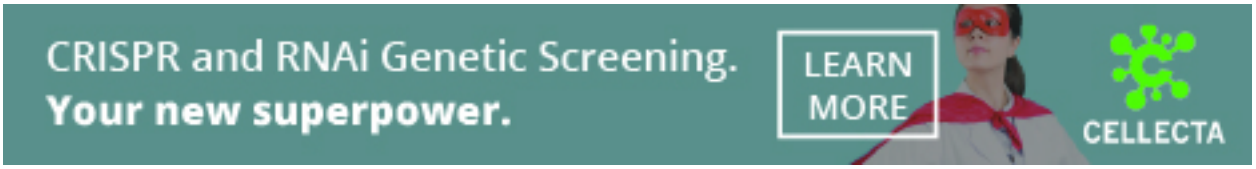
Exon Nomenclature And Classification of Transcripts (ENACT) provides a systematic framework to annotate exon attributes

Paras Verma, Deeksha Thakur, Deepanshi Awasthi, et al.

Genome Res. 2025 35: 1440-1455 originally published online May 7, 2025
Access the most recent version at doi:10.1101/gr.279878.124



Method

Exon Nomenclature And Classification of Transcripts (ENACT) provides a systematic framework to annotate exon attributes

Paras Verma, Deeksha Thakur, Deepanshi Awasthi, and Shashi Bhushan Pandit

Bioinformatics Center, Department of Biological Sciences, Indian Institute of Science Education and Research (IISER) Mohali, Knowledge City, Sector-81, SAS Nagar 140306, India

Isoform diversity is known to enhance a gene's functional repertoire by producing protein variants with distinct functional implications. Despite numerous studies on transcriptome diversifying processes (alternative splicing/transcription), understanding their extent and correlated impact on proteome diversity remains limited owing to dearth of subsequent proteogenomic consequences. To coalesce the genomic information embedded in exons with isoform sequences, we present an innovative framework, "Exon Nomenclature And Classification of Transcripts" (ENACT). This centralizes exonic loci such that protein sequence information is integrated (onto the available/annotated or new transcripts) while enabling tracking and assessing splice-site variability through unique yielded descriptors. The resulting annotation from the ENACT framework enables exon features to be tractable, facilitating a systematic analysis of isoform diversity. Our findings and case studies unveil systemic exon inclusion roles in regulating diversity in coding region. Correspondingly, annotation of protein-coding genes and associated transcripts from *C. elegans*, *D. melanogaster*, *D. rerio*, *M. musculus*, and *H. sapiens* are publicly accessible in a dedicated resource.

[Supplemental material is available for this article.]

Gene architecture in eukaryotes facilitates generation of more than one mRNA per gene through a differential combination of exons. The exonic region determination is under regulation of several co- and post-transcriptional mechanisms, in which alternative splicing (AS) plays a central role (Baralle and Giudice 2017; Tapial et al. 2017; Verta and Jacobs 2022); this, along with alternative transcription (ATR) and translation (ATL) mechanisms, drives transcriptome and proteome diversity. Among these processes, AS is the most widely studied, which enables the inclusion or exclusion of specific exons, creating multiple splice variants within isoforms. ATR, on the other hand, introduces or limits exon content through different promoters or terminators, affecting the 5′ or 3′ end of isoforms (Ni et al. 2010; Kamieniarz-Gdula and Proudfoot 2019). ATL mechanisms occur during and after transcription and contribute to proteome diversity through processes such as leaky scanning, reinitiation, inclusion of upstream open reading frame (uORF) usage, and utilization of varied ribosomal entry sites (Kochetov 2008; Lee et al. 2012; Tamarkin-Ben-Harush et al. 2014; Johnstone et al. 2016). Although these processes collectively shape transcriptome and proteome diversity, determining their combined or individual impact is challenging and becomes more so in organisms with advanced complexity like in humans, in which only about one-third of exons are protein coding (Aspden et al. 2023). Notably, ATR contributes approximately fourfold higher varying nucleotides to the transcript region than AS (Shabalina et al. 2014), complicating the assessment of their impact on coding sequence (CDS). These complexities are not limited to humans but extend to mice, in which splicing patterns from noncoding regions were recapitulated for human Chr 21 (Deveson et al. 2018).

Considering the multifaceted roles of eukaryotic gene architecture in influencing transcriptional and translational processes, previous studies unveiled several attributes concerning the evolution of introns and exons, in which, specifically, a reduction in exon length and an increase in their count were noted with the rise in organismal complexity (Zhu et al. 2009; Koralewski and Krutovsky 2011; Movassat et al. 2019). Importantly, evolutionary studies suggest that the splicing process evolves faster than gene expression and often diverges from gene expression to facilitate distinct functions, thereby enabling faster adaptation in species (Barbosa-Morais et al. 2012; Merkin et al. 2012). However, most previous studies focused primarily on investigating AS, with a limited emphasis on ATR, ATL, or their combined effects on the evolution of exonic attributes and proteome variations. Although extensive insights have been gained into AS at the transcriptome level (Wang et al. 2023; Zhao et al. 2023), its footprint at the proteome level has lagged largely owing to experimental limitations. This disparity might result from different sensitivity requirements to quantify proteins and their abundances, as previous discussed and reviewed (Blencowe 2017; Tress et al. 2017; Manuel et al. 2023). Investigations into the impact of AS on proteome have highlighted that AS introduces disordered protein regions and enables tissue-specific identities (Buljan et al. 2012). However, subsequent studies, such as the one by Reyes and Huber (2018), showed that tissue-specific signatures are driven primarily by ATR rather than AS. These findings, along with evidence that one-third of exons serve as CDS (Aspden et al. 2023) and that the majority of alternative nucleotides are introduced within or near untranslated

Corresponding author: shashibp@iisermohali.ac.in
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279878.124.

35:1440–1455 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/25; www.genome.org

regions (UTRs) (Shabalina et al. 2010, 2014), prompt questions about the extent of AS contributions in diversifying the proteome.

Existing approaches to analyze splicing and associated events have relied predominantly on pairwise comparisons of transcript architectures. For instance, the numeric symbolic notation has been employed by Foissac and Sammeth (2007) and Sammeth et al. (2008) to identify AS events and subtypes, whereas splice graph-based data structures have been employed by Xing et al. (2006) and Vaquero-Garcia et al. (2016) to uncover similarity and complex events in transcriptome data. Modifications of associated approaches, specifically in the splice graphs, have been developed to uncover the splicing complexity (Vaquero-Garcia et al. 2016), discover unique exon junctions, and improve computational efficiency in transcriptomic analyses (Sterne-Weiler et al. 2018). Although such approaches are robust in efficiently elucidating AS events, they primarily focus on their detection and not on distinguishing them for CDS, UTR, and their intervening regions. This distinction is important for understanding these regions' unique roles in proteome and transcriptome diversity, as emphasized in the literature (Shabalina et al. 2014). The lack of exons' regional specification, particularly CDS, also points to the underexplored role of ATL in shaping exonic attributes. Moreover, existing approaches have primarily focused on pairwise notations rather than simplifying polymorphic representation of exonic loci at gene level as these undergo modifications from AS and related processes. Given that these processes impact yielded proteins, which are essential phenotypic determinants with a pivotal role in cellular complexity, it is crucial to understand the basis of how their encoding exons are influenced to diversify proteome. These findings underscore the need for tools and methods that enable comprehensive study of splicing at both gene-specific and genome-wide levels. Recognizing this need, Reixachs-Solé and Eyras (2022) emphasized the importance of systematically integrating transcriptome and proteome data.

To achieve the integration of this distinct information, we developed the exon nomenclature and classification of transcripts (ENACT) framework. ENACT's systematic translation-focused, exon-centric design streamlines the computational tracking of exonic loci while facilitating the manual and automatic interpretation of their features. The attributed exonic loci facilitate the inference of AS events and associated processes, including splice-site variations, coding/noncoding regional distinction, and embedded variation in protein-coding potential. Through comprehensive accommodation of intron–exon definitions with protein sequence information, ENACT addresses the intricacies promulgated by AS, alternate transcription, and alternate translation processes and enables a detailed assessment of how resulting exon variations influence proteome diversity.

Results

Exon entity description in ENACT

The ENACT framework centralizes exon entities by processing isoform exon compositions, extracting and associating attributes from isoforms to exons, and defining the exon architecture of a gene to comprehensively represent isoforms. As exons in isoforms would overlap, we first establish a reference set of nonoverlapping exons (*RSOEx*) and assign them ordinal positions. These will serve as anchors and enable discernment of overlapped exons or polymorph variants concerning their genomic coordinates (see Methods). Implementing the former is nontrivial owing to the preponderance of alternate exons in the mammalian genome and multiple splice-site variations involving complex combinations. The ENACT algorithm systematically addresses these splicing complexities through exon architecture representation for genes after processing exonic loci in isoforms and their variations (see Methods) (Box 1).

Subsequently, for assignment of ordinal position to exons, the procedure characterizes exons for their occurrence, splice-site variations, and isoform-specific amino acid contributions. These feature characteristics are recorded while ENACT processes isoform exon composition (see Methods). As each exon entity in our framework is primarily characterized by its genomic coordinates, changes at 5′ or 3′ or both ends will yield its consideration as a new entity; however, its relatedness is maintained through the commonly assigned ordinal position (see Methods). The feature characterization and ordinal position are embedded into a six-character alphanumeric Exon Unique IDentifier (EUID) (Fig. 1A), which is assigned to every nonredundant exonic entity, defined from its genomic coordinates. With an objective for easy inferential characterization of exon and its transcript participation, EUID can be categorized into three blocks by referencing position as $EUID^k$, where k ranges from one to six (see Methods, section "EUID construction"). The blocks constitute the following:

1. Exon translational feature (Block-I). This block includes $EUID^1$ and $EUID^2$ and is referred to as the exon's protein-coding global and local scope. The global scope describes whether an exon is coding ("T") or noncoding ("U") based on the presence or absence of coding genomic coordinates in isoforms (Fig. 1A). Some exons can be noncoding in some transcripts while coding in others, and these are classified with global scope value "D," noting dual ("D") exons (Fig. 1A, top). Other global scope descriptors are described in the Methods section. Briefly, these include "M" and "R," where "M" corresponds to an exon with coding scope; however, coding genomic coordinate is of only 1 nucleotide (nt), and it cannot contribute amino acid sequence on its own (see Methods, section "Amino acid coding (translational) attribute of exons defined in *RSOEx* and $Exon_{variants}$"), whereas "R" corresponds to an intron retention (IR) exon (see Methods, section "Intron retention").

   Local scope tracks and annotates variations in amino acid sequence, which can arise owing to frameshift or alternate translation initiation/termination. This feature is denoted using a numeric descriptor and accounts for amino acid sequence variations contributed by an exon entity. Notably, the local scope tracks sequence contribution and its variations within an exon entity but does not account for variations across splice variants. In Figure 1B, the splice variants of "Ref" exon 2 are shown as individual entities, with each adopting independent distinct Block-I features. Splice variants' different local scopes do not infer protein sequence variations from reference exon 2; however, their relation from reference is only indicated from the identical ordinal position.
2. Exon's prevalence feature and ordinal position (Block-II). This block includes $EUID^3$ and $EUID^4$, providing information on the prevalence feature and exonic ordinal position, respectively. As splice variants of exons are tracked relative to reference exon's ordinal position, an exon entity is called constitutive ("G") when it occurs without variants in all protein-coding transcripts, constitutive-like ("F") when exon and its variants appear in all transcripts, or alternate ("A") otherwise (Fig. 1A, middle).

Box 1. ENACT Algorithm

Functions:

$len(exon^l)$ = *nucleotide length of* l^{th} *exon*
$genSort(x) = [x_1, x_2, x_3, \ldots, x_n]$ *where* $n = |x|$, *AlphanumericSort*(x)
$sorted(ExSet) = [exon^1, exon^2, \ldots, exon^n]$ *where* $n = |Exset|$, $gc(exon^1) \le gc(exon^2) \le \ldots \le gc(exon^n)$

$$maxexon(isfs) = \left\{ k : isf_k = \max_i (EXONCod_i) \right\}$$

$$checkOL(i, j) = \begin{cases} 1, & (s_i < e_j) \wedge (e_i < s_j), \\ 0, & \end{cases}$$

selectExon(*GoExList*):
  $exons\_grt\_30 = \{elem \in GoExList \text{ if } len(elem) > 29\}$
  $\begin{cases} qualifierexon = \text{argmin}\ (exons\_grt\_30), \text{ if } exons\_grt\_30 \neq \emptyset \\ qualifierexon = \text{argmax}\ (GoExList), \text{ if } exons\_grt\_30 = \emptyset \end{cases}$
  return(*qualifierexon*)

defineSuboverlapExons(*NolEx*):
  NolEx = *lengthSort*(*NolEx*); *k* = 1
  while (*NolEx* ≠ ∅){
    GoEx[*k*] = *NolEx*[1]
    GoEx[*k*] = {*GoEx*[*k*] ∪ *NolEx*[*i*] |∀*i* *NolEx*[2:], *if checkOL*(*GoEx*[*k*], *NolEx*[*i*]) = 1}
    NolEx = *NolEx* − *GoEx*[*k*];
    k = *k* + 1
  return (*GoEx*)

ENACT main set definitions:

ISF_{gene} = {*NP_∗ and XP_∗ listed isoforms of a gene*}
$ISFCurated$ = {*isf* ∈ ISF_{gene}: *isf is NP_∗ isoforms*}
$EXON_k$ = {*exon* : *exon is part of isoform k, where* $k \in ISF_{gene}$}
$EXONCod_k$ = {*exon* : *exon* ∈ $EXON_k$ ∧ *coding exons of isoform k,* $k \in ISF_{gene}$}
$gc(exon^l) = (s_l, e_l)$, s_l *and* e_l *are start and end genomic coordinates of* l^{th} *exon, respectively*

$$RISO = \begin{cases} risf = maxexon(genSort(ISFCurated)), & |ISFCurated| > 0 \\ risf = maxexon(genSort(ISF_{gene})), & |ISFCurated| = 0 \end{cases}$$

$Exon_{RISO}$ = {*exon*: *exon* ∈ $EXON_k$, *where* $k = RISO$}
$RSOEx_{RISO} = Exon_{RISO}$
$NRExon$ = {*exon* : *exon* ∈ $EXON_k$, $k = 1$, $|ISF_{gene}|$ *and nonredundant for* $gc(exon)$}
$NolEx = \{exon_{nol} \subset NRExon_{IsfSet}: exon_{nol} = j, \text{ if } checkOL(i, j) = 0, \forall i \in sorted(Exon_{RISO}), \forall j \in sorted(NRExon_{IsfSet})\}$
$OlEx = \{exon_{ol} \subset NRExon_{IsfSet}: exon_{ol} = j, \text{ if } checkOL(i, j) = 1, \forall i \in sorted(Exon_{RISO}), \forall j \in sorted(NRExon_{IsfSet})\}$
$Exon_{variants} = OlEx$
GoEx = *defineSuboverlapExons*(*NolEx*)
$NolExA = \{GoE[k]: |GoEx[k]| = 1, \forall k = 1 \ldots |GoEx|\}$
$NolExB = \{GoEx[k]: |GoEx| > 1, \forall k = 1 \ldots |GoEx|\}$
$Qualifier_{exon} = \{Qexon: Qexon = selectExon(NolExB[i]), \forall i \in (1 \ldots |NolExB|)\}$
$RSOEx = \{Exon_{RISO} \cap NolExA \cap Qualifier_{exon}\}$
$Exon_{variants} = \{Exon_{OlEx} \cup (NolExB - Qualifier_{exon})\}$

3. Exon splice-site variability (Block-III). This block includes EUID[5] and EUID[6] and denotes alternate splice-site subtypes and their unique occurrence, respectively. Because splice variants of an exon share the same ordinal index, this block specifies the nature of splice-site variation. We denote "n," "c," and "b" codes to indicate variation in 5′, 3′, and both splice-sites, respectively (Fig. 1A, bottom). Their unique recurrence specific for subtypes (n/c/b) is captured as a count in the last character of EUID.

These block attributes in the EUID descriptor comprehensively characterize exon entities. Although most exons and their variants in protein-coding transcripts are annotated using this six-character EUID, specific exon instances that retain introns (intron retention [IR]) require a special EUID notation (see Methods). Thus, ENACT annotates all exon occurrences in a gene.

Although exonic features can be informed from specific block attributes, their combination provides a platform to describe various complex scenarios of exon variations. Because Block-II's ordinal index and Block-III's splice-site attribute define an exon entity, every unique splice-site variant is considered a distinct exon instance. These splice-site variations can have different protein-coding Block-I scopes, as illustrated in Figure 1B, which shows them as independent entities that can acquire different protein-coding features from their reference. For example, the reference exon is coding (T.1.*.2.0.0) in Figure 1B; however, its splice variants can have variable translational features of either coding, noncoding, or dual state depending on their properties in encoded transcripts. Figure 1C illustrates this variability in four human genes, showing that splice variants adopt different coding attributes.

Therefore, the EUID system facilitates distinguishing various translational features introduced by exonic variations at the transcript level while tracking their position within the gene. These exon features and block attributes can be extracted as shown above, and we will use them in subsequent sections to inform events involving exonic variations.

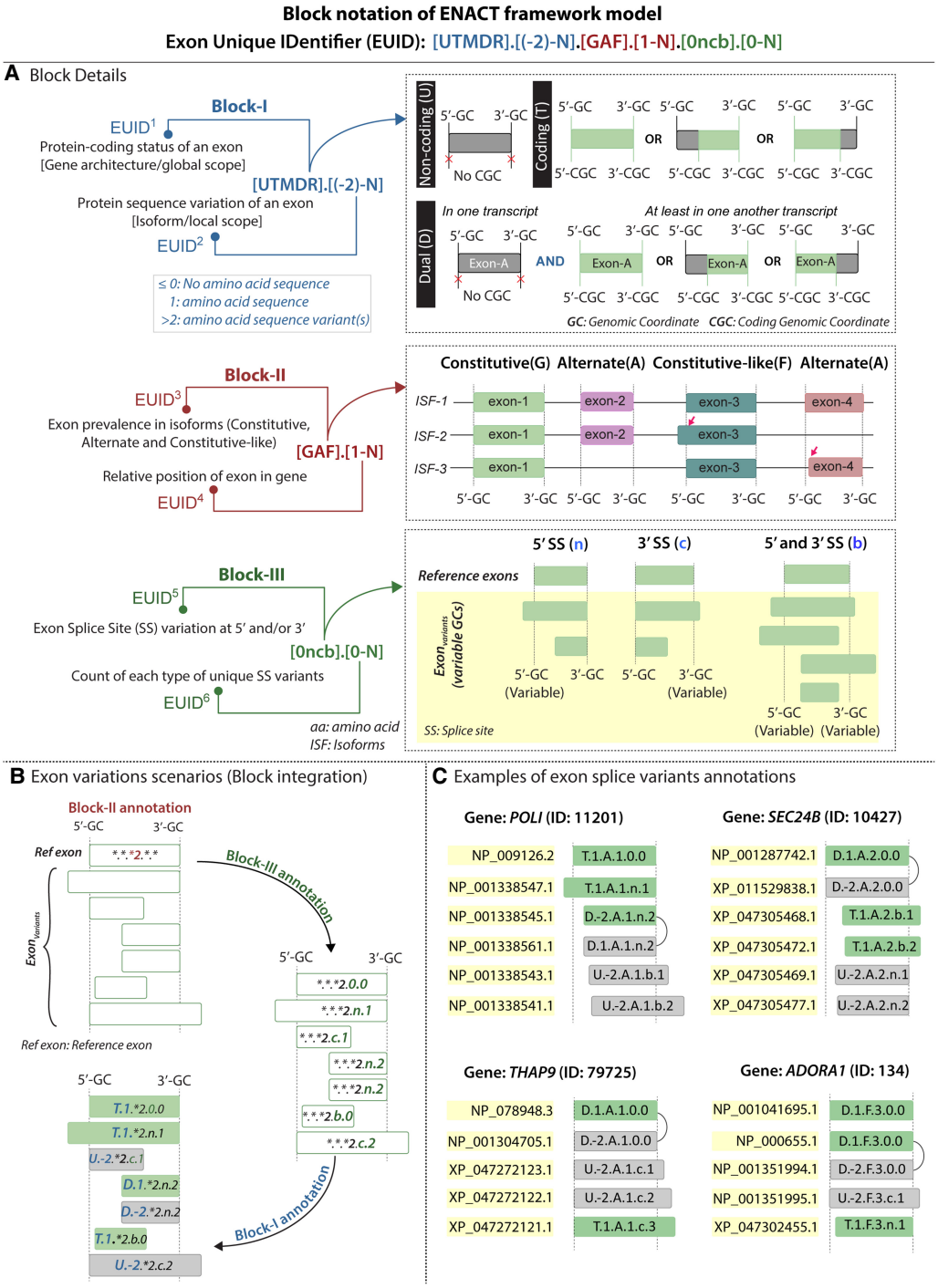


Figure 1. Overview of ENACT exon nomenclature. (*A*) Blocks of EUID represent several exon feature variations. Block-I global scope defines translational features of exons with unchanged genomic coordinates (GCs) by considering their coding genomic coordinates (CGCs). Exons are depicted as having coding, noncoding, and dual scope, where the "dual" is assigned to exons that take part in both coding and noncoding states in different isoforms. Block-I local scope depicts amino acid sequence contributions, where values of greater than one inform sequence variation from an exon and values of zero or less inform no amino acid sequence. The latter category involves (1) UTR exons that have no CGC and no amino acid assigned, depicted as "−2," (2) exons having only 1 nt in CGC, which cannot contribute amino acid independently, depicted by local scope "0" and often global scope "M," and (3) exons having >1 nt in CGC but no assigned amino acid, albeit rare in occurrence depicted by local scope value "−1." Block-II represents exon occurrence as "constitutive" ("G"), occurrence in all isoforms with unchanged splice sites; "facultative" ("F"), occurrence in all isoforms with varying splice sites; and otherwise "alternate" ("A"). Different splice-site variants (5′, 3′, or both) are annotated with n, c, and b instances, tracked numerically for unique variants, and constituting Block-III in EUID. "N" in EUID descriptor (*top* row) is numeric character denoting occurrence count for respective block attributes, and "ISF" corresponds to isoform. (*B*) Exon annotation with each block is illustrated, focusing on transitions between coding, noncoding, and dual states from the reference state (coding) after it undergoes splice-site variations. Gray and light green colors indicate the exon's coding and noncoding scope. The GC and CGC are shown with the dashed line. An asterisk indicates that appropriate code will be inserted based on relevant information. (*C*) Variants exhibiting changes in Block-I scope owing to splice-site variation at respective exon positions are shown for genes *POLI*, *SEC24B*, *THAP9*, and *ADORA1* (from *left* to *right*). Exon variants undergoing "dual" scope changes are specified with both "coding" and noncoding subtypes, connected by an arrow on *right*. The ID mentioned in brackets refers to Entrez gene identifiers. Corresponding abbreviations used in figure are expanded at their respective positions.

Interpretation of ATR, ATL, and AS events

ENACT framework facilitates the featurization of exons and their attribute extraction, as described before. This embedding enables systematic assessment of intra-transcript exonic composition as impacted by ATR, ATL, and AS. To illustrate these event depictions through ENACT, we focus primarily on comparing EUIDs between human gene transcripts. Additionally, we have demonstrated a hypothetical gene (Fig. 2), as it is difficult to find all such events occurring within a single gene. In Figure 2, top row shows its RSOEx and panels A through D represent exons to illustrate splicing events.

The hypothetical gene's exon architecture depiction from ENACT shows that it comprises 10 exons (Fig. 2). The Block-I global scope (EUID[1]) for assigned EUIDs indicates that three of 10 exons are noncoding with EUID[1]: "U" (**U**.-2.G.1.0.0, **U**.-2.F.9.0.0, and **U**.-2.A.10.0.0). Of the rest, six belong to the coding scope with EUID[1]: "T" (**T**.1.F.3.0.0, **T**.1.A.4.0.0, **T**.1.F.5.0.0, **T**.1.A.6.0.0, **T**.1.A.7.0.0, **T**.1.G.8.0.0), and one belongs to "dual" state with EUID[1]: "D" (**D**.-2.A.2.0.0). To note their splice-site variability, Block-II informs that two exons are "constitutive" at positions 1 and 8 with a prevalence feature (EUID[3]) of value "G" (U.-2.**G**.1.0.0, T.1.**G**.8.0.0); three are "constitutive-like" at positions 3, 5, and 9 with value "F" (T.1.**F.3**.0.0, T.1.**F.5**.0.0, U.-2.**F.9**.0.0); and other five are alternate at positions 2, 4, 6, 7, and 10 with value "A" (D.-2.**A.2**.0.0, T.1.**A.4**.0.0, T.1.**A.6**.0.0, T.1.**A.7**.0.0, U.-2.**A.10**.0.0). The splice variants for alternate exons (EUID[3]: "A") at positions 2, 4, and 7 and for constitutive-like exons

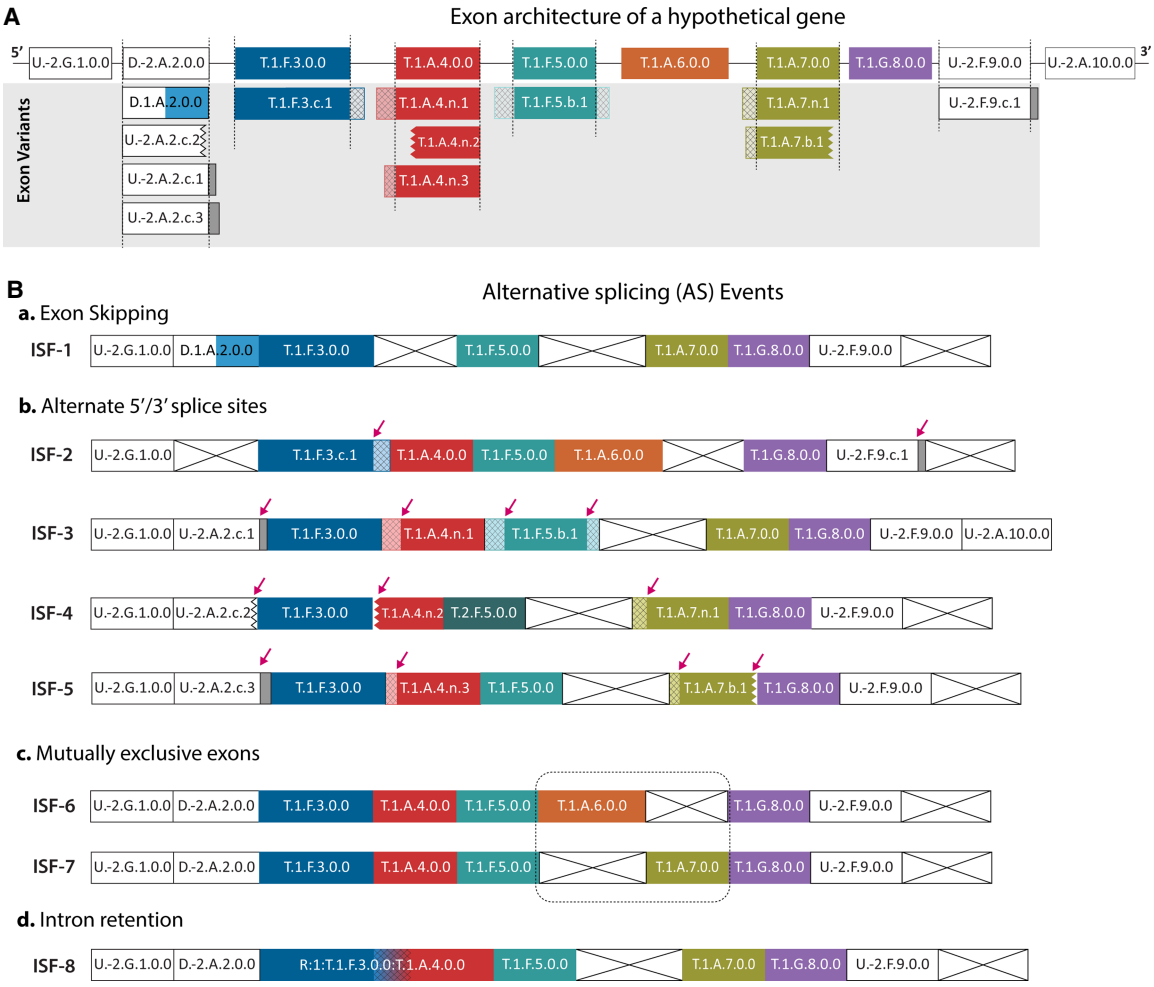


Figure 2. Schematic illustrating events involving alternative splicing (AS), alternative transcription (ATR), and alternative translation (ATL). Figure represents a hypothetical example of various AS, ATR, and ATL splicing events. (*A*) *Topmost* row represents the reference exons (*RSOEx*; see Methods), arranged in the 5′-to-3′ direction. Gray region depicts the splice-site and amino acid sequence variants related to respective *RSOEx* from equivalent positions. (*B*) Inference of AS, ATR, and ATL events is facilitated by their inter-transcript comparison. ATR and ATL events involve a change in UTR and CDS, in which ATR events are inferred from alternate first or last exons and ATL events are inferred from alternate first or last coding exons or variations in Block-I local scope within exon for different isoforms. AS events involving ES and A(ss) are depicted based on their Block-II and Block-III attributes. Red arrows highlight splice-site variations for Block-III changes. ISF-6 and ISF-7 show a MXE event involving exon 6 and exon 7 in dashed round rectangles, and ISF-8 shows an IR event from gradation of two exon colors (exon 3 and exon 4). Exons are represented as rectangular boxes with a unique identifier (EUID) assigned to them within transcripts. Variability in Block-I features of an exon is shown through variable colors. Noncoding exons are shown in white, and coding exons are filled in various colors to make a distinction from each other. Variations in splice sites are represented by crosshatched filled rectangles (for extension) and jagged ends (for shortening). The skipped exon is shown with a crossed empty rectangle box.

(EUID[3]: "F") at positions 3, 5, and 9 are depicted from EUID's Block-III features in their isoform occurrences.

Inference of ATR initiation and termination

ATR initiation and termination leave imprint by either introducing different terminal exons or varying boundaries of existing ones. In 5′ UTR, ATR initiation has the potential to modulate nascent transcript processing, translational efficiency, and their concerning processes, as noted by Churbanov et al. (2005) and Resch et al. (2009). Similarly, in 3′ UTR, ATR termination can affect the localization of the resulting protein product and polyadenylation site preferences, as discussed by Batt et al. (1994). From perspectives of gene exon architecture, alterations of termini exons or their skipping are better indicators of ATR activity than splice-site variations occurring in terminal exons, as also noted in the study by Pal et al. (2011).

ENACT can capture ATR events from alternate first or/and last exonic loci by comparing only the ordinal position of exon (denoted in Block-II) between two transcripts. This approach enhances the ability to infer ATR-associated changes, circumventing the complexities arising from alternate splice-site choices of an exon. For example, in Figure 2, all transcripts share a common 5′ start exon, constitutive (Block-II prevalence feature "G") in the UTR, indicative of a consistent transcription initiation site. However, variability in 3′ UTR exons is apparent from an inter-transcript comparison of ISF-1 and ISF-3, which shows ATR termination in the latter owing to exon-10 inclusion. We illustrate the ATR initiation event in Supplemental Figure S1 for human *GLRX2*'s transcripts NP_0057150.2 and NP_932066.1 involving exon 1 and exon 2.

Inference of ATL initiation and termination

ATL initiation and termination events can occur within a single exon or across different exons in isoforms. When ATL sites occur in different exons for initiation and termination, their event elucidation becomes similar to ATR but with an explicit focus on identifying alternate first or last "coding" exons. Conversely, when ATL sites occur within the same exon across separate isoforms, Block-I's local scope becomes informative about sequence change. As ENACT framework assigns isoform instance-specific protein-coding attributes after centralizing exons, the corresponding EUID's Block-I notes these ATL events. Specifically, these ATL events are inferred from local scope variations in equivalent EUIDs, with unchanged genomic coordinates for global scope values of "T" and "D." The following examples illustrate how ATL changes are depicted using EUID's block attributes:

- ATL initiation in *CDK7* gene (Fig. 3A). The first ordinal exon exhibits a "dual" coding scope (Block-I, global scope "D") with different translation initiation sites. The coding genomic coordinates of ISF: NP_001790.1 and ISF: NP_001311000.1 start with 92 and 216, respectively. Although genomic coordinates are unchanged, the translation start site change results in different amino acid sequences captured by the Block-I local scope (EUID: D.**1**.F.1.0.0 and D.**2**.F.1.0.0).
- Alternate first coding exons in gene *A2M* (Fig. 3B). ISF: NP_001334352.2 and NP_001334354.2 utilize different first coding exons, exon 2 and exon 5, respectively. In ISF: NP_001334352.2, exon 5 is part of continuing CDS; however, in ISF: NP_001334354.2, it becomes the first coding exon. This change is marked with different coding genomic coordinates and results in a truncated amino acid sequence compared with the reference exon (noted by Block-I local scope value 1 and 2, EUID: T.**1**.G.5.0.0 and EUID: T.**2**.G.5.0.0). Importantly, exon positions 1 and 2 are also sites of ATR initiation at which ISF: NP_001334352.2 starts from exon 1 and others from exon 2. A similar ATL instance can also be noted in Figure 2, in which exon 2 is an alternate first coding exon in ISF-1 while being a coding instance of "dual" (Block-I), and exon 3 is the first coding exon in other isoforms.
- Alternate first coding exons with different ATL sites in gene *ADSL* (Fig. 3C). We illustrate three isoforms of the *ADSL* gene in which exons at positions 1 and 2 alternate as first coding exons, with exon 2 exhibiting different ATL initiation sites. In ISF: NP_001350769.1, exon 1's translation initiation site is chosen, whereas in ISF: NP_001304852.1 and XP_047297124.1, distinct ATL sites in exon 2 are chosen. This leads to three different amino acid sequence variations from exon 2, noted by their EUID Block-I local scope values ranging from one to three. When ATL from exon 1 is chosen, exon 2 is part of continuing CDS (EUID: T.**1**.G.2.0.0). When two different ATLs in exon 2 are chosen, in their respective isoforms, it shows two distinct sequences contributions (EUIDs: T.**2**.G.2.0.0, and T.**3**.G.2.0.0).

  It can be additionally noted that exons 1 and 2 are also sites of ATR initiation, and exons 14 and 15 are sites of ATR termination.

These examples show that ENACT effectively captures ATL events through its detailed Block-I notations in EUIDs when their genomic coordinates are unchanged. However, in cases in which genomic coordinates vary, splice variants will emerge, and it will be nontrivial to infer ATL initiation/termination, as splice variants may adopt the same or different Block-I (protein-coding) status. As shown in Figure 1B, splice variants 2.c.1 and 2.n.2 transition from coding scope of reference (T.1.*.2.0.0) to UTR (U.-2.*.2.c.1) and dual scope (D.*.*2.n.2)). These transitions demonstrate the comprehensiveness of ENACT in characterizing the intricate relationship between splicing choices and translation site preferences.

Inference of AS events

AS modulates the exonic composition of transcript primarily through 4 major events: (1) exon skipping (ES), (2) alternative splice sites (A(ss)) at 5′ and/or 3′ ends of exons, (3) mutually exclusive events (MXE), and (4) IR. A(ss) and IR events are primarily inferable from EUIDs, as splice-site variations are dedicated attributes in Block-III. Conversely, MXE and ES events involve exon presence or absence inter-relationships and are inferred from pairwise transcript comparisons. Depiction of these events through the transcript's EUID comparisons is detailed below:

- Alternate splice site. ENACT assigns Block-II prevalence feature tag of "F" or "A" to exons exhibiting splice-site variations. Their specific splice variant subtype and unique occurrence are informed through Block-III's EUID[5] ("*n*"/"*c*"/"*b*") and EUID[6]. For example, in Figure 2B, exons at positions 3 and 9 with prevalence feature "F" display their 3′ splice-site variations in ISF-2 (EUID[5]: "*c*," EUID[6]:1) from their reference instances in ISF-1. Exon 5 shows splice-site variations at both the 5′ and 3′ splice sites in ISF-3 (EUID[5]: "*b*," EUID[6]:1) compared with its reference instance in ISF-2.

  Similarly, exons 2, 4, and 7 with prevalence feature "A" express different splice-site variations in isoforms. Exon 2 has three different 3′ splice-site variants in ISF-3 (EUID: U.-2.A.2.**c.1**), ISF-4 (EUID: U.-2.A.2.**c.2**), and ISF-6 (EUID: U.-2.A.2.**c.3**), all noted by

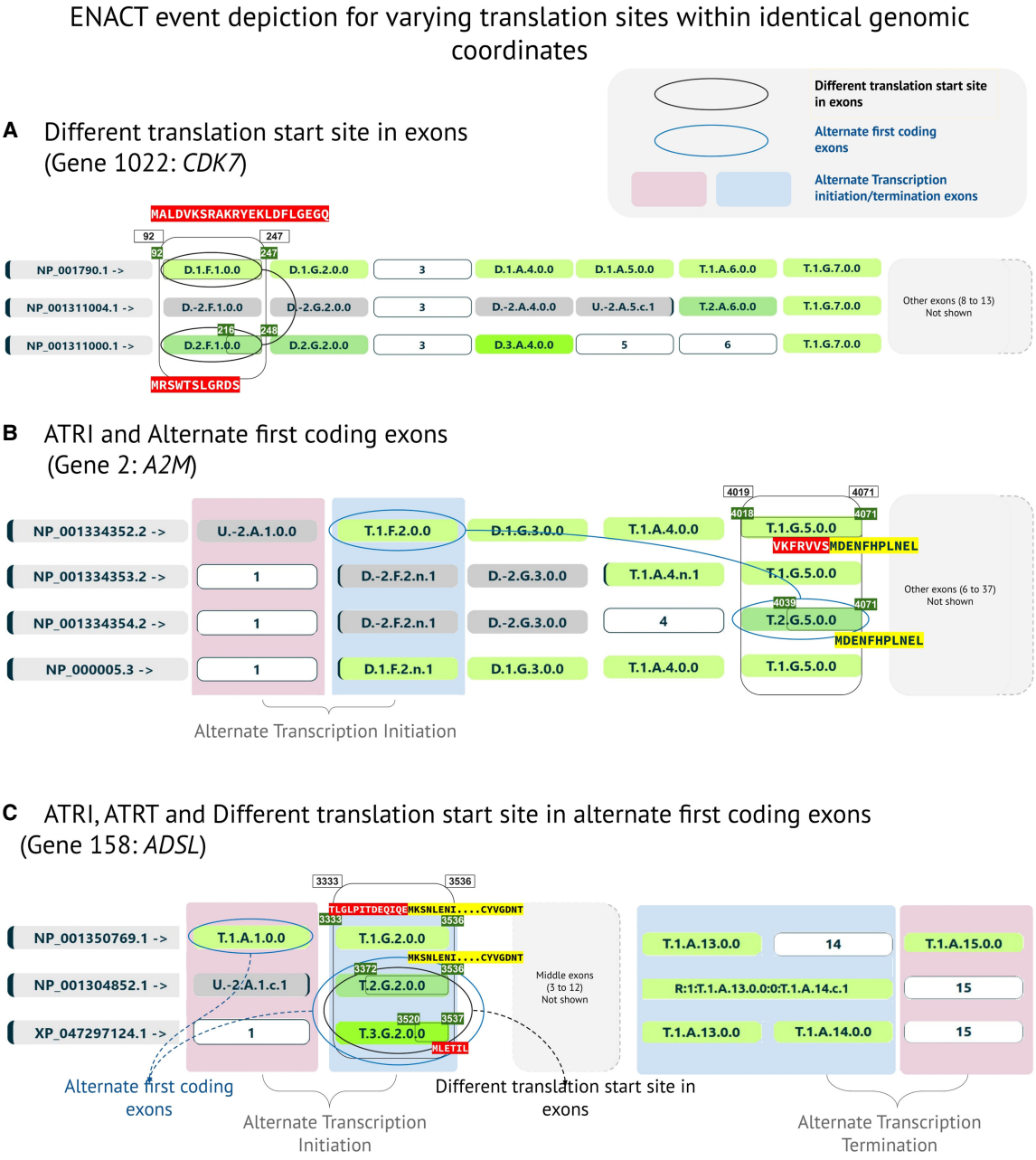


Figure 3. ENACT depiction of translation start/end site and transcription start/end site variations. (*A*) Sequence changes in the first exon of human *CDK7* gene are shown for ISF: NP_001790.1 and ISF: NP_001311000.1 owing to a change in translation initiation site as indicated by different 5′ coding genomic coordinates (black oval). This variation is captured in Block-I local scope within the EUID, as it has different values for exon 1 in two isoforms. The amino acid sequences encoded by each exon entity are also shown. (*B*) The utilization of alternate first coding exons is shown for human *A2M* gene isoforms involving exon 2 and exon 5 (blue oval). This indicates alternate translation initiation sites in isoforms through different exons. The sequence variations in exon 5 are shown and are captured in Block-I's local scope. This variation leads to considerable N-terminal sequence truncation in ISF: NP_001334354.2. (*C*) This panel illustrates variations in both alternate first exons (blue ovals) and alternate translation sites (black oval) in three isoforms of human *ADSL* gene. Exons 1 (ISF: NP_001350769.1) and 2 (ISF: NP_001304852.1 and XP_047297124.1) in the coding subspace are alternate first coding exons in respective isoforms. Exon 2 exhibits two translation initiation sites in ISF: NP_001304852.1 and XP_047297124.1, depicted using Block-I local scope values of two and three. Exon alignment obtained from ENACTdb represents alternate translation initiation sites. Genomic boundary coordinates for exons of interest across isoforms are indicated with values at the *top* of the rectangular box. Coding genomic coordinates are mentioned in the exon block with an olive-colored text background. Vertical pink and light blue backgrounds for exon positions highlight the alternate first or last exons related to alternate transcription in different isoforms.

EUID[5]: "*c*." The distinct occurrence of individual nonrepetitive splice sites is marked by an increment of count in EUID[6]. Exon 4 also exhibits variation at the 5′ splice site, as noted by EUID[5]: "*n*" in ISF-3, ISF-4, and ISF-5, with each unique occurrence tracked in EUID[6], and it can be inferred that this exon has three variants of "n" type. Lastly, exon 7 undergoes 5′ splice-site variation in ISF-4 and both 5′ and 3′ splice-site variations in ISF-5, captured by EUID[5]: "*n*" and "*b*," respectively.

- ES. Block-II prevalence feature "A" in EUIDs typically indicates an ES event for the respective exon. The "A" notation also includes splice variant(s) of alternate exon. This embedding enables consideration of two types of exon-skipping events: (1) loci with unchanged splice sites (ES(only)) and (2) loci undergoing splice-site variations (ES + A(ss)). These subtypes can be segregated using combined inference of Block-III and Block-II's "A" category. For ES(only) event, EUID[5] of corresponding exon will remain "0" across all its transcript instances, whereas for "ES + A(ss)," at least one transcript instance will have EUID[5] as "*n*" or "*c*" or "*b*."

  In Figure 2A, exon loci 6 and 10 undergo ES(only) (absent in Fig. 2A), in which exon 6 is included in ISF-2 and ISF-6, and exon 10 is included in ISF-3 (Fig. 2B). On other hand, exons 2, 4, and 7 undergo ES + A(ss) events in their respective isoforms (Fig. 2B) with defined splice variant subtype informed in Block-III.
- MXEs. MXEs involve exons that do not co-occur in the same transcripts. Although detecting and interpreting these MXE events may seem straightforward, complexities arise when MXE candidate exon positions exhibit splice-site variations (captured in Block-III). This can propose multiple choices and could introduce ambiguity in MXE selection criteria. As ENACT centralizes exon entities from genomic coordinates, splice variants are traced to the ordinal exon position. This enables easy MXE detection and prevents consideration of two exon variants at ordinal position in this categorization. Accordingly, exons 6 and 7 are considered MXEs in Figure 2C (ISF-6 and ISF-7), rather than considering every splice variant at exon 7 as an MXE with exon 6.
- IR. IR events are special cases identified with the letter "R" at the beginning of EUID. For such instances of exon fusion, EUID is expanded to depict the starting and terminating exons between which the inner region is contained as intron. For example, in Figure 2D, the descriptor "R:1:T.1.F.3.0.0:0:T.1.A.4.0.0" represents IR between exons 3 and 4 in ISF-8. The colon-separated instances depict the following:
  - The letter "R" denotes these instances as IR cases.
  - "1" indicates local protein-coding potential, identical information to Block-I (local scope).
  - The EUID of start exon in which IR begins (in this case, exon 3).
  - "0" indicates the first instance of IR from the start exon. If another retention initiates from this exon (in this case, exon 3), value would be incremented by one.
  - The EUID of the end exon up to which the intron genomic region has been retained (in this case, exon 4).

  An *AIF1* gene instance illustrating such cases is shown in Supplemental Figure S2.

Through the above-demonstrated instances of ATR, ATL, and AS, EUIDs reflect the comprehension in their cataloging after carefully integrating splice-site usage and coding region distinction with variability in ENACT.

Inference of protein indels and substitutions using ENACT

The utility of EUIDs in identifying protein sequence variations, particularly indels and substitution among isoforms from its coalescence of exon's splicing information and their isoform protein sequences, is discussed below.

Inferring protein sequence indel and substitution from Block-I attributes

EUID's Block-I attributes provide insights into coding scope of exons. Their comparison in isoforms for equivalent ordinal positions of exons can reveal changes in amino acid sequences, particularly for exons involving alternate translation sites. For instance, the "dual" exons, which participate differentially in coding and noncoding regimes, can be analyzed for differences in local scope coding variations (EUID[2]: >1) to inform substitutions (differences in amino acid sequences within the unchanged genomic coordinates of exon). Additionally, a comparison of the dual exon's coding and noncoding local scope variations (EUID[2]: "−2" and "≥1") can inform protein indels (presence or absence of amino acid sequences in the unchanged genomic coordinates of exon). Similarly, for coding exons (global scope: "T"), comparisons of their coding subtypes (EUID[2]: >1) also inform substitutions. The following illustrations highlight these cases in detail.

- Indel at N-terminal. Figure 4A illustrates the inference of protein indels in *CDK7* isoforms from dual exons. EUIDs Block-I attributes indicate the coding subtypes of dual exons in ISF: NP_001790.1 and noncoding subtypes in ISF: NP_00 1311004.1 at equivalent positions 1, 2, and 4 (local scope: "1" and "−2"). This leads to a truncated N-terminus in ISF: NP_001311004.1 compared with ISF: NP_001790.1.
- Substitution at N-terminal and indel. Figure 4B demonstrates substitutions in *CDK7* isoforms resulting from the same dual exons at positions 1, 2, and 4. In the N-terminus of ISF: NP_001311000.1, an alternate translation initiation site in exon 1 causes a change in its reading with respect to NP_001790.1, leading to sequence differences in dual exons at these positions. These are indicated in Block-I local scope values of respective EUIDs and are also evident from the protein sequence alignment of isoforms, showing unaligned instances of D.**1**.F.1.0.0 with D.**2**.F.1.0.0, D.**1**.G.2.0.0 with D.**2**.G.2.0.0, and D.**1**.A.4.0.0 with D.**3**.A.4.0.0). The altered reading frame in ISF: NP_001311000.1 is rescued later by skipping exons 5 and 6 (indel event).
- Substitution in the middle. Figure 4C highlights amino acid substitutions inferred from EUIDs in the middle of isoform's sequences through comparison of splice variants with reference isoform instances. In isoforms NP_001271455.1 and NP_001300882.1 of the *AURKB* gene, sequence variation arises from splice-site change in exon 6 (Block-III: "b"), which adapts a new protein-coding scope of Block-I ("D") from the reference Block-I ("T"). Importantly, a 5′ splice-site change in exon 7 (Block-III: "n") yields a more extended sequence that remains unaligned to exon 6 (Fig. 4C). These EUID (Block-III) variations in the middle, combined with other unchanged EUIDs, indicate the sequence modifications. In the protein sequence alignment, middle sequence substitution is observed in which the splice variant of exon 6 (EUID: D.2.F.6.**b**.1) has a truncated sequence (Fig. 4C, yellow region) compared to the reference exon 6 (EUID: T.1.F.6.**0**.0).

These examples illustrate that ENACT algorithmically identifies and denotes exonic variations in EUIDs, whose inter-isoform comparisons unveil complex splicing patterns. The listed events have the potential to provide a greater understanding of inter-transcript composition change at the level of protein sequences.

Functional diversity inference illustration by ENACT entities

In the previous section, we demonstrated the interpretation of isoform composition with their EUID entities. Here, we extend the

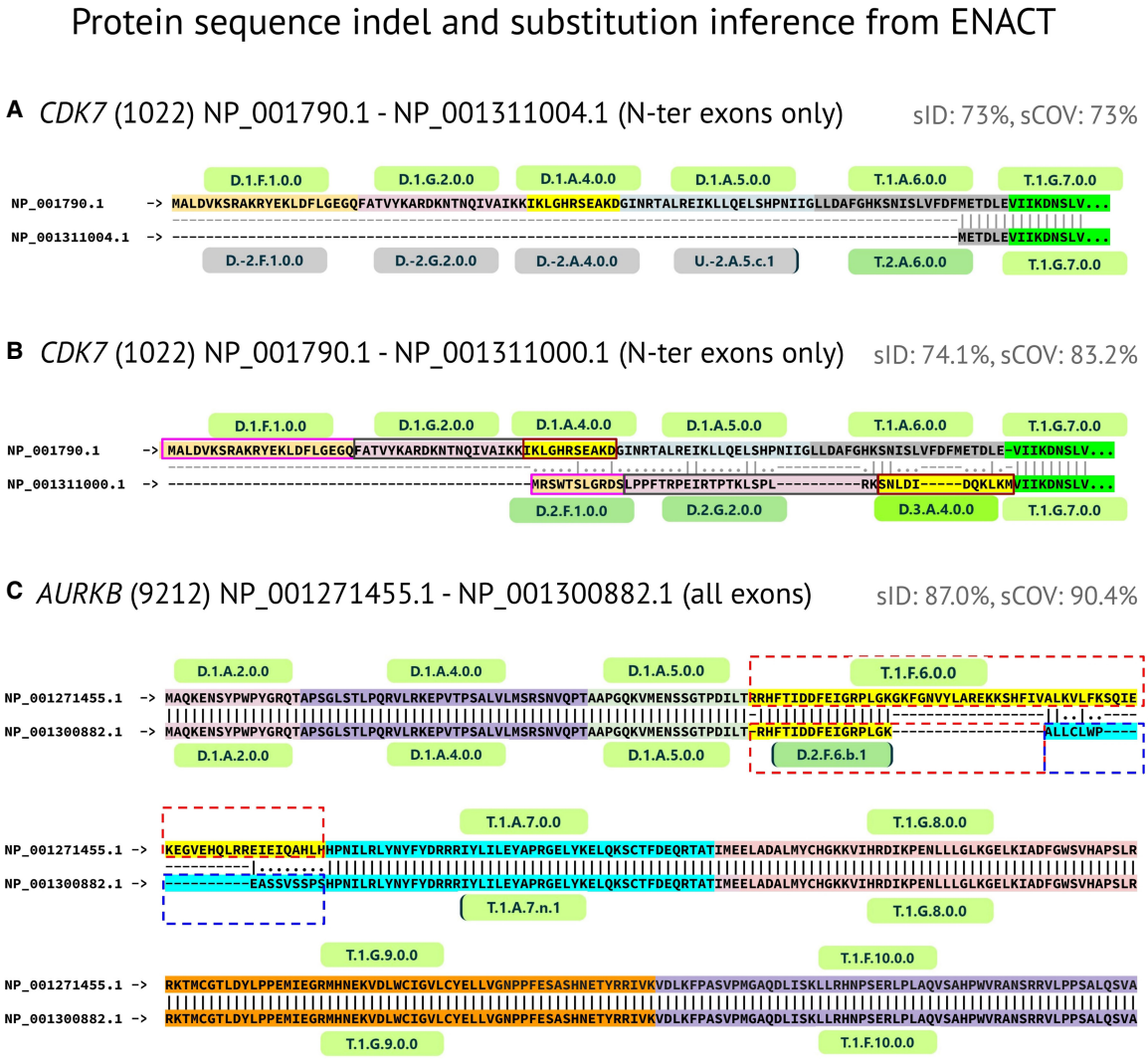


Figure 4. Features of ENACT EUIDs in inferring protein indels and substitutions. Protein sequence alignment of isoforms, overlayed with exon entity, for three genes are shown to highlight indels/substitutions. (*A*) N-terminal indel introduced by "dual" exons in *CDK7* gene, in which their noncoding instances (exons 1, 2, and 4, and a splice variant of 5) are missed in ISF: NP_001311004.1 sequence respective to ISF: NP_001790.1, resulting in isoforms having 73% of sequence identity and coverage. (*B*) N-terminal substitution (*CDK7*), in which dual exons 1, 2, and 4 contribute different amino acid sequences in ISF: NP_001311000.1 in comparison to ISF: NP_001790.1 owing to alternative translation initiation site (also depicted in Fig. 3A). The altered reading frame in ISF: NP_001311000.1 is rescued after skipping of the fifth and sixth exons. (*C*) EUID-enabled substitution inference is illustrated in the middle of the ISF: NP_001300882.1 sequence compared with ISF: NP_001271455.1 of the *AURKB* gene. Ordinal exon position 6 shows splice-site alteration (recorded by Block-III subtype "b," occurrence value: 1) in ISF: NP_001300882.1. In ISF: NP_001300882.1, this instance contributed a different sequence, which aligns with the N-terminal of the reference instance. Another event record in ISF: NP_001300882.1 is noteworthy, in which exon 7 undergoes 5′ splice-site alteration (recorded by Block-III, subtype "n," occurrence value: 1) and contributes an extended amino acid sequence. Exon 6 and exon 7 events are represented in red and blue dashed boxes. Protein sequence alignment was performed using the Needleman–Wunsch algorithm for isoform pairs. "sID" and "sCOV" refer to sequence identity and coverage, respectively. The amino acid sequences are highlighted in distinct colors to demarcate their respective exon with their corresponding EUIDs, which are listed *above* for the top isoform and *below* for the other isoform in the alignment. No EUIDs are specified for skipped exons.

significance of comprehending protein diversity in isoforms introduced through AS and related processes. We selected two genes, *ADAM8* and *WNK4*, as case studies to demonstrate the cointerpretation of functional variations using ENACT's exon-centric annotation framework.

ADAM8

The *ADAM8* gene encodes membrane-anchored disintegrin and metalloprotease family proteins. This protein belongs to the proteinases family and plays a role in cleaving the extracellular domain of several cell surface proteins and receptors (Fourie et al. 2003). It is also involved in various cellular functions, including inflammation, immunomodulation, neutrophil activation/mobility, immune cell migration, osteoclast stimulating factor, and neurodegeneration (Yamamoto et al. 1999; Schlomann et al. 2000; Romagnoli et al. 2014). The *ADAM8* domain architecture consists of an N-terminal prodomain, a catalytic metalloproteinase domain, a disintegrin domain for interaction with integrins, a cysteine-rich domain, a transmembrane region, and a C-terminal

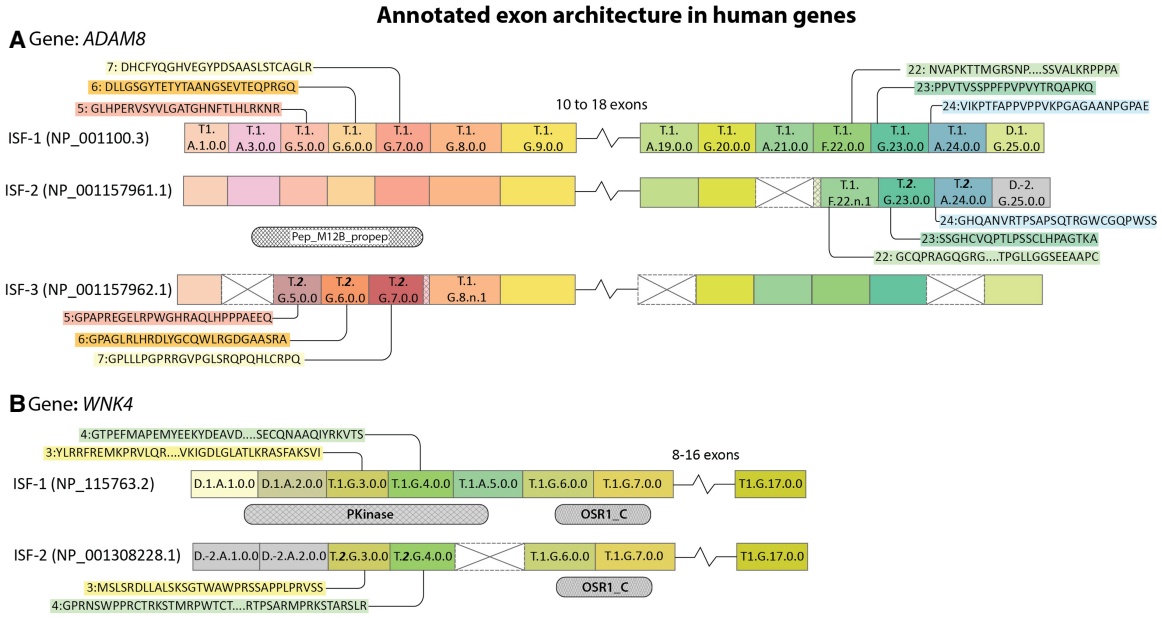


Figure 5. ENACT annotation of *ADAM8* and *WNK4* isoforms and exons. (*A*) In *ADAM8* isoforms, ISF-2 undergoes a reading frame shift in the C-terminal region, involving Block-I local scope variation in exons 23 and 24. A similar reading frame shift occurs in ISF-2 in the N-terminal region, resulting in a loss of the "Pep_M12B_propep" domain. This change in reading frame is noted through Block-I local scope variation in exons 5, 6, and 7. (*B*) In *WNK4*, ISF-2 lacks a "Pkinase" domain in the N-terminal region and undergoes reading frame shift in exons 3 and 4, as indicated by their Block-I local scope values. The NCBI gene identifier is shown for each isoform, and exons are represented as colored rectangle boxes with their respective EUID. The skipped exon is shown with a crossed empty rectangle box. The small extension of exon rectangle boxes with crisscross-filled lines represents extended exon boundaries owing to alternate splice sites (5′/3′). Exons not showing any variation are *left* unlabeled. The break shows that exons intervening in the region do not change in the isoforms. The jagged edge of a rectangle represents alternate 5′ or 3′ splice site. The isoforms sharing Pfam domains for a region are shown *below* the exon layout. Changes in Block-I local scope are complemented with sequence specifications.

domain involved in protein-protein interaction (Knolle and Owen 2009).

ADAM8's gene architecture comprises 23 coding, one noncoding, and one "dual" exon. Among coding exons, 17 are constitutive or constitutive-like, whereas others are alternate. We analyzed three isoforms harboring a combination of AS events (Fig. 5A). The reference isoform (ISF-1; NP_0011003.3) has Pep_M12B_propep, reprolysin (metalloproteinase), disintegrin, and ADAM_CR (cysteine-rich) Pfam domains lying before transmembrane region. In ISF-2, skipping of exon 21 is combined with 5′ splice site of exon 22 (Block-III: "*n*"). This leads to exonic substitution and frameshift, impacting subsequent exons 23 and 24 (local scope: "2") with premature termination. Compared with ISF-1, ISF-2 has 79% global sequence identity and lacks proline-rich regions, which is required for protein–protein interaction. It will be worth exploring the functional impact of ISF-2, as it is expressed in metastatic lung cancer cell lines (Knolle and Owen 2009).

Skipping of exons 2 to 4 in ISF-3 induces amino acid substitution and frameshift in exons 5 to 7 (local scope: "2"). Notably, the reading frame is restored in exon 8 by 5′ splice-site event (Block-III: "*n*"). Because of this, ISF-3 lacks a prodomain in the N-terminal region, suggesting constitutive metalloproteinase activity. As a consequence, ISF-3 loses two of four glycosylation sites from the prodomain (Srinivasan et al. 2014) but preserves the conserved glutamate (158E), which is essential for prodomain's catalytic removal (Hall et al. 2009). Further experimental studies would provide information on the enzymatic activity and biological role of ISF-3.

WNK4 gene

The *WNK4* gene belongs to the "with no lysine (WNK)" group of serine/threonine kinases (STK) in eukaryotic organisms. These have been named because of their atypical positioning of catalytic lysine in subdomain II instead of subdomain I, as in other STKs. *WNK4* is expressed primarily in the kidney, where *WNK4* and other members of the family have a role in modulating the balance between sodium chloride reabsorption and renal potassium ion secretion (Murillo-de-Ozores et al. 2021) by regulating the activities of cation-coupled cotransporters (SLC12, NCC), ion channels (ENaC) and ion exchangers (Moriguchi et al. 2005; San-Cristobal et al. 2008). Mutations in the *WNK4* are associated with a rare genetic hypertension disorder called pseudohypoaldosteronism type 2 (PHA2).

WNK4 contains Protein kinase and "Oxidative-stress-responsive Kinase1" C-terminal (OSR1_C) domains, in which OSR1_C encompasses the Pask-Fray 2 (PF2) domain. The PF2 region interacts with RFX[VI] motif and suppresses the activity of kinase domain (Murillo-de-Ozores et al. 2021). Other than these domains, the rest of the *WNK4* protein sequence is largely intrinsically disordered.

A total of 13 isoforms are listed in the NCBI RefSeq database. Of these, two isoforms are reviewed (Fig. 5B) and considered for further analysis. In ISF-1: NP_115763.2, the first two exons harbor coding subtypes of "dual" exons (global scope: "D") and contribute to the assignment of kinase domain (Fig. 5B) to the N-terminal region. In contrast, ISF-2 has a noncoding contribution from these "dual" exons, leading to an alternate translation initiation site in exon 3 (local scope: "2"). These events lead to amino acid deletion at exons 1 and 2 and substitution at exons 3 and 4 because of a shift

in the reading frame. The reading frame in ISF-2 is subsequently restored by exon 5 skipping (indel event). Because of this, ISF-2 lacks kinase domain's integrity but retains the OSR1_C (PF2) domain, which could act as a potential sequestering factor by interacting with the SPAK/OSR1 protein.

Thus, the exonic variations were localized to their specific protein sequences and functional domains through an integrated ENACT annotation framework. The resulting EUIDs represent isoform compositions and are analyzed alongside predicted protein features of Pfam domains, disordered regions, and secondary structures. These are integrated for user-friendly visualization in ENACTdb (Verma et al. 2024). Although we have focused on selected isoforms, the comparison above could be performed for any isoforms. ENACT preserves unique exon combinations and variations specific to each isoform and reflects them through individual EUIDs. We also discussed an example of the *GLRX2* gene (Supplemental Fig. S1) that shows tissue-specific expression patterns involving different exons with ENACT depiction.

Our approach provides detailed insights into how exon variation affects isoform functionality, highlighting the value of integrated annotation framework in functional genomics and understanding protein diversity.

Discussion

In the present study, we standardized exon relative position across all protein-coding isoforms, focusing on their variations, such as indels and splice-site polymorphs, and associating them with the coding potential of exonic loci. By systematically annotating exonic loci through EUIDs, ENACT allows inference of their multifaceted roles within gene isoforms and a detailed understanding of how exonic diversity impacts protein variations and functionality. Previous approaches, such as ASTALAVISTA (Foissac and Sammeth 2007; Sammeth et al. 2008), perform comprehensive and sensitive event categorization; however, it only considers them in a pairwise manner. ENACT executes an exon-centric process distinct from such methods and is uniquely capable of capturing and tracking protein-coding information for exons across all its occurring isoforms. Leveraging this strength, ENACT unveils the polymorphic nature of exon while also inferring alternate transcription, translation, and splicing events. These examples demonstrate ENACT's ability to detail respective events and combinations in transcripts through information extraction from EUIDs. The embedded coding potential will also provide insights into exon conservation and the evolution of spliceosome-based exonic recognition. Collectively, this could advance our understanding of how "alternate" to "constitutive" or vice versa transitions in exons (Koren et al. 2007; Lev-Maor et al. 2007) shape the functionally relevant protein structural elements. Importantly, ENACT's systematic accrual of exon attributes from protein-coding isoforms bridges the gap between isoform-centric and gene-level analysis, which is necessary for investigating the expansion, fusion, or fission of equivalent exons in orthologous genes. Moreover, given the observation of varied splicing patterns for comparable gene expression (Jacobs and Elmer 2021; Singh and Ahi 2022; Verta and Jacobs 2022), ENACT will provide a platform to investigate their differences from the analysis of common denominator protein products.

By categorizing exons in CDS, UTR, or transitional regions ("dual"), ENACT allows footprint studies of alternative promoters, terminators, and translation sites. Their combined influence on modulating inter-isoform protein diversity has been illustrated through case studies on *ADAM8* and *WNK4*. This demonstrated ENACT's ability to analyze domain-associated splicing events integrated within ENACTdb (Verma et al. 2024). Complimenting above, gene-wide regional segregation of exons and their coinvestigation with splice events has the potential to enable how they shape exon architecture that allows regulations (Shabalina et al. 2010, 2014) from transcriptional and translational components. For instance, *ADAM8* and *WNK4* showed how specific AS events rescue reading-frame changes induced by ES, preserving domain functionality. Extending these investigations could provide essential insights into the balance between protein structural and functional constraints in evolution of AS.

ENACT provides a platform to explore broader correlations between exon/intron rearrangements in CDS, 5′ UTR, and 3′ UTR regions, building on the previous work of Shabalina et al. (2010). Briefly, Shabalina et al. (2010) noted a positive correlation between (1) alternate transcription initiation and AS in 5′ UTR and (2) AS in the CDS region and alternate transcription termination in 3′ UTR, but they noted an anticorrelation between (3) AS in CDS and AS with alternate transcription in 5′ UTR. Although the biological relevance of the first two observations aligns with existing literature, the anticorrelation (observation 3) remains challenging to explain and rationalize. These anticorrelations suggest modulation of upstream UTR without associated splicing pattern changes. A careful analysis of *WNK4* indicates another mechanism related to observation 3. Here, AS in CDS (exon-5 skipping) complements ATL (exon 3), with no AS and ATR in 5′ UTR. Because AS and ATR in 5′ UTR also have potential to influence translation and could introduce ATL (Cenik et al. 2010; Palaniswamy et al. 2010; Kramer et al. 2013; Weber et al. 2023), in *WNK4*, the ATL is driven by exon 3 compensated for this role. Moreover, the ATL followed by ES of exon 5 restores the reading frame change introduced by the former (exon), resulting in partial N-terminal variability, demonstrating the impact of their combined action leading to partial diversification and shedding light on the rudimentarily understood aspect of their anticorrelation. This is speculative and requires a comprehensive study for further validation, and it could pave the way for a detailed examination to rationalize the mechanistic basis of anticorrelation (Shabalina et al. 2010).

Conclusively, by integrating isoform data and protein sequence features with exon-centric annotations, ENACT advances our understanding of gene architecture and its functional implications. It is an invaluable resource for experimentalists and computational biologists aiming to decipher the functional repertoire embedded in transcriptional and translational processes.

Limitations of the ENACT framework

The current framework focuses on centralizing coding/noncoding exons derived only from protein-coding genes. These centralized exonic positions were annotated with their genomic coordinates and protein sequence variations. In the present framework, exons specific to untranslated transcripts of protein-coding genes and exons from non-protein-coding genes are not annotated.

ENACT defines exons as polymorphic loci and integrates these with the encoded protein sequence, facilitating exon tracking in gene architecture. This process utilizes well-annotated gene models consisting of splice architecture and transcript translation scope data. Hence, it limits the annotation of transcripts available in NCBI or similar databases. Metadata associated with exonic/genetic architecture, including but not limited to epigenetic marks and regulatory elements, can be incorporated into the annotation. However, ENACT currently does not integrate such

information; instead, it focuses on assessing the impact of exons in protein sequence variations within genes.

Methods

Overview of ENACT framework

Unique indexing of exon to construct gene architecture and define alternate/constitutive features to exon

RISO selection

For a given gene, we select an isoform having the maximum number of coding exons from a curated set of isoforms (NCBI RefSeq proteins having "NP_" prefix) and define it as a Reference ISOform (RISO). If the number of coding exons is identical in two or more isoforms, then the one with the longest amino acid ("aa") length is selected as RISO. If a gene has no "NP_" prefixed isoforms, RISO can be chosen from all known isoforms using similar criteria (Fig. 6A; Box 1).

Defining reference set of exons

Initially, exons of RISO constitute the *RSOEx*, which is subsequently populated with nonredundant exons from other isoforms (*NRExon*) based on overlap of genomic coordinates (GC) (Box 1; for representation, see Fig. 6B1). This procedure of exon selection (for *RSOEx*)

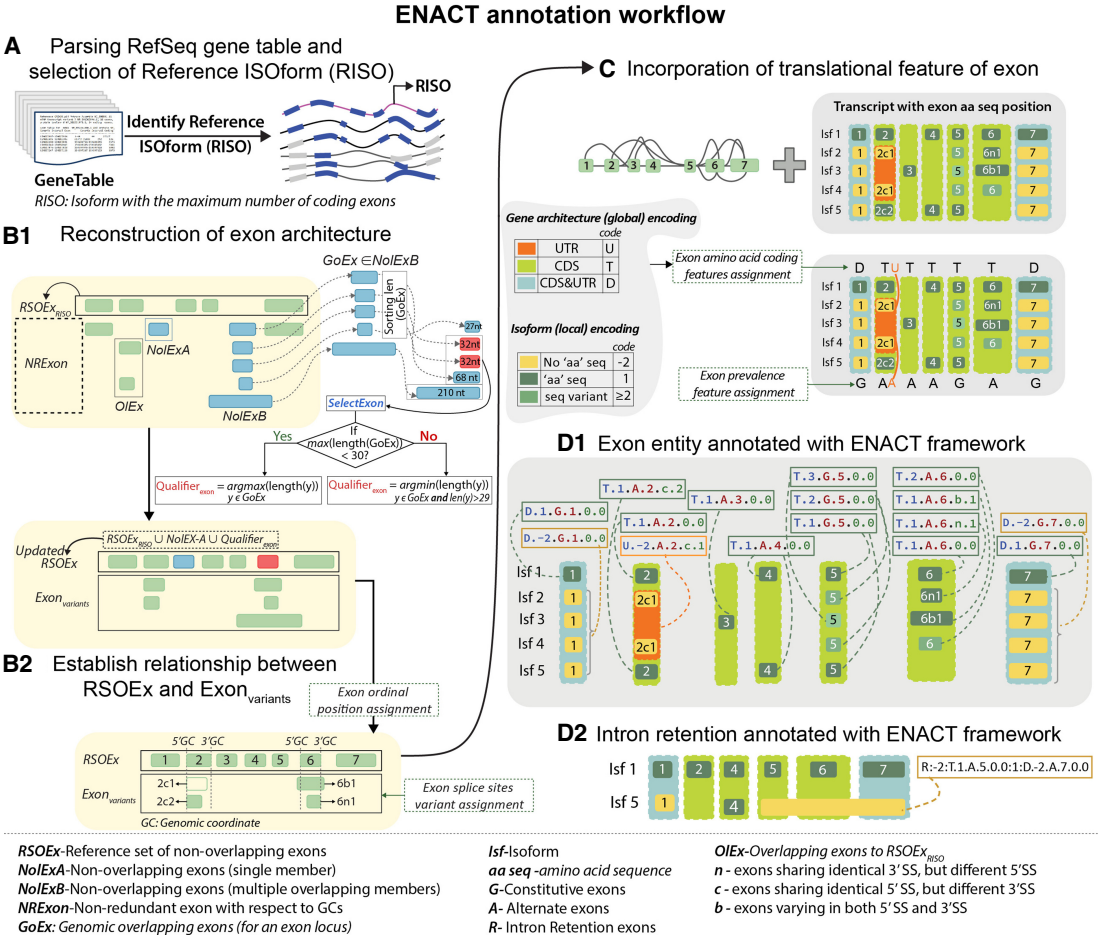


Figure 6. Overview of ENACT annotation workflow. Figure illustrates the main steps of the ENACT algorithm (*A–D*). (*A*) Curated models of protein-coding transcripts or isoforms with their genomic and coding genomic coordinates are collected in gene table format from NCBI RefSeq (blue and gray colors represent coding and noncoding exons, respectively). A reference isoform (RISO) is selected based on the maximum number of coding exons (see Methods). (*B1*) The initial set of reference exons ($RSOEx_{RISO}$) consists of exons from RISO, which are compared with nonredundant exons (*NRExon*) from other isoforms using genomic coordinates. This yields a set of overlapping (*OIEx*) and nonoverlapping exons (*NoIEx*). *NoIEx* entities without varying genomic coordinates are categorized as *NoIExA* and are added to *RSOEx*, whereas other *NoIEx* entities (*NoIExB*) are further processed to yield subsets that show genomic coordinate overlaps for a locus, referred to as *GoEx*. One such *GoEx* entity is illustrated in *B1* (penultimate exon). *GoEx* sets are iterated to select representative exons ($Qualifier_{exon}$) for the *RSOEx*, based on exon length using "selectExon" procedure (see Box 1). The updated RSOEx includes *NoIExA* and $Qualifier_{exon}$ from *GoEx* subgroups. Similarly, the $Exon_{variants}$ set includes initial *OIEx* and exons from *GoEx* subgroups, excluding their $Qualifier_{exons}$. (*B2*) $Exon_{variants}$ entities are compared with *RSOEx* to define alternate splice-site variants: 5′, 3′, and 5′ with 3′ are characterized as "n," "c," and "b," respectively. (*C*) The *RSOEx* and their associated splice variants from the $Exon_{variants}$ set are sorted based on genomic coordinates and assigned ordinal positions (from one to the number of exonic loci in *RSOEx*). After establishing updated *RSOEx* and others as splice-site variants, exon translation attributes are assigned considering coding genomic coordinates. Further, prevalence attributes are noted based on locus prevalence across isoforms: (G) constitutive, (A) alternate. This procedure annotates each exon in *RSOEx* and $Exon_{variants}$ with its relative position, translational (protein-coding potential) feature, occurrence, and splice-site variations. (*D1*) Exon attributes are combined to construct a six-character alphanumeric notation defined as an exon unique identifier (EUID). For the representative example, EUIDs for each exon are shown. (*D2*) The intron retention instances are identified in step B and annotated with intron retention (IR) codes.

involves the segregation of overlapping exons (*OlEx*; with GC overlap to *RSOEx*) from nonoverlapping exons (*NolEx*; without GC overlap to *RSOEx*). Overlapping exons (*OlEx*) to $RSOEx_{RISO}$ are first extracted and added to the $Exon_{variants}$ set. The *NolEx* group consists of singleton exons (*NolexA*), which are directly added to *RSOEx*, and another subgroup (*NolExB*) having exons with self GC overlaps (for description, see the routine *defineSuboverlapExons* in Box 1; shown in Fig. 6B1).

NolExB subgroups are further processed to identify exons with self-overlapping genomic coordinates, referred to as *GoEx*. From this *GoEx* subset, $Qualifier_{exon}$ is selected based on following criteria: an exon of minimum length of at least 10 "aa" (30 nt) or the next longer length is preferred, provided the longest member of *GoEx* has a length >29 nt. If the longest *GoEx* entity does not exceed 30 nt in length, then the longest exon in the *GoEx* subset is chosen as representative ($Qualifier_{exon}$) for *RSOEx* set (Fig. 6B1). After $Qualifier_{exon}$ selection, other *GoEx* members are moved to $Exon_{variants}$ set.

Subsequently, *RSOEx* exons are sorted based on their genomic coordinates and are assigned ordinal positions (see Fig. 6B2). Further details of ENACT procedure and iteration-specific depiction of representative exon choice from *GoEx* are provided for the human gene *RUNX1T1* in Supplemental Figure S3.

Length choice of ≥30 nt

The current criterion prioritizes an empirical length threshold of ~30 nt (10 amino acids) for inclusion in *RSOEx*. This criterion will not discard any exon and is used only to choose the representative entity from the *GoEx* set. We have carefully chosen this, as considering representative length of exon shorter than 10 "aa" could result in members of *GoEx* (processed later from $Exon_{variants}$) undergoing IR, particularly if intron length between exons is smaller and comparable to the length of $Exon_{Variant}$ instances. This will be common in lower organisms that favor the intron definition model (Talerico and Berget 1994; Rogozin et al. 2012). Conversely, choosing a larger value than 10 "aa" could cause the smaller *GoEx* members to be represented as split exons, which will be particularly relevant for higher organisms with exon definition preference (Zheng 2004), in which alternative exon usage is the dominant splicing event, and splice-site relationships are more complex.

Importantly, as length threshold affects only the alternative exons, choosing uniform criteria to select representative *RSOEx* entities strengthens ENACT's ability to analyze orthologous exonic positions, including the regions undergoing constitutive to alternate transitions (Lev-Maor et al. 2007).

Prevalence of exon entities

Upon constructing *RSOEx* and their linear indexing, ENACT next depicts alternate/constitutive properties of exons, defined by their GC consistency in transcripts. An exon is considered alternate (shown as "A") if it lacks uniform presence in all transcripts and is constitutive (depicted using "G") otherwise. Additionally, certain exonic positions are depicted by "F," representing these as occurring in all isoforms, but with splice-site variations.

Relationship definition between RSOEx and $Exon_{variants}$

The previous step identifies reference exon (*RSOEx*) and variant exon sets ($Exon_{variants}$). Here, we define specific splice-site relations (5′ and/or 3′ exon) between their exons, considering genomic coordinate overlap. We define splice sites and their variations by considering the exonic definition model (De Conti et al. 2013), as ENACT focuses on featurization of them and their protein encoding potential.

Splice-site variability

Each exonic entity in $Exon_{variants}$ (referred using index "*i*") is compared with those in *RSOEx* (referred using index "*k*") and is assigned notation as follows:

- (n) It denotes a different 5′ splice site (5ss) but an identical 3′ splice site (3ss) for *i*th entity of $Exon_{variant}$ to the *k*th entity of *RSOEx*.
- (c) It denotes identical 5ss but different 3ss for *i*th entity of $Exon_{variant}$ to *k*th entity of *RSOEx*.
- (b) It denotes different 5ss and 3ss for *i*th entity of $Exon_{variant}$ to *k*th entity of *RSOEx*.
- (0; default) It denotes identical 5ss and 3ss for *i*th entity of $Exon_{variant}$ to *k*th entity of *RSOEx*.

The above notations (*n*/*c*/*b*/0) will depict splice site variability and represent either the extension or shortening of the exon length (see Fig. 1). It should be noted that when an exonic entity ($Exon_{variants}$) showed GC overlap to more than one entity in *RSOEx*, we defined them as IR events, which is described in results and discussed later (see Methods, section "Intron retention").

Occurrence of splice-site changes

GC overlaps facilitate splice-site relationship inference between entities of $Exon_{variants}$ and *RSOEx*. To accommodate and acknowledge more than one GC overlapped entity ($Exon_{variants}$) being related to the *RSOEx* instance, we count and track their occurrences. For instance, in Figure 6B2, two $Exon_{variants}$ entities have identical 5ss and altered 3ss compared with the *RSOEx* instance in the top row, inferring two distinct "*c*" variations. Their distinct occurrence (1 and 2) is notated to inform different splice sites, specific for subtypes "*n*," "*c*," and "*b*." Additionally, this value for the *RSOEx* entity will be represented by zero. Importantly, the number of *n*/*c*/*b* events at a specific exon locus (ordinal position "N") can be obtained by extracting this attribute.

Amino acid coding (translational) attribute of exons defined in RSOEx and $Exon_{variants}$

The translational attribute of the exon is depicted at two levels (see Fig. 1, Block-I). The first level describes global scope, denoting the region of gene architecture to which the exonic locus belongs, and the second level describes local scope and denotes its isoform-specific amino acid contribution (Fig. 6C). Global scopes are denoted by letters "T," "U," "D," "M," and "R," and local scopes have been defined by numeric identifiers "−2" to "n," with details as follows:

- "**T**" (translated). This depicts the exonic locus containing coding genomic coordinates (CGCs) in all its isoform instances. Hence, it is a part of CDS regime in gene architecture. An amino acid sequence of the current locus need not be uniform among its occurring isoforms and may yield different sequences, considering alternate promoter-driven alternate translation initiation in an N-terminal, non-3n ES-driven reading frame change in the middle, or truncation in a C-terminus (alternate translation termination). Accordingly, subtype local scope is necessary to encompass complete translational features and is defined as below:
  - 1: Reference "aa" contribution of an exon.
  - ≥2: Increment counter denoting the number of different "aa" variants observed for an exon with the same GCs (because of alternate translation or reading frameshift).
  - −1: Depicts premature stop codon in the upstream exon; hence, concerning locus lacks "aa" contribution despite having CGC.

- "**U**" (untranslated). This depicts the exonic locus lacking CGC for all its isoform instances and is a part of UTR regime throughout gene architecture. The numeric tag "−2" denotes isoform-specific scopes (local) of such instances.
- "**D**" (dual). These are exonic loci that are part of the CDS and UTR regime, thus showing inconsistent CGC (variable and none) in its occurring isoforms. Additionally, for a locus to have a global scope "D," it should have at least two local scope instances with tags of "−2" and "1."
- "**M**" (micro). This depicts locus having CGC of "1 nt" and indicates single-nucleotide protein-coding exon in the CGC region. The "0" local tag marks these instances.
- "**R**" (retention). This denotes exon undergoing IR, described in the later section (see Methods, section "Intron retention").

A further procedure extension of ENACT that maps the protein translation block to *RSOEx* and $Exon_{variants}$ entities is shown in Supplemental Figure S3G.

EUID construction

The attributes discussed in the method sections "Amino acid coding (translational) attribute of exons defined in *RSOEx* and $Exon_{variants}$" (local and global scopes), "Unique indexing of exon to construct gene architecture and define alternate/constitutive features to exon" (subsections "RISO selection" and "Defining reference set of exons"; prevalence and linear position), and "Relationship definition between *RSOEx* and $Exon_{variants}$" (subsections "Splice-site variability" and "Occurrence of splice-site changes"; splice-site subtype and unique occurrence) comprise Block-I, II, and III attribute groups, respectively. The exonic features are encoded as alphanumeric characters joined sequentially by "." (in the order of Block-I, Block-II, and Block-III) as demonstrated in Figures 1 and 6D1, yielding a six-character long alphanumeric string descriptor termed an exon unique identifier (EUID). Briefly, the first block denotes translational attribute, and the second block indexes the relative linear position of the exon in a gene, followed by a feature of exon occurrence. The third block depicts the exon splice-site variations.

The subset of EUID characters can be used to construct subfeatures by invoking $EUID^{k}$, where "k" ranges from one to six and represents one or more attributes or their combinations.

Intron retention

The IR events previously filtered from $Exon_{variants}$ (see Methods, subsection "Splice-site variability") (Fig. 6B1) to have GC overlap with more than one *RSOEx* entity are treated as a special case of exon nomenclature in which a six-character EUID is insufficient to capture details of their retention event (as it involves retention of exon/intron region between exons). We combine EUIDs of two exons with three other identifiers separated by a colon (":") to construct IR-EUID (Fig. 6D2). The first identifier is the alphabet "R" to recognize that the exon is involved in IR, followed by a digit describing its amino acid coding attribute (the same notation is used as described before in the Methods, section "Amino acid coding (translational) attribute of exons defined in *RSOEx* and $Exon_{variants}$"). The third and fifth identifiers are exon EUIDs, between which the intron/exon region is retained to form the IR exon. The fourth identifier is a numeric character showing the number of retention events observed involving exons and their variants. We use "0" as the default value of this counter. For instance, in Figure 6D2, the IR identifier "R:-2:T.1.A.5.0.0:0:D.-2.A.7.0.0" depicts the noncoding exon (second digit delimited by ":" in IR identifier) of the IR exon encompassing from exon at position 5 (T.1.A.**5**.0.0) to exon at position 7 (D.-2.A.**7**.0.0). It is the first instance starting from exon 5 (fourth digit delimited by ":" in IR identifier). Other instances of IR are illustrated in ISF-8 of Figure 2 and Supplemental Figure S2.

Data availability

Using ENACT, we have annotated protein-coding genes of five representative organisms viz. *Caenorhabditis elegans* (worm), *Drosophila melanogaster* (fruit fly), *Danio rerio* (zebrafish), *Mus musculus* (mouse), and *Homo sapiens* (human) (see Supplemental Text S4). These annotations are available on the webserver at https://www.iscbglab.in/enactdb (Verma et al. 2024). ENACT annotations for additional representative organisms will be included and maintained through the ENACTdb resource.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by the Indian Institute of Science Education and Research Mohali; Bioinformatics Center, Department of Biotechnology under the Ministry of Science and Technology, Government of India (BT/PR40419/BTIS/137/36/2022) and the National Network Project, Department of Biotechnology under the Ministry of Science and Technology, Government of India (BT/PR40198/BTIS/137/56/2023). We acknowledge the computing facility Param Smriti formed under a National Supercomputing Mission.

Author contributions: P.V. and D.T. were involved in methodology development, data analysis, and visualization. D.A. helped investigate and analyze the review. S.B.P. conceptualized, designed, and supervised the study. All authors were involved in result interpretations and writing of manuscript.

References

Aspden JL, Wallace EWJ, Whiffin N. 2023. Not all exons are protein coding: addressing a common misconception. *Cell Genom* **3:** 100296. doi:10.1016/j.xgen.2023.100296

Baralle FE, Giudice J. 2017. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol* **18:** 437–451. doi:10.1038/nrm.2017.27

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Çolak R, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338:** 1587–1593. doi:10.1126/science.1230612

Batt DB, Luo Y, Carmichael GG. 1994. Polyadenylation and transcription termination in gene constructs containing multiple tandem polyadenylation signals. *Nucleic Acids Res* **22:** 2811–2816. doi:10.1093/nar/22.14.2811

Blencowe BJ. 2017. The relationship between alternative splicing and proteomic complexity. *Trends Biochem Sci* **42:** 407–408. doi:10.1016/j.tibs.2017.04.001

Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM. 2012. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell* **46:** 871–883. doi:10.1016/j.molcel.2012.05.039

Cenik C, Derti A, Mellor JC, Berriz GF, Roth FP. 2010. Genome-wide functional analysis of human 5′ untranslated region introns. *Genome Biol* **11:** R29. doi:10.1186/gb-2010-11-3-r29

Churbanov A, Rogozin IB, Babenko VN, Ali H, Koonin EV. 2005. Evolutionary conservation suggests a regulatory function of AUG triplets in 5′-UTRs of eukaryotic genes. *Nucleic Acids Res* **33:** 5512–5520. doi:10.1093/nar/gki847

De Conti L, Baralle M, Buratti E. 2013. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* **4:** 49–60. doi:10.1002/wrna.1140

Deveson IW, Brunck ME, Blackburn J, Tseng E, Hon T, Clark TA, Clark MB, Crawford J, Dinger ME, Nielsen LK, et al. 2018. Universal alternative

splicing of non-coding exons. *Cell Syst* **6:** 245–255.e5. doi:10.1016/j.cels.2017.12.005
Foissac S, Sammeth M. 2007. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res* **35:** W297–W299. doi:10.1093/nar/gkm311
Fourie AM, Coles F, Moreno V, Karlsson L. 2003. Catalytic activity of ADAM8, ADAM15, and MDC-L (ADAM28) on synthetic peptide substrates and in ectodomain cleavage of CD23. *J Biol Chem* **278:** 30469–30477. doi:10.1074/jbc.M213157200
Hall T, Leone JW, Wiese JF, Griggs DW, Pegg LE, Pauley AM, Tomasselli AG, Zack MD. 2009. Autoactivation of human ADAM8: A novel pre-processing step is required for catalytic activity. *Biosci Rep* **29:** 217–228. doi:10.1042/BSR20080145
Jacobs A, Elmer KR. 2021. Alternative splicing and gene expression play contrasting roles in the parallel phenotypic evolution of a salmonid fish. *Mol Ecol* **30:** 4955–4969. doi:10.1111/mec.15817
Johnstone TG, Bazzini AA, Giraldez AJ. 2016. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J* **35:** 706–723. doi:10.15252/embj.201592759
Kamieniarz-Gdula K, Proudfoot NJ. 2019. Transcriptional control by premature termination: a forgotten mechanism. *Trends Genet* **35:** 553–564. doi:10.1016/j.tig.2019.05.005
Knolle MD, Owen CA. 2009. ADAM8: a new therapeutic target for asthma. *Expert Opin Ther Targets* **13:** 523–540. doi:10.1517/14728220902889788
Kochetov AV. 2008. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays* **30:** 683–691. doi:10.1002/bies.20771
Koralewski TE, Krutovsky KV. 2011. Evolution of exon-intron structure and alternative splicing. *PLoS One* **6:** e18055. doi:10.1371/journal.pone.0018055
Koren E, Lev-Maor G, Ast G. 2007. The emergence of alternative 3′ and 5′ splice site exons from constitutive exons. *PLoS Comput Biol* **3:** e95. doi:10.1371/journal.pcbi.0030095
Kramer M, Sponholz C, Slaba M, Wissuwa B, Claus RA, Menzel U, Huse K, Platzer M, Bauer M. 2013. Alternative 5′ untranslated regions are involved in expression regulation of human heme oxygenase-1. *PLoS One* **8:** e77224. doi:10.1371/journal.pone.0077224
Lee S, Liu B, Lee S, Huang SX, Shen B, Qian SB. 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci* **109:** E2424–E2432. doi:10.1073/pnas.1207846109
Lev-Maor G, Goren A, Sela N, Kim E, Keren H, Doron-Faigenboim A, Leibman-Barak S, Pupko T, Ast G. 2007. The "alternative" choice of constitutive exons throughout evolution. *PLoS Genet* **3:** e203. doi:10.1371/journal.pgen.0030203
Manuel JM, Guilloy N, Khatir I, Roucou X, Laurent B. 2023. Re-evaluating the impact of alternative RNA splicing on proteomic diversity. *Front Genet* **14:** 1089053. doi:10.3389/fgene.2023.1089053
Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338:** 1593–1599. doi:10.1126/science.1228186
Moriguchi T, Urushiyama S, Hisamoto N, Iemura S-I, Uchida S, Natsume T, Matsumoto K, Shibuya H. 2005. WNK1 regulates phosphorylation of cation-chloride-coupled cotransporters via the STE20-related kinases, SPAK and OSR1. *J Biol Chem* **280:** 42685–42693. doi:10.1074/jbc.M510042200
Movassat M, Forouzmand E, Reese F, Hertel KJ. 2019. Exon size and sequence conservation improves identification of splice-altering nucleotides. *RNA* **25:** 1793–1805. doi:10.1261/rna.070987.119
Murillo-de-Ozores AR, Rodríguez-Gama A, Carbajal-Contreras H, Gamba G, Castañeda-Bueno M. 2021. WNK4 kinase: from structure to physiology. *Am J Physiol Renal Physiol* **320:** F378–F403. doi:10.1152/ajprenal.00634.2020
Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J. 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods* **7:** 521–527. doi:10.1038/nmeth.1464
Pal S, Gupta R, Kim H, Wickramasinghe P, Baubet V, Showe LC, Dahmane N, Davuluri RV. 2011. Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res* **21:** 1260–1272. doi:10.1101/gr.120535.111
Palaniswamy R, Teglund S, Lauth M, Zaphiropoulos PG, Shimokawa T. 2010. Genetic variations regulate alternative splicing in the 5′ untranslated regions of the mouse glioma-associated oncogene 1, Gli1. *BMC Mol Biol* **11:** 32. doi:10.1186/1471-2199-11-32
Reixachs-Solé M, Eyras E. 2022. Uncovering the impacts of alternative splicing on the proteome with current omics techniques. *Wiley Interdiscip Rev RNA* **13:** e1707. doi:10.1002/wrna.1707
Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV. 2009. Evolution of alternative and constitutive regions of mammalian 5′ UTRs. *BMC Genomics* **10:** 162. doi:10.1186/1471-2164-10-162
Reyes A, Huber W. 2018. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res* **46:** 582–592. doi:10.1093/nar/gkx1165
Rogozin IB, Carmel L, Csuros M, Koonin EV. 2012. Origin and evolution of spliceosomal introns. *Biol Direct* **7:** 11. doi:10.1186/1745-6150-7-11
Romagnoli M, Mineva ND, Polmear M, Conrad C, Srinivasan S, Loussouarn D, Barillé-Nion S, Georgakoudi I, Dagg Á, McDermott EW, et al. 2014. ADAM8 expression in invasive breast cancer promotes tumor dissemination and metastasis. *EMBO Mol Med* **6:** 278–294. doi:10.1002/emmm.201303373
Sammeth M, Foissac S, Guigó R. 2008. A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol* **4:** e1000147. doi:10.1371/journal.pcbi.1000147
San-Cristobal P, Ponce-Coria J, Vázquez N, Bobadilla NA, Gamba G. 2008. WNK3 and WNK4 amino-terminal domain defines their effect on the renal Na^+-Cl^- cotransporter. *Am J Physiol Renal Physiol* **295:** F1199–F1206. doi:10.1152/ajprenal.90396.2008
Schlomann U, Rathke-Hartlieb S, Yamamoto S, Jockusch H, Bartsch JW. 2000. Tumor necrosis factor alpha induces a metalloprotease-disintegrin, ADAM8 (CD 156): implications for neuron-glia interactions during neurodegeneration. *J Neurosci* **20:** 7964–7971. doi:10.1523/JNEUROSCI.20-21-07964.2000
Shabalina SA, Spiridonov AN, Spiridonov NA, Koonin EV. 2010. Connections between alternative transcription and alternative splicing in mammals. *Genome Biol Evol* **2:** 791–799. doi:10.1093/gbe/evq058
Shabalina SA, Ogurtsov AY, Spiridonov NA, Koonin EV. 2014. Evolution at protein ends: major contribution of alternative transcription initiation and termination to the transcriptome and proteome diversity in mammals. *Nucleic Acids Res* **42:** 7132–7144. doi:10.1093/nar/gku342
Singh P, Ahi EP. 2022. The importance of alternative splicing in adaptive evolution. *Mol Ecol* **31:** 1928–1938. doi:10.1111/mec.16377
Srinivasan S, Romagnoli M, Bohm A, Sonenshein GE. 2014. *N*-Glycosylation regulates ADAM8 processing and activation. *J Biol Chem* **289:** 33676–33688. doi:10.1074/jbc.M114.594242
Sterne-Weiler T, Weatheritt RJ, Best AJ, Ha KCH, Blencowe BJ. 2018. Efficient and accurate quantitative profiling of alternative splicing patterns of any complexity on a laptop. *Mol Cell* **72:** 187–200.e6. doi:10.1016/j.molcel.2018.08.018
Talerico M, Berget SM. 1994. Intron definition in splicing of small *Drosophila* introns. *Mol Cell Biol* **14:** 3434–3445. doi:10.1128/mcb.14.5.3434-3445.1994
Tamarkin-Ben-Harush A, Schechtman E, Dikstein R. 2014. Co-occurrence of transcription and translation gene regulatory features underlies coordinated mRNA and protein synthesis. *BMC Genomics* **15:** 688. doi:10.1186/1471-2164-15-688
Tapial J, Ha KCH, 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 Res* **27:** 1759–1768. doi:10.1101/gr.220962.117
Tress ML, Abascal F, Valencia A. 2017. Alternative splicing may not be the key to proteome complexity. *Trends Biochem Sci* **42:** 98–110. doi:10.1016/j.tibs.2016.08.008
Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, Lynch KW, Barash Y. 2016. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *eLife* **5:** e11752. doi:10.7554/eLife.11752
Verma P, Thakur D, Pandit SB. 2024. Exon nomenclature and classification of transcripts database (ENACTdb): a resource for analyzing alternative splicing mediated proteome diversity. *Bioinform Adv* **4:** vbae157. doi:10.1093/bioadv/vbae157
Verta JP, Jacobs A. 2022. The role of alternative splicing in adaptation and evolution. *Trends Ecol Evol* **37:** 299–308. doi:10.1016/j.tree.2021.11.010
Wang R, Helbig I, Edmondson AC, Lin L, Xing Y. 2023. Splicing defects in rare diseases: transcriptomics and machine learning strategies towards genetic diagnosis. *Brief Bioinform* **24:** bbad284. doi:10.1093/bib/bbad284
Weber R, Ghoshdastider U, Spies D, Duré C, Valdivia-Francia F, Forny M, Ormiston M, Renz PF, Taborsky D, Yigit M, et al. 2023. Monitoring the 5′UTR landscape reveals isoform switches to drive translational efficiencies in cancer. *Oncogene* **42:** 638–650. doi:10.1038/s41388-022-02578-2
Xing Y, Yu T, Wu YN, Roy M, Kim J, Lee C. 2006. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res* **34:** 3150–3160. doi:10.1093/nar/gkl396
Yamamoto S, Higuchi Y, Yoshiyama K, Shimizu E, Kataoka M, Hijiya N, Matsuura K. 1999. ADAM family proteins in the immune system. *Immunol Today* **20:** 278–284. doi:10.1016/S0167-5699(99)01464-4

Zhao F, Yan Y, Wang Y, Liu Y, Yang R. 2023. Splicing complexity as a pivotal feature of alternative exons in mammalian species. *BMC Genomics* **24:** 198. doi:10.1186/s12864-023-09247-y

Zheng ZM. 2004. Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *J Biomed Sci* **11:** 278–294. doi:10.1007/BF02254432

Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D. 2009. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics* **10:** 47. doi:10.1186/1471-2164-10-47

Received August 2, 2024; accepted in revised form March 14, 2025.