomics.

Over the past decades, tremendous effort has been put into constructing and improving the annotations of transcriptomes, especially the model organisms. For illustration, the major consortia for the annotation of the human species include RefSeq (Li et al., 2021), Ensembl (Aken et al., 2016), CHESS (Pertea et al., 2018), and MANE (Morales et al., 2022). These annotations were primarily conducted in a data-driven manner, where one common approach is to perform assembly from RNA-seq data (Salzberg, 2019; Raghavan et al., 2022). The assemblies are often additionally augmented or validated by experimental data. For example, NCBI annotations, including RefSeq, also consider transcript sequences, reads in the SRA database, CAGE-Seq, amino acid sequences,

and curated data from other sources (Li et al., 2021; NCBI, n.d.). The Ensembl annotation consolidates information from cDNAs, protein sequences, RNA-seq, and manual curations (Aken et al., 2016). CHESS is based on a large-scale RNA-seq of nearly ten thousand samples (Pertea et al., 2018). The MANE annotation constitutes a consensus between RefSeq and Ensembl with manual curations (Morales et al., 2022). Despite the significant number of computational tools, pipelines, and manual curations, the transcriptome annotations are not complete even for model organisms (Salzberg, 2019; Zhang et al., 2020; Zerbino et al., 2020). Humans, the undoubtedly most-studied species, have had a continually increasing number of recorded genes and transcripts from GRCh37 to GRCh38 (Schneider et al., 2017) and to T2T-CHM13 (Nurk et al., 2022). Annotations for other model organisms, e.g. mouse or *Drosophila*, are also incomplete, as novel transcripts were found with higher sequencing depth and more comprehensive sequencing experiments (Leung et al., 2021; Tian et al., 2021; Alfonso-Gonzalez et al., 2023).

In this work, we propose an evolution-inspired mathematical model for alternative splicing. Based on this model, we develop a tool called TENNIS (Transcript EvolutioN for New Isoform Splicing) that predicts missing isoforms in an annotation (without using any external sequencing data). Our model characterizes the AS evolution trajectory based on two simple premises. First, evolution does not create new splicing isoforms out of thin air, rather, it modifies and adapts existing ones; and second, evolution takes baby steps, namely, each isoform is derived from its predecessor through a single AS event. The problem of identifying missing isoforms is formulated as an optimization problem following the parsimony principle: find the minimum number of transcripts whose inclusion connects all observed AS isoforms, such that each pair of adjacent isoforms differs by a single AS event. We applied TENNIS to transcriptome annotations of various species to validate our evolution-inspired model, and evaluated its performance in predicting missing isoforms from both real and simulated datasets.

Results

Overview of the TENNIS model

TENNIS is built on an evolution-inspired model of alternative splicing. We consider transcript groups consisting of isoforms that share the same transcription start site (TSS) and transcription

end site (TES), meaning they originate from identical pre-mRNAs and differ solely due to alternative splicing. Our model rests on two premises: (1) new splicing isoforms arise by modification of existing ones, and (2) each isoform derives from its predecessor through one of the four AS events—cassette exon (CE), alternative 5' splice site (A5), alternative 3' splice site (A3), or intron retention (IR) (Fig. 1A).

To formalize this model, we represent each transcript as a binary vector encoding the inclusion (1) or exclusion (0) of genomic regions delineated by splice sites (Fig. 1B,C). Two transcripts are connected by an edge if they differ by exactly one AS event. For example, in Fig. 1D, transcripts t_1 and t_3 differ only in the third genomic region and are thus connected. It is easy to identify this event as an A3 event. Importantly, AS events including or excluding multiple consecutive partial exons—such as those arising from alternative 5' or 3' splice sites—are treated as single events despite spanning over multiple genomic regions (Fig. 1E). Under this model, all isoforms within a transcript group should form a connected graph, provided none is missing.

When annotated isoforms fail to form a connected graph, it indicates that the annotation may be missing intermediate transcripts. TENNIS identifies such gaps and predicts the minimum number of novel isoforms needed to restore connectivity. We formulate this task as an optimization problem and design an algorithm, TENNIS-SAT, that transforms the subroutine problem into a Boolean satisfiability (SAT) instance: given the binary matrix representation of a transcript group, TENNIS-SAT determines whether adding k novel isoforms suffices to produce a connected graph, and if so, returns their binary representations. For transcript groups with exactly two connected components, an optimal greedy algorithm provides an efficient exact solution. For more complex cases, TENNIS-SAT iteratively tests increasing values of k until a solution is found or computational limits are reached (see Methods for details). We denote the set of single-isoform transcript groups as $\mathcal{T}_S$ and the set of multi-isoform transcript groups as $\mathcal{T}_M$. Multi-isoform transcript groups are further classified into $\mathcal{T}_M^k$ $(k = 0, 1, \cdots)$ and $\mathcal{T}_M^{@}$ groups, meaning they require k additional isoforms or exceed computational resources.

Most transcript groups satisfy the AS evolution model

In a well-annotated transcriptome, we expect most transcript groups to satisfy our evolution-inspired model. To verify, we analyzed 7 transcriptome annotations from 6 model species: human (GRCh38

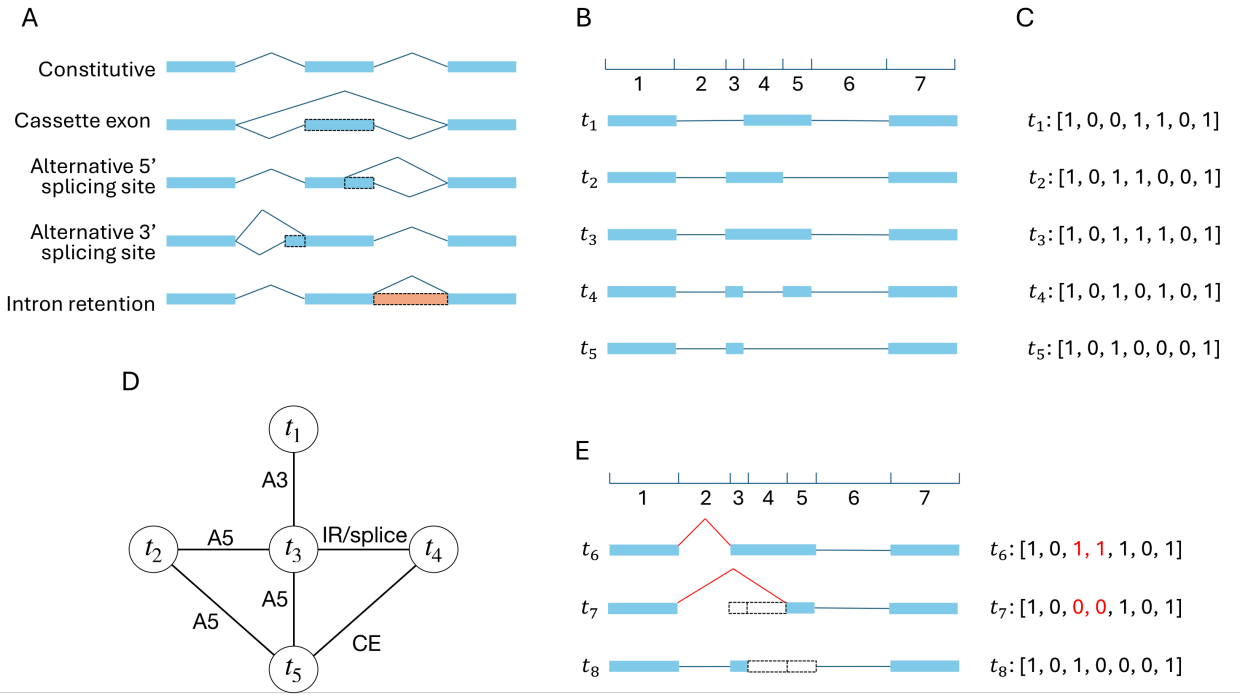

Figure 1: Alternative splicing events and transcript-group representation. (A) Constitutive isoform splicing and four basic alternative splicing types. Blue rectangle: exon; Peach rectangle: retained intron; Dashed rectangle: alternative (partial) exon/intron; Blue polyline: splice junction. (B) Splice sites of all transcripts divide the genome into several sub-regions. This example shows a group of 5 transcripts (t_1, t_2, t_3, t_4, t_5) that divides the genome into 7 regions. (C) Each transcript is encoded as a binary vector indicating which regions are spliced in (1) or spliced out (0). This example encodes each panel B transcript as a vector of length 7. The whole transcript group is represented as a 5×7 binary matrix. (D) Potential AS events between panel B transcripts. Only exactly one basic AS event is considered between each pair of transcripts. t_1 is convertible to t_3 by one A3 event, t_2 to t_3 by one A5 event, t_2 to t_5 by one A5 event, t_3 to t_4 by one intron retention or splicing event, t_3 to t_5 by one A5 event, t_4 to t_5 by one cassette exon event. (E) Skipping multiple consecutive partial exons is one AS event. The red splice junctions and binary bits illustrate that converting t_6 to t_7 is one A3 event, but two consecutive partial exons are skipped. Similarly, converting t_6 to t_8 is one A5 event (junctions not shown). However, converting t_7 to t_8 requires two basic AS events, because the changed partial exons (3rd and 5th) are not consecutive and one is skipped while the other is included.
CE: cassette exon; A5: alternative 5' splice sites; A3: alternative 3' splicing site; IR: intron retention.

RefSeq and GENCODE), mouse (GRCm39), *Drosophila* (dm6), zebrafish (GRCz11), maize (Zm-B73-REFERENCE-NAM-5.0) and *Arabidopsis* (TAIR10) (Supplemental Table S1). For each transcriptome, we first partition all multi-exon transcripts into transcript groups, as described in Methods. TENNIS is applied to all transcript groups, and according to the outcomes, they are partitioned into 7 categories: $\mathcal{T}_S, \mathcal{T}_M^0, \cdots, \mathcal{T}_M^4, \mathcal{T}_M^@$.

The statistics are reported in Table 1. We found that 70%-87% of the transcript groups fall into

Species	Annotation	$\mathcal{T}_M^0$	$\mathcal{T}_M^1$	$\mathcal{T}_M^2$	$\mathcal{T}_M^3$	$\mathcal{T}_M^4$	$\mathcal{T}_M^{@}$	$\mathcal{T}_M$	$\mathcal{T}_S$
Human/GRCh38	GENCODE	11960(62%)	4277(22%)	1654(9%)	676(3%)	322(2%)	536(3%)	19425	125777(87%)
Human/GRCh38	RefSeq	20852(78%)	3951(15%)	1169(4%)	435(2%)	199(1%)	278(1%)	26884	63541(70%)
Mouse	GRCm39	17178(84%)	2321(11%)	599(3%)	190(1%)	92(0%)	102(0%)	20482	49524(71%)
Drosophila	dm6	2433(78%)	451(14%)	116(4%)	46(1%)	28(1%)	42(1%)	3116	17938(85%)
Zebrafish	GRCz11	7455(88%)	673(8%)	155(2%)	62(1%)	21(0%)	61(1%)	8427	41836(83%)
Maize	NAM-5.0	16799(88%)	1612(8%)	332(2%)	142(1%)	53(0%)	46(0%)	18984	83398(81%)
Arabidopsis	TAIR10	1481(88%)	147(9%)	33(2%)	10(1%)	2(0%)	2(0%)	1675	9105(84%)

Table 1: Summary statistics of the number of transcript groups in each category.

the $\mathcal{T}_S$ category, meaning they have just one (multi-exon) transcript. For these groups, transcript isoform diversity arises from distinct TSS and/or TES usage rather than from alternative splicing within a group. This observation aligns well with previous studies that TSS and TES are the major sources of transcriptome diversity (Reyes and Huber, 2018), and the selection of TSS and TES is coordinated (Alfonso-Gonzalez et al., 2023; Calvo-Roitberg et al., 2025). Human RefSeq has the lowest single-transcript group rate (70%) while human GENCODE has the highest single-transcript group rate (87%). We note that GENCODE has many more transcript groups than RefSeq (145202 vs. 90429) but fewer of them are multi-transcript groups (26884 vs. 19425). This indicates GENCODE annotated more genes and alternative TSS/TES isoforms but fewer AS isoforms per gene.

Among the transcript groups with multiple transcripts (i.e., $\mathcal{T}_M$), the majority (78%–88%, except human/GENCODE) satisfy our model (i.e., are in $\mathcal{T}_M^0$), supporting the rationale of this model. Human GENCODE is an outlier, with 62% of transcript groups ending up in $\mathcal{T}_M^0$. This might be due to a combination of GENCODE over-annotating some transcripts with alternative TSS/TES isoforms and some transcript groups being incomplete. Among transcript groups that do not satisfy our model, the majority are in $\mathcal{T}_M^1$, i.e., for most groups, only one additional isoform is required to make it complete. Lastly, approximately only 1% of transcript groups require more than 4 transcripts to meet our model or time out in 15 minutes for 6 out of the 7 annotations (it is 3% of groups for human GENCODE). Consistent with this observation, for dm6, all transcript groups with an MST-based upper bound of at most 4 were solved within 15 minutes, and the MST-based upper bound is typically equal to or close to the optimal number of novel isoforms across annotations (Supplemental Fig. S1). This suggests that the 15-minute timeout threshold and the default maximum of 4 missing isoforms serve as a sufficient balance between completeness and efficiency for the great majority of transcript groups.

TENNIS-predicted isoforms are validated by long-read RNA-seq data

It is of great interest to test whether TENNIS is able to predict correct novel isoforms. Since TENNIS predicts novel/missing isoforms from the reference transcriptome annotations without additional input, if those isoforms can be cross-validated by other data sources such as RNA-seq or external databases, then they are likely to be true positives. In this way, we demonstrate the accuracy and applicability of TENNIS.

We chose the *Drosophila* transcriptome as an example, which is relatively small and well-studied. We retrieved an assembly of high-depth long-read RNA-seq data from a previously published dataset (Alfonso-Gonzalez et al., 2023). We used GffCompare (Pertea and Pertea, 2020) to compare the predicted isoforms from TENNIS against this assembly. GffCompare considers two multi-exon isoforms matching if they have the same intron-chain, which is a widely accepted practice. A TENNIS-predicted novel isoform is considered "matched" if it shares the same intron chain as a transcript from a different source. Otherwise, the prediction is considered "unmatched". In this experiment, we consider "matched" as true-positive and "unmatched" as false-positive. Accordingly, the number of matched predictions is proportional to sensitivity, and the frequency of matched predictions is proportional to precision.

We also set up baseline comparisons through randomized approaches and exon usage-based approaches. Only transcript groups requiring novel isoform prediction, namely those in $\bigcup_{k=1}^{4} \mathcal{T}_M^k$, were used in these baseline experiments; for dm6 this corresponds to 641 groups, whereas the remaining $\mathcal{T}_M^0$ groups were already complete under our model and therefore did not require prediction. Specifically, within each eligible group, all constitutive exons were always included, and alternative exons were combined to create novel, previously unobserved isoform predictions. We used two random baselines. In the first one, referred to as "Rand1", one isoform per eligible transcript group is randomly generated; in the second one, termed "Rand$X$", k novel isoforms per group in $\mathcal{T}_M^k$ were produced, where the value of k is obtained by TENNIS. To reduce random noise, both "Rand1" and "Rand$X$" experiments were repeated 5 times. Their means and standard deviations were reported. In addition to the random baselines, we designed two more realistic baseline methods that consider exon usage frequencies. We retrieved percent spliced in (PSI) values of alternative exons in *Drosophila* dm6 annotation from VastDB (Tapial et al., 2017) and computed the average PSI

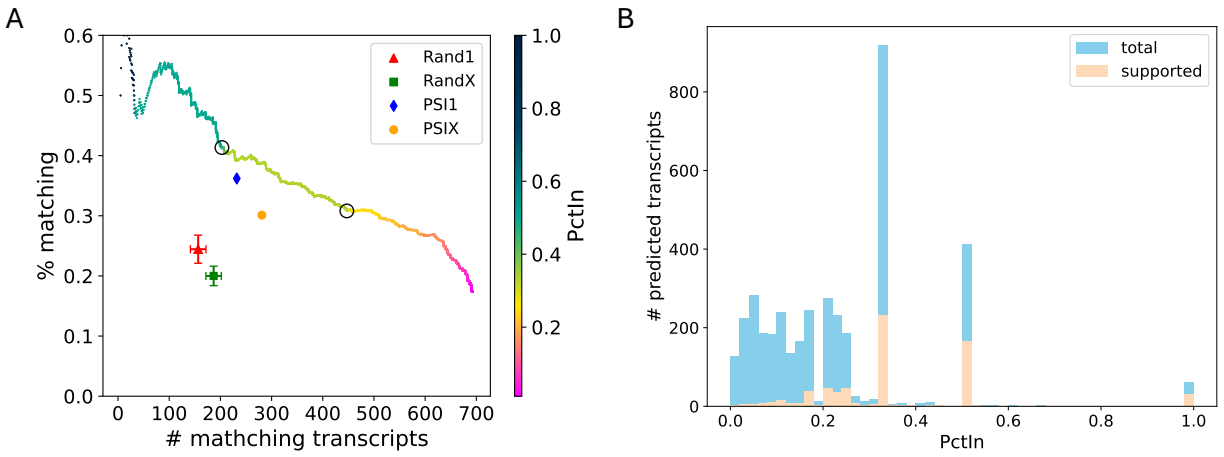

Figure 2: TENNIS augmentations on the *Drosophila* transcriptome dm6. (A) The number and percentage of transcript predictions matched by long-read RNA-seq, sorted in descending order of PctIn and then splicing probability score. The color gradient indicates PctIn values, with circles highlighting critical thresholds (PctIn = 0.5, 0.333). Baseline methods include random (Rand1, Rand$X$) and PSI-based (PSI1, PSIX) approaches. Error bars for random baselines represent standard deviation over 5 repetitions. (B) Histogram of PctIn values for predicted transcripts. Three local peaks were observed at 0.333, 0.5 and 1.0 PctIn.

values of each exon across tissues. We designed baseline methods "PSI1" and "PSIX" to construct novel transcripts by selecting exon combinations with the highest product of PSI values. Similar to Rand1 and Rand$X$, PSI1 outputs one transcript per group and PSIX outputs k transcripts per $\mathcal{T}_M^k$ group. These PSI-based methods provide more realistic baselines by considering the likelihood of exon inclusion based on experimental data.

By considering isoforms identified from a real experiment (Alfonso-Gonzalez et al., 2023) as a reference set, we reasoned that the number and rate of isoform predictions matched by those isoforms are good approximations of the relative sensitivities and precisions of methods for isoform prediction. Here, we present a plot of matching rate versus number of matches for TENNIS and the four baseline methods (Fig. 2A).

Both randomized baselines have lower matching rates and fewer matched isoforms than TENNIS. The Rand1 baseline outputs 641 multi-exon isoform predictions, where, averaged over 5 replicates, only 156.6 (24.5%) are matched by long reads. The Rand$X$ baseline outputs 933 multi-exon isoform predictions, where 186.8 (20.0%) isoforms are matched. TENNIS has 640 (resp. 682) matched isoforms at the same matching-rate level as Rand1 (resp. Rand$X$) and matching rates of 46.6% (resp. 45.3%) at the same number of matched isoforms as Rand1 (resp. Rand$X$). The PSI-based baselines

substantially outperform the random baselines, where 232 (36.2%) PSI1 transcripts and 281 (30.1%) PSIX transcripts are matched, demonstrating that exon usage information improves transcript prediction. Nevertheless, TENNIS still outperforms the PSI-based methods. At equivalent numbers of matched isoforms, TENNIS achieves 8.6% higher matching rate than PSI1 and 28.6% higher than PSIX. At equivalent matching-rate levels, TENNIS achieves 36.2% more matching transcripts than PSI1 and 81.5% more than PSIX. Our evolution-inspired model captures additional information beyond exon-usage frequencies. As the PctIn (Percentage In) level of a predicted isoform is defined as the number of SAT solutions containing this isoform divided by the total number of solutions for that transcript group, a higher PctIn level indicates a higher confidence that the predicted isoform indeed is missing from the transcript group. We sorted TENNIS-predicted transcripts in descending order of their PctIn values, and then, if tied, by their splicing probability score.

At the PctIn level of 0.5 and 0.33, TENNIS reported matching rates/numbers of 41.4%/203 and 30.8%/447, respectively (circled points in Fig. 2). Additionally, at the two extremes, TENNIS reported a matching rate of 50% for the intersection of all potential solutions and 693 matched isoforms for the union of all potential solutions. These observations demonstrate that novel transcripts that have a higher chance of being from the evolution trajectory are more likely to be true positives, which supports the evolution-inspired model of TENNIS.

Although the PctIn values indeed range from 0 to 1, their distribution displays significant skewness with discrete peaks occurring at 0.333, 0.50, and a smaller local peak at 1.0 (Fig. 2B). Transcripts with PctIn values of 0.333 (resp. 0.50 or 1.0) may come from a transcript group T in which each has three (resp. two or one) optimal solutions from SAT. Note this concept is different from $\mathcal{T}_M^k$ which describes the number of missing isoforms. In other words, a transcript group T may need only one isoform to form a connected graph (thus, in $\mathcal{T}_M^1$), but may have two possible optimal configurations for this isoform by SAT. Correspondingly, transcripts with lower PctIn values are from a transcript group T with more optimal SAT solutions. The latter group is harder to solve, and predicted isoforms from such groups are less favorable. Transcripts with PctIn values of 0.333 (resp. 0.50 or 1.0) have a matching rate of 25% (resp. 40% or 51%), much higher than that of transcripts with lower PctIn values (9.6%). Therefore, we show that PctIn values of 0.5 and 0.333 can generally serve as two good thresholds for filtering TENNIS predictions. We also evaluated TENNIS performance stratified by the number of novel transcripts per group ($\mathcal{T}_M^1$, $\mathcal{T}_M^2$, etc.), showing that

groups requiring fewer novel isoforms achieve higher matching rates (Fig. S2).

TENNIS-predicted novel transcripts have adequate expression levels. We quantified the mean expression levels of matched TENNIS transcripts in several Nanopore long-read RNA-seq datasets from Alfonso-Gonzalez et al. (2023) (accession SRR19355640, SRR19355639, SRR19355636) using Oarfish (Zare Jousheghani et al., 2025). We compared the count per million (CPM) of TENNIS transcripts to CPMs of transcripts from the same transcript group. The mean TENNIS isoforms CPM to group total CPM ratio is 20%. Also, 45% of TENNIS transcripts have higher CPMs than their group average, and 90.4% have at least 10% of their group average CPM (Fig. S3).

We also investigated how many of the TENNIS-predicted transcripts preserve open reading frames (ORFs) of their original gene. ORFanage (Varabyou et al., 2023) was employed to compare TENNIS-predicted transcripts with dm6 reference annotation. More than 80% long-read-matched TENNIS transcripts have greater than 90% in-frame length percent identity (ILPI) with their best reference transcript, meaning 90% of the ORFs from the original gene were preserved in-frame, and more than 68% of all TENNIS-predicted transcripts (including those unmatched by long reads) preserve 90% of the original ORFs (Fig. S4). These results imply that TENNIS transcripts may constitute a substantial proportion of the transcriptome in real samples.

Notably, not all genes or transcripts are expressed, and not all expressed transcripts are captured by sequencing. Hence, using assemblies and quantification from real RNA-seq as a ground truth tends to underestimate the total number of true-positive genes and/or transcripts. In other words, transcript predictions unmatched by an assembly may be false-positive predictions or may be unexpressed or unsequenced in the experiments. To estimate the coverage of genes in our "ground-truth" (namely, the assembly from Alfonso-Gonzalez et al. (2023)), we compared it with dm6 annotations, in addition to TENNIS outputs. GffCompare (Pertea and Pertea, 2020) reported this long-read assembly overlaps with only 54.0% loci (based on exon overlapping) in dm6 annotation and 63.0% loci in TENNIS. Hence, the number of true positives is most likely underestimated for TENNIS to a noticeable level.

TENNIS accurately retrieves isoforms in a removal simulation

To further validate TENNIS's ability to detect missing transcripts, we conducted a simulation using a removal and retrieval approach. From genes containing three or more annotated isoforms, we

randomly removed one isoform. The removed isoform could not be the shortest isoform, and its removal could not reduce the total number of exons (i.e. the exon spliced out in all other isoforms) in the group, so that retrieval of this isoform is not impossible. This experimental design aimed to assess both the matching rate and the number of matched transcripts of TENNIS in identifying missing transcripts. GffCompare was used for evaluation and the removed transcripts are regarded as ground truth.

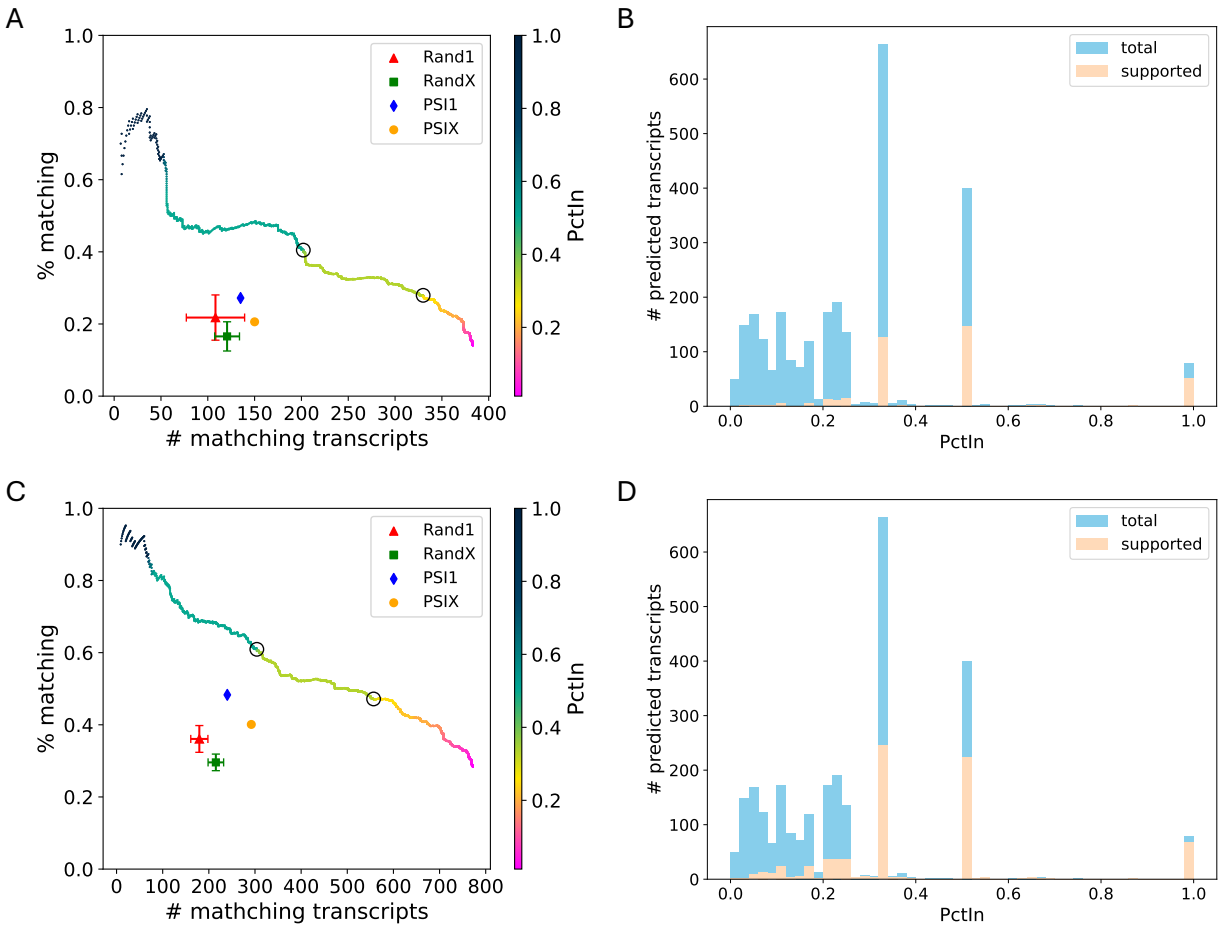

Figure 3: TENNIS outperforms baseline methods in transcript-retrieval simulations. (A) The number and percentage of transcript predictions validated against exactly removed isoforms. The color gradient indicates PctIn values, with circles highlighting critical thresholds (PctIn = 0.5, 0.333). Baseline methods include random (Rand1, Rand$X$) and PSI-based (PSI1, PSIX) approaches. Error bars represent standard deviation over 5 repetitions. (B) Histogram of PctIn values for predicted transcripts validated against exactly removed isoforms. (C) The number and percentage of transcript predictions validated against combined ground truth (exactly removed isoforms + long-read RNA-seq assembly), using the same color scheme as panel A. (D) Histogram of PctIn values for predicted transcripts validated against the combined ground truth.

A total of 796 multi-isoform groups were used for this removal simulation. TENNIS classified the 796 $\mathcal{T}_M$ groups to 262(33%) $\mathcal{T}_M^0$, 342(43%) $\mathcal{T}_M^1$, 102(13%) $\mathcal{T}_M^2$, 30(4%) $\mathcal{T}_M^3$, 23(3%) $\mathcal{T}_M^4$, and 37(5%) $\mathcal{T}_M^@$ groups. The percentages of all $\mathcal{T}_M^k$ classes increased, compared to Table 1. This is expected since we removed one isoform from each group. The presence of $\mathcal{T}_M^0$ groups indicates that some groups have a more "connected" graph and that not all non-terminal vertices are cut vertices. Besides, those 796 groups do not necessarily satisfy the evolution model prior to the removal. Therefore, the missing isoform identification problem is further complicated.

TENNIS achieved high matching rates and numbers of matched transcripts, which are considerably better than those of baseline approaches, in this simulated removal-and-retrieval experiment (Fig. 3A). At PctIn values of 0.5 or 0.333, TENNIS has a matching rate of 40.6% or 27.9% and 202 or 329 matched transcripts. The matching rates and numbers of matched transcripts for Rand1 are 21.8% and 108.2, while those for Rand$X$ are 16.6% and 120.6, averaged over 5 replicates. The matching rates and numbers of matched transcripts for PSI1 are 27.2% and 135, while those for PSIX are 20.6% and 150. At an equivalent matching-rate level or number of matched transcripts, TENNIS substantially outperforms all four baseline methods, as demonstrated by its curve lying above the baseline data points in Fig. 3A.

Considering the presence of multiple solutions and the potential incompleteness of annotations, we also evaluated the predictions using the combined ground truth, i.e. union of removed transcripts and the long-read RNA-seq assembly (Fig. 3C). Hence, previously predicted transcripts that are not matched by the exactly removed transcripts could potentially be validated by real sequencing data. At a PctIn level of 0.5 (resp. 0.333), TENNIS successfully predicts 304 (resp. 556) matched isoforms with a matching rate of 61.0% (resp. 47.1%). The baseline approaches Rand1 and Rand$X$ respectively identified only 179.4 and 215.2 matched isoforms with matching rates of 36.1% and 29.6%, averaged across 5 replicates. PSI1 had 240 (48.3%) transcripts matched and PSIX had 292 (40.1%) transcripts matched. At an equivalent matching-rate level or number of matched transcripts, the performance of TENNIS is again higher than that of those four baselines.

The distribution of PctIn values mirrors the pattern observed in real data, exhibiting local peaks at 0.33, 0.5, and 1.0 (Fig. 3B and D). The matching rates of isoforms with those PctIn values are 19%, 37%, 67% if validated by exact removed isoforms, and 37%, 56%, 87% if validated by combined ground truth.

We also evaluated TENNIS performance stratified by the number of novel transcripts per group ($\mathcal{T}_M^1$, $\mathcal{T}_M^2$, etc.). Similar to the case with real data, when a group requires fewer novel isoforms, the matching rates are higher (Fig. S5 and S6). Additionally, more complex simulation experiments were performed by extending the simulation to remove 2 or 3 transcripts per group, and TENNIS consistently outperformed all baseline methods in these more challenging scenarios (Fig. S7).

Cross-validation on human annotations

We investigated whether TENNIS predictions on one annotation can be validated using another annotation of the same species. We conducted this experiment using human GENCODE and RefSeq annotations. Since GffCompare considers multi-exon transcripts with identical intron-chains as "matching", we removed transcript groups that do not share identical "first-exon donor and last-exon acceptor" combinations between RefSeq and GENCODE, because TENNIS-predicted isoforms from those groups will never match with the other annotation.

We compared TENNIS-predicted transcripts from GENCODE (referred to as TENNIS_Gencode) with human RefSeq annotation. At the level of PctIn = 0.5, TENNIS_Gencode has 672 transcripts matched / 26.78% matching rate with RefSeq annotation. An additional 528 TENNIS_Gencode transcripts (21%) matched with TENNIS_RefSeq (TENNIS-predicted transcripts from RefSeq), meaning those 528 transcripts are potentially missing from both annotations and can be recovered by our method. Those results indicate that a substantial number of TENNIS-predicted transcripts can be validated using a different annotation dataset. Serving as a comparison baseline, 15.3% of multi-exon transcripts in GENCODE annotation and 42.5% in RefSeq annotation match each other.

Discussion

A comprehensive transcriptome annotation is essential for many bioinformatics and biomedical studies. While significant resources have been invested in improving these annotations through the invention of new methods, pipelines, and manual curations, the great majority of these annotations, if not all, are data-driven (Li et al., 2021; Aken et al., 2016; Pertea et al., 2018; Morales et al., 2022). Little attention has been paid to modeling the annotated isoforms particularly from an evolutionary perspective. We fill this critical gap with TENNIS, an evolution-inspired model for

characterizing annotated transcripts, together with an algorithm that infers missing isoforms in an annotation. The model of TENNIS is simple: isoforms in a transcript group are connected in a single component using the four basic AS events, provided no isoform is missing. When this condition is not satisfied, TENNIS seeks the minimum number of isoforms to make them connected, using a novel SAT formulation that guarantees to find all optimal solutions.

We analyzed seven transcriptome annotations of model organisms using TENNIS. We showed that the majority of transcript groups are single-isoform, accounting for about 80% of all multi-exon groups. The evolution-inspired model is satisfied by 62%-88% of transcript groups in various species' annotations, supporting the propositions of our model. We also evaluated the validity of TENNIS isoform predictions by comparing them with a long-read RNA-seq assembly and through a simulation experiment of removal and retrieval. In both settings, TENNIS outperformed the randomized and PSI-based baseline methods. After controlling the same number of matched transcripts or matching rate, TENNIS showed approximately a 70%-200% increase in matching rate and 250%-330% increase in the number of matched transcripts over the baselines in the experiments. Furthermore, the analysis revealed that if an isoform appears in multiple optimal solutions with a higher percentage (PctIn), then it is more likely to represent a true isoform. The PctIn metric can thus serve as an effective criterion for filtering predicted isoforms.

The missing isoforms identified by TENNIS should be interpreted as transcripts that are plausibly annotatable under our model. Their identification does not imply that these isoforms are actively expressed in extant species. Rather, it is entirely possible that some of these transcripts were expressed historically but have since been lost, silenced, or otherwise suppressed during evolution in specific lineages or clades. Empirical validation of these predicted isoforms should rely on experimental evidence, such as assembling and quantifying RNA-seq data.

While TENNIS is inspired by evolutionary principles, we note that it does not directly compare transcriptomes across species. Cross-species validation of predicted isoforms remains challenging because homologous genes may have divergent transcript structures and sequences. Genes that are highly conserved tend to have conserved transcripts across species, while less conserved genes may have different structures that are difficult to align. Nevertheless, TENNIS provides meaningful insights when comparing different annotations of the same species, as demonstrated by our cross-annotation validation between GENCODE and RefSeq. A direct cross-species isoform evolution

analysis would be an interesting direction for future work.

TENNIS is best applied when a gene has multiple isoforms with the same transcription start site (TSS) and transcription end site (TES). This condition ensures that the transcripts originate from the same pre-mRNA and their diversity results purely from alternative splicing. For under-annotated genomes, where many genes have only one annotated transcript and are hence categorized as $\mathcal{T}_S$, users may benefit from performing RNA-seq assembly prior to applying TENNIS, thereby leveraging sequencing data to first identify multiple isoforms per gene.

The assumption made by TENNIS is minimal, yet demonstrates strong prediction capability in identifying missing isoforms. It therefore holds great potential to model complex evolutionary trajectories. A promising future enhancement for TENNIS is the integration of additional prior knowledge and features related to (AS) events, such as the lengths and sequences of introns/exons, their splicing patterns, and conservation of ORFs of original genes. Previous studies have established that constitutive exons typically exhibit greater lengths and are flanked by shorter introns, while alternatively spliced exons are more likely to be shorter and accompanied by longer introns (Ast, 2004; Lev-Maor et al., 2007; Kim et al., 2007b). Also, ref. (Lev-Maor et al., 2007) revealed a positive correlation between expression level and evolutionarily conserved transcripts, which are often ancestral. As a mathematical model to characterize observed isoforms, TENNIS fills the research gap of AS evolution and provides valuable insights for various research areas including alternative splicing mechanisms, comparative transcriptomics, and phylo-transcriptomics studies.

Methods

An evolution-inspired model

TENNIS models the evolution of alternative splicing (AS) within a transcript group (defined below) based on two premises: (1) AS isoforms evolve sequentially, with each isoform being derived from a predecessor; and (2) each isoform must originate from its parent through a single AS event (CE, A3, A5, or IR) per evolutionary step. The rationale behind the second premise is that AS events arise independently through mutations in splicing sites or regulatory elements and it is less likely to have two mutations occur simultaneously (Ast, 2004; Zhang et al., 2017). Consequently, all isoforms of a transcript group should be connected via single AS events. If not, then it indicates that isoform(s)

are missing from the annotation or have lost function and therefore are not present in the current annotation.

The framework of TENNIS is as follows. It takes a transcriptome or an assembly, i.e., a set of annotated or assembled transcripts in GTF format, as input. It first partitions all transcripts into transcript groups (defined below). Within each group, it constructs the evolutionary relationship using a graph, determines evidence of missing isoforms, and if evidence is present, identifies the missing isoforms.

We focused on analyzing AS isoforms originating from the same pre-mRNA. That is, TENNIS groups transcripts that share the same alternative transcription start site (TSS) and alternative transcription end site (TES) together, referred to as a "transcript group". We denote by $\mathcal{T}_S$ the set of transcript groups with just a single transcript, and by $\mathcal{T}_M$ the set of transcript groups with two or more transcripts. Fig. 1B shows an example of a transcript group with 5 transcripts. Although TSS and TES are two events that also produce diverse transcripts, the pre-mRNAs are already different for transcripts with such events (Marasco and Kornblihtt, 2023). Hence, the AS processes are more different between transcripts with alternative TSS or TES (Alfonso-Gonzalez et al., 2023; Reyes and Huber, 2018).

Next, TENNIS builds a graph for each transcript group. In the graph, the collection of nodes represents all isoforms and the collection of edges represents whether two isoforms are convertible via a single AS event. For example, Fig. 1D illustrates the graph for transcripts in Fig. 1B. Details of the construction of graphs are described in the next subsection. We say that a transcript group does not present evidence of missing isoform(s) if the graph is connected (i.e., the group consists of a single connected component). Otherwise, TENNIS recruits a minimal number of additional nodes to make all components connected. These reconstructed nodes/transcripts are regarded as missing isoforms. This step is modeled as an optimization problem and solved by transforming it into a satisfiability (SAT) formulation, detailed in the TENNIS-SAT section below.

Constructing evolution trajectory and identifying missing isoforms

Let T be a transcript group. TENNIS encodes each isoform in T as a binary vector, depicting exonic or intronic regions. First, genomic coordinates of all splicing sites of all isoforms in T are collected and then the genome is split into smaller regions according to those coordinates (Fig. 1B). Let n

be the number of resulting genomic regions. Clearly, each exon or intron spans either one region or several consecutive regions. Hence, every isoform in T can be described using indices of spliced-in regions (i.e. exonic regions) and indices of spliced-out regions (i.e. intronic regions). Therefore, by encoding the exonic region as 1 and intronic regions as 0, an isoform is encoded as a length-n binary vector. For example, a 1 at position i of the binary vector means the i-th genomic region is covered by an exon in this isoform and vice versa (see Fig. 1C). Assume T contains m isoforms. Then T can be represented as an $m \times n$ binary matrix, denoted as M.

The benefits of binary encoding of isoforms are that, besides clarity and conciseness, all simple AS events can be represented as the flip of a bit or several consecutive bits. While an exon is split into multiple smaller regions (in this case, called partial-exons) due to A3/A5 events in another isoform, this exon is accordingly coded as multiple 1's. Partial-introns are defined likewise. Hence, the A3/A5 event can be represented as a flip of the bits of those corresponding partial-exons to partial-introns. CE and IR are also represented as 1-to-0 or 0-to-1 flips. Equivalently, all simple AS events can be considered as the inclusion or exclusion of one or multiple consecutive regions.

Given an $m \times n$ binary matrix M representing a transcript group, a graph will be constructed. Each annotated isoform is denoted as a vertex. An edge may be added between two vertices if their isoforms can convert to each other by one AS event, *i.e.* a flip of consecutive 0's to 1's or consecutive 1's to 0's. Edges are undirected, since 0-to-1 and 1-to-0 flips are symmetric. This also reflects the invertible property of the basic events.

Provided no transcript is missing, all vertices should be in one connected component. In this case, we say that transcript group T satisfies our evolution-inspired model, and call T a transcript group in $\mathcal{T}_M^0$. Otherwise, one or more isoforms are said to be missing from T. It is important to note that, due to the minimality of our model, neither direction of the reasoning is decisive, that is, it is possible that T misses some isoforms but the resulting graph remains connected, and it is also possible that the graph is not connected but T does not miss any unannotated isoform.

In the case that the graph contains more than one connected component, TENNIS will reconstruct missing isoforms. We formulate this task as an optimization problem, that is, to find a minimum number of isoforms such that adding them to T results in a graph with just one connected component. This is a parsimonious assumption — that a minimal number of isoforms are missing from the annotations. We design an algorithm, termed TENNIS-SAT (M, k), described in detail

in the next section. TENNIS-SAT takes matrix M and an integer $k \geq 1$ as input, and answers if adding k isoforms suffices to make the resulting graph connected, and if yes, also returns the binary representation of the k additional isoforms. Using TENNIS-SAT as a subroutine, starting with $k = 1$, TENNIS employs an iterative approach that calls TENNIS-SAT(M, k) in each iteration and increases k, until either the subroutine returns yes (and the k isoforms) or a maximum iteration number is reached. As a compromise of computational time and accuracy, the default maximum iterations, which is also the maximum number of missing isoforms that TENNIS attempts to reconstruct, is 4. According to our experiments, with this threshold, the model can explain more than 97% of the investigated transcript groups (Table 1). Transcript group T will be assigned to a category $\mathcal{T}_M^k$, if TENNIS determines that a minimum of k transcripts are missing from T, $k = 1, 2, 3, 4$. T will be assigned to category $\mathcal{T}_M^@$ if the maximum iteration is reached, which means T misses more than 4 transcripts, or TENNIS fails to finish in 15 minutes. Optionally, users can replace this fixed default with an MST-based per-group upper bound, as described later. Additional computational considerations on why conserved positions cannot simply be collapsed are discussed in Supplemental Note Suppl. Note S4.

It is common that multiple optimal solutions exist. This means, for a transcript group T in $\mathcal{T}_M^k$, different sets of k isoforms may make the resulting graph connected. For example, in Fig. 1D, if both t_2 and t_4 were missing, then either of them would be an optimal solution of size 1. TENNIS is able to return all optimal solutions. This offers an additional critical signal to decide if a constructed missing isoform is correct or not. The intuition is that if there are multiple optimal solutions, and an isoform appears in all of them, then it is more likely to be truly missed than isoforms that appear in only one solution. Therefore, for each reconstructed isoform in the union of all optimal solutions, we introduce a measure "Percentage In (PctIn)", defined as the number of solutions containing this isoform divided by the total number of solutions. In the above example, both t_2 and t_4 will be in the output with a PctIn value of 0.5.

TENNIS-SAT

Given an $m \times n$ binary matrix M representing all isoforms in a transcript group T, and an integer k representing the maximum number of missing isoforms allowed to be added, we use a SAT formulation to decide if adding k isoforms is sufficient to connect the graph. Similar to existing isoforms,

the unknown novel isoforms are represented as a vector of binary variables. For simplicity, they are appended to M, and all isoforms are represented by rows in M. This means that M_i is a length-n vector of known binary values for $i = 1, \cdots, m$, while M_i is a length-n vector of unknown binary variables for $i = m + 1, \cdots, m + k$.

Since the aim is to construct a connected graph, the presence of a spanning tree in the graph is necessary and sufficient. The spanning tree can be more efficiently represented in SAT by treating it as a rooted tree. In the constructed tree, each vertex denotes a row of M (i.e. one isoform) and this tree should have $m + k$ vertices and at most $m + k$ levels. Otherwise, the problem is infeasible. The high-level idea of the SAT formulation is trying to put each vertex, including both given and missing ones, to a certain level of the tree and construct their parent-child relationship. It is worth noting that such a parent-child relationship is solely for the convenience of construction, it does not indicate the direction of evolution — recall that our model is an undirected graph that primarily concerns about the presence/absence of isoforms, no effort has been made to infer the actual evolutionary trajectory.

We now provide the implementation details for the above idea. Recall that an SAT formulation consists of a set of boolean/binary variables and a conjunction of clauses where each clause is a disjunction of literals (boolean variables or their negations). We first introduce boolean variable $D_{i,j}$ to denote whether an edge exists between vertex i and vertex j, $i \neq j$. So $D_{i,j}$ is True if and only if the i-th isoform is derivable from the j-th isoform via exactly one simple AS event. Let a helper binary variable $d_{i,j,k}$ denote the number of (extra) events needed to convert $M_{i,k}$ from $M_{j,k}$ ($i \neq j$), i.e. flipping the bit of the k-th region. Since we only permit one AS event between direct parent-child isoforms, $D_{i,j}$ is set to True if and only if exactly one variable in $\{d_{i,j,k} \mid 1 \leq k \leq n\}$ is True. Enforcing the condition "exactly one variable in a set must be True" can be implemented as SAT clauses detailed in Suppl. Note S1.

Consider a simplified case when all exons are represented by exactly one region, *i.e.* no partial-exons. Then $d_{i,j,k}$ is set to True if and only if $M_{i,k} \neq M_{j,k}$ (Suppl. Note S2). However, when partial-exons exist, skipping multiple consecutive partial-exons is also regarded as one event because it takes the same number of splicing to skip one exon or multiple consecutive partial-exons (Fig 1E). Thus, we set $d_{i,j,k}$ to True, if and only if both conditions are true: (1) $M_{i,k} \neq M_{j,k}$; and (2) $M_{i,k} \neq M_{i,k-1}$ or $M_{j,k} \neq M_{j,k-1}$, $2 \leq k \leq n$; (second condition not required when $k = 1$). Otherwise the difference

between $M_{i,k}$ and $M_{j,k}$ has been compensated at or before position $k-1$, so the penalty should not be double-counted. Those two conditions can be modeled by clauses in Suppl. Note S3.

After properly representing edges with $D_{i,j}$, we can fit vertices into a tree. Let boolean variable $L_{i,j}$ denote whether vertex i is on level j of this tree. $L_{i,j}$ should satisfy the following constraints: First, a vertex appears exactly once in the tree, which means for the i-th isoform, exactly one of the variables in $\{L_{i,j} \mid \forall j\}$ is set to True. Second, exactly one vertex, *i.e.* the root, is on level 1, namely, exactly one variable in $\{L_{i,1} \mid \forall i\}$ is True. Both require the constraint "exactly one variable in a set is True", which again can be modeled with the approach in Suppl. Note S1.

Next, we add constraints governing the spanning tree edges. The idea is that if a vertex i is present at level g ($g \geq 2$), then there must exist a node j on level $g-1$ that has an edge connecting to vertex i, namely, $D_{i,j}$ is True. Let binary variable $C_{i,j,g}$ denote if vertex i is on level g of the tree and is preceded by vertex j on level $g-1$ through one simple AS event of edge $D_{i,j}$. Therefore, $C_{i,j,g}$ can only be set to True if all three variables $D_{i,j}$, $L_{i,g}$ and $L_{j,g-1}$ are True. However, the reverse direction does not always hold because $D_{i,j}$ may be true for different pairs of vertices i and j. Intuitively, a vertex i can have multiple potential parents in the graph, but we only choose one in the constructed spanning tree. These constraints can be modeled by 3 SAT clauses:

$$(\overline{C_{i,j,g}} \vee D_{i,j}) \bigwedge (\overline{C_{i,j,g}} \vee L_{i,g}) \bigwedge (\overline{C_{i,j,g}} \vee L_{j,g-1}).$$

Last, every vertex i must be either the root vertex in the spanning tree or located on level ≥ 2. So we have the following constraints for each i: exactly one variable from the set $\{L_{i,1}\} \cup \{C_{i,j,g} \mid g \geq 2, j \neq i\}$ is True. Again, Suppl. Note S1 models these constraints.

TENNIS implements the SAT formulation via the pySAT interface (Ignatiev et al., 2018) and solves the problems using the Glucose SAT solver (Audemard and Simon, 2018). We configure it to time out after 15 minutes to balance computational efficiency and accuracy.

Optimal greedy algorithm for two connected components

When a transcript group has exactly two connected components, an optimal solution can be computed using a simple greedy algorithm without resorting to the SAT formulation. This algorithm is both faster and guaranteed to find the minimum number of missing isoforms needed to connect

the two components.

Let C_1 and C_2 be the two connected components in the graph constructed from matrix M. The optimal greedy algorithm proceeds by first computing the pairwise distance $d(i, j)$ for all pairs where $i \in C_1$ and $j \in C_2$. Here, $d(i, j) = \sum_k d_{i,j,k}$ counts the number of AS events separating isoforms i and j, as defined in the TENNIS-SAT formulation above. Next, the algorithm identifies the minimum distance $d^* = \min_{i \in C_1, j \in C_2} d(i, j)$ across all such pairs ($d^* \geq 2$, otherwise C_1 and C_2 would be connected). The minimum number of missing isoforms needed to connect the two components is then $d^* - 1$. Finally, for each pair (i', j') achieving the minimum distance d^*, the algorithm enumerates all possible paths of intermediate isoforms connecting i' to j', where each path (excluding node i' and j') represents a distinct optimal solution. The minimal number of novel transcripts needed to connect the two components is the same as that needed to connect transcripts i' and j'. When partial-exons do not exist, we can connect i' and j' through a path of a series of single-event edges by sequentially flipping the d^* differing bits one at a time. This creates $d^* - 1$ intermediate isoforms, each differing from its neighbors by exactly one bit flip (representing one simple AS event), thereby yielding one optimal solution. If partial-exons exist (defined in the above formulation of TENNIS-SAT as consecutive bits), we flip differing consecutive bits instead of one bit each time. Hence, the above $d^* - 1$ intermediate-isoform solution still holds.

Upper bound computation for novel transcripts

To balance computational efficiency and performance, TENNIS uses a fixed upper bound of 4 novel isoforms by default. Optionally, the upper bound of the number of novel transcripts in a group can be computed by an algorithm based on the minimal spanning tree (MST).

We construct a weighted hypergraph H where each node represents a connected component, and edges between nodes are weighted by the minimum distance between any two isoforms from the respective components. A minimum spanning tree (MST) of H guarantees an efficient way to connect all components using an upper bound on the number of novel isoforms. If the MST has edges with weights $w_1, w_2, \ldots, w_{c-1}$ (where c is the number of connected components and w_i is at least 2), then the upper bound is $\sum_{i=1}^{c-1}(w_i - 1)$. This bound is guaranteed to be achievable because we can independently apply the two-component greedy algorithm to connect components according to the MST structure.

Computing the MST-based upper bounds has a few benefits. First, it enables per-group flexibility: rather than applying a single fixed global cutoff (e.g., $k \leq 4$), users can set tighter or looser iteration limits based on group-level complexity such as the number of annotated isoforms or exons. Second, when the bound is 1, the greedy algorithm is guaranteed to find the optimum and the SAT solver is bypassed entirely, saving computation. Third, when the SAT solver times out before converging, the precomputed bound characterizes the remaining search space: without it, a timeout leaves open whether the current k is simply infeasible or whether more time would suffice; with it, users can make an informed decision to relax the time limit or switch to the greedy solver. The greedy solver is attempted first for all cases; if it succeeds, the result is returned immediately. Otherwise, the computed upper bound guides the iterative SAT solver to focus on the feasible solution space.

Splicing probability scoring for predicted transcripts

To further refine the ranking of TENNIS-predicted isoforms, we computed a splicing probability score for each predicted transcript based on exon inclusion rates. This score quantifies the likelihood of an isoform given the included exons. The scoring scheme takes the predicted transcripts and the collected experimentally measured PSI (percent spliced-in) values of exons as inputs.

Given an $m \times n$ binary matrix M representing all isoforms in a transcript group T, let M_i represent the i-th isoform as a binary vector of length n, and let j index the exons such that $M_{i,j} \in \{0, 1\}$ indicates the inclusion of exon j. Let p_j denote the inclusion probability (PSI) of the j-th exon. The splicing probability $sp(M_i)$ is computed as:

$$sp(M_i) = \prod_{j=1}^{n} p_j^{M_{i,j}} (1 - p_j)^{(1-M_{i,j})}$$

This formulation treats the inclusion of each exon as a Bernoulli trial, where p_j is the probability of inclusion ($M_{i,j} = 1$) and $1 - p_j$ is the probability of exclusion ($M_{i,j} = 0$). Likewise, baseline methods PSI1 and PSIX aim to construct 1 or k novel isoforms whose $sp(\cdot)$ is maximized for $\mathcal{T}_M^k$ ($k \geq 1$) transcript groups.

Code availability

TENNIS is implemented in Python and is freely available as open-source software under the BSD-3-Clause license at `https://github.com/Shao-Group/tennis`. Scripts, documentation, and data descriptions for reproducing the experiments in this manuscript are available at `https://github.com/Shao-Group/tennis-test` and in the Supplemental Code file.

Competing interests

No competing interests were declared.

Acknowledgements

The authors thank Tasfia Zahin for help in data collection and processing. X.C.Z. and M.S. conceived of the project. X.C.Z., K.C., and M.S. designed the algorithm. X.C.Z. and K.C. implemented the software. X.C.Z. and I.M.K. performed the experiments. All authors analyzed the results. All authors wrote, reviewed, and approved the manuscript. M.S. supervised the project. This work is supported by the US National Science Foundation (2145171 to M.S.) and by the US National Institutes of Health (R01HG011065 to M.S.).

References

Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Banet JF, Billis K, Girón CG, Hourlier T, et al.. 2016. The Ensembl gene annotation system. *Database: The Journal of Biological Databases and Curation* **2016**: baw093.

Alfonso-Gonzalez C, Legnini I, Holec S, Arrigoni L, Ozbulut HC, Mateos F, Koppstein D, Rybak-Wolf A, Bönisch U, Rajewsky N, et al.. 2023. Sites of transcription initiation drive mRNA isoform selection. *Cell* **186**: 2438–2455.e22.

Ast G. 2004. How did alternative splicing evolve? *Nature Reviews Genetics* **5**: 773–782.

Audemard G and Simon L. 2018. On the glucose sat solver. *International Journal on Artificial Intelligence Tools* **27**: 1840001.

586 Birzele F, Csaba G, and Zimmer R. 2008. Alternative splicing and protein structure evolution.
587 *Nucleic Acids Research* **36**: 550–558.

588 Calvo-Roitberg E, Carroll CL, Kim G, Sanabria V, Venev SV, Mick ST, Paquette JD, Uriostegui-
589 Arcos M, Dekker J, Fiszbein A, et al.. 2025. mRNA initiation and termination are spatially
590 coordinated. *Science* **390**: eado8279.

591 Chen M and Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from molecular
592 and genomics approaches. *Nature Reviews Molecular Cell Biology* **10**: 741–754.

593 Hoyos LE and Abdel-Wahab O. 2018. Cancer-specific splicing changes and the potential for splicing-
594 derived neoantigens. *Cancer Cell* **34**: 181–183.

595 Ignatiev A, Morgado A, and Marques-Silva J. 2018. PySAT: A Python toolkit for prototyping with
596 SAT oracles. In *SAT*, pp. 428–437.

597 Keren H, Lev-Maor G, and Ast G. 2010. Alternative splicing and evolution: diversification, exon
598 definition and function. *Nature Reviews Genetics* **11**: 345–355.

599 Kim E, Goren A, and Ast G. 2007a. Alternative splicing: current perspectives. *BioEssays* **30**:
600 38–47.

601 Kim E, Magen A, and Ast G. 2007b. Different levels of alternative splicing among eukaryotes.
602 *Nucleic Acids Research* **35**: 125–131.

603 Leung SK, Jeffries AR, Castanho I, Jordan BT, Moore K, Davies JP, Dempster EL, Bray NJ, O'Neill
604 P, Tseng E, et al.. 2021. Full-length transcript sequencing of human and mouse cerebral cortex
605 identifies widespread isoform diversity and alternative splicing. *Cell Reports* **37**: 110022.

606 Lev-Maor G, Goren A, Sela N, Kim E, Keren H, Doron-Faigenboim A, Leibman-Barak S, Pupko T,
607 and Ast G. 2007. The "Alternative" Choice of Constitutive Exons throughout Evolution. *PLOS*
608 *Genetics* **3**: e203.

609 Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F,
610 Derbyshire MK, Durkin AS, et al.. 2021. RefSeq: Expanding the Prokaryotic Genome Annotation
611 Pipeline reach with protein family model curation. *Nucleic Acids Research* **49**: D1020–D1028.

612 Marasco LE and Kornblihtt AR. 2023. The physiology of alternative splicing. *Nature Reviews*
613 *Molecular Cell Biology* **24**: 242–254.

614 Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen LL, Chen R, Dean C, Dinger
615 ME, Fitzgerald KA, et al.. 2023. Long non-coding RNAs: Definitions, functions, challenges and
616 recommendations. *Nature Reviews Molecular Cell Biology* **24**: 430–447.

617 Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, Cox E, Davidson C, Ermolaeva
618 O, Farrell CM, et al.. 2022. A joint NCBI and EMBL-EBI transcript set for clinical genomics
619 and research. *Nature* **604**: 310–315.

620 NCBI. n.d. The NCBI Eukaryotic Genome Annotation Pipeline.
621 https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/process/. Accessed: 2024-10-22.

622 Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N,
623 Uralsky L, Gershman A, et al.. 2022. The complete sequence of a human genome. *Science* **376**:
624 44–53.

625 Pan Q, Shai O, Lee LJ, Frey BJ, and Blencowe BJ. 2008. Deep surveying of alternative splicing
626 complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40**:
627 1413–1415.

628 Pertea G and Pertea M. 2020. GFF utilities: GffRead and GffCompare. *F1000 Research* **9**.

629 Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang YC, Madugundu AK, Pandey
630 A, and Salzberg SL. 2018. CHESS: a new human gene catalog curated from thousands of large-
631 scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biology* **19**:
632 208.

633 Raghavan V, Kraft L, Mesny F, and Rigerte L. 2022. A simple guide to de novo transcriptome
634 assembly and annotation. *Briefings in Bioinformatics* **23**: bbab563.

635 Reyes A and Huber W. 2018. Alternative start and termination sites of transcription drive most
636 transcript isoform differences across human tissues. *Nucleic Acids Research* **46**: 582–592.

Salzberg SL. 2019. Next-generation genome annotation: We still struggle to get it right. *Genome Biology* **20**: 92.

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, et al.. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **27**: 849.

Scotti MM and Swanson MS. 2015. Rna mis-splicing in disease. *Nature Reviews Genetics* **17**: 19–32.

Singh P and Ahi EP. 2022. The importance of alternative splicing in adaptive evolution. *Molecular Ecology* **31**: 1928–1938.

Statello L, Guo CJ, Chen LL, and Huarte M. 2021. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology* **22**: 96–118.

Sugnet CW, Kent WJ, Ares M, and Haussler D. 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* pp. 66–77.

Tao Y, Zhang Q, Wang H, Yang X, and Mu H. 2024. Alternative splicing and related rna binding proteins in human health and disease. *Signal Transduction and Targeted Therapy* **9**.

Tapial J, Ha KC, Sterne-Weiler T, Gohr A, Braunschweig U, Hermoso-Pulido A, Quesnel-Vallières M, Permanyer J, Sodaei R, Marquez Y, et al.. 2017. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Research* **27**: 1759–1768.

Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, Kariyawasam H, Du MRM, Schuster J, Wang C, et al.. 2021. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biology* **22**: 310.

Varabyou A, Erdogdu B, Salzberg SL, and Pertea M. 2023. Investigating open reading frames in known and novel transcripts using ORFanage. *Nature Computational Science* **3**: 700–708.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, and Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476.

Wang Y, Liu J, Huang B, Xu YM, Li J, Huang LF, Lin J, Zhang J, Min QH, Yang WM, et al.. 2014. Mechanism of alternative splicing and its regulation. *Biomedical Reports* **3**: 152–158.

Wright CJ, Smith CWJ, and Jiggins CD. 2022. Alternative splicing as a source of phenotypic diversity. *Nature Reviews Genetics* pp. 1–14.

Zare Jousheghani Z, Singh NP, and Patro R. 2025. Oarfish: Enhanced probabilistic modeling leads to improved accuracy in long read transcriptome quantification. *Bioinformatics* **41**: i304–i313.

Zerbino DR, Frankish A, and Flicek P. 2020. Progress, Challenges, and Surprises in Annotating the Human Genome. *Annual review of genomics and human genetics* **21**: 55.

Zhang D, Guelfi S, Garcia-Ruiz S, Costa B, Reynolds RH, D'Sa K, Liu W, Courtin T, Peterson A, Jaffe AE, et al.. 2020. Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Science Advances* **6**: eaay8299.

Zhang SJ, Wang C, Yan S, Fu A, Luan X, Li Y, Sunny Shen Q, Zhong X, Chen JY, Wang X, et al.. 2017. Isoform Evolution in Primates through Independent Combination of Alternative RNA Processing Events. *Molecular Biology and Evolution* **34**: 2453–2468.

Zhang W, Guenther A, Gao Y, Ullrich K, Huettel B, and Tautz D. 2024. Plasticity and evolutionary dynamics of alternative RNA splicing.