Supplementary information for manuscript entitled

"Exon Nomenclature and Classification of Transcripts (ENACT) - Systematic framework to annotate exon attributes"

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

Running Title: ENACT - exon annotation framework

Keywords: Alternative splicing, exon nomenclature, exon inclusion, intron retention

*To whom correspondence should be addressed
Shashi Bhushan Pandit
Associate Professor
Bioinformatics Center,
Department of Biological Sciences
Indian Institute of Science Education and Research (IISER) – Mohali,
Knowledge City, Sector-81, SAS Nagar, Manauli PO 140306, India.
Email: shashibp@iisermohali.ac.in

S1: Exon architecture of human *GLRX2* gene

We present a case study of *GLRX2* gene isoforms with their composed EUIDs from ENACT (Figure S1). This case illustrates large independence in ENACT's exon centralization procedure while having seed representation from Reference isoform (RISO, having largest number of coding exons, see Methods in the main manuscript). In *GLRX2* gene, differential exon usage was noted in different organs [1, 2]. The expression of isoform NP_932066.1 (ENST00000367439) decreased two-fold, while that of isoform NP_057150.2's (ENST00000367440) increased significantly in testis tissue. Importantly, exons 1 and 2 are mutually exclusive (multiple exon alignment not shown). Both isoforms have the same exon counts, where NP_057150.2 consists of alternate coding exon-1 (Block-II: A) and NP_932066.1 as alternate coding exon-2 (Block-II: A). As NP_057150.2 is one amino acid longer than NP_932066.1, ENACT selects it as the reference. Due to this selection, the algorithm does not exclude alternate coding exon 2, even though it is absent in NP_057150.2; instead, this forms part of the reference set of exons (RSOEx) and is assigned a relevant ordinal position.

We also identified alternate exons that introduce different functional regions by utilizing ENACTdb. For example, NP_057150.2 (abundant in testis tissue) contains a disordered region absent in the other isoform, as shown in the Nightingale view of ENACTdb [3] in Figure S2-B.

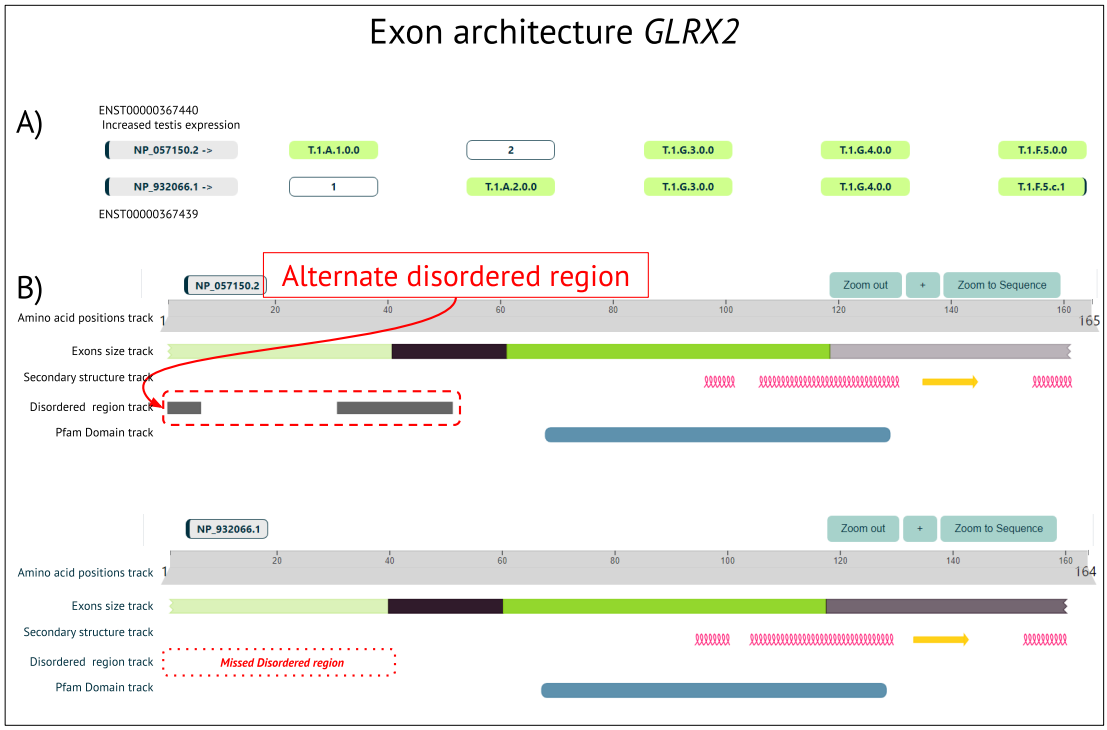


Figure S1: ENACT annotated exon architecture of human gene *GLRX2* and functionality inference from ENACTdb. A) Exon composition comparison between testis

overexpressed isoform NP_0057150.2 and pancreas overexpressed isoform NP_932066.1. B) Nightingale's view of ENACTdb with domain/disorder overlaid on exon features, highlighting likely functional changes introduced by alternate exons.

S2: Annotation of intron retention instance in human gene *AIF1*

In Figure S2, we show intron retention/exon fusion involving genomic coordinates of exons 4 and 5 or gene *AIF1* in a single row and that of fused exon construct in another row. Genomic coordinates have been shown using the vertical bar. Other exons have not been shown in Figure S2 for clarity. ENACT gives the following notation to this fused exon (background color only to distinguish attributes):

R:1:T.1.A.4.0.0:0:T.1.A.5.c.1

- The first letter 'R' (cyan background) denotes these instances as intron retention cases. (Other block-1 instances were U: UTR exons, T: CDS exons, D: exons intervening in UTR and CDS, and M: coding exons having only '1nt' as coding genomic coordinate).
- Following this, 1 (green background) indicates local protein-coding potential (This is local scope of Block-1, where value -2 indicate no coding genomic coordinate and amino acid ('aa') sequence, -1 means coding genomic coordinate but no 'aa' sequence, 0 means coding genomic coordinate of only 1 nt and hence will not have dedicated 'aa' sequence, 1 means it does have coding genomic coordinates and assigned 'aa' sequence and >1 values mean different 'aa' sequence than reference instance). (For other Block-I notations, see Figures 1 and 6 in the main manuscript)
- Yellow EUID indicates former exon, from where intron retention begins
- Magenta value 0 indicates the 0^{th} instance of intron retention beginning from this yellow EUID exon; had another retention starting from exon 4, this value would have been incremented by 1.
- Red EUID indicates to which exon this exon has fused/retained the intronic part, which in this case is exon 5.

The value 5.c.1 in red EUID indicates splice site alteration in the last exon. As illustrated in the figure, it can be noted that genomic coordinates for this entity extended right from the exon 5 by 4 nt.

Comparing the amino acid ('aa') sequence for this entity, it can be seen in the figure that this fusion retained the original frame and added a yellow highlighted sequence. The functional context of this retention and added sequence can be studied

from ENACTdb, where it was found to extend contribution to disordered region (data not shown).

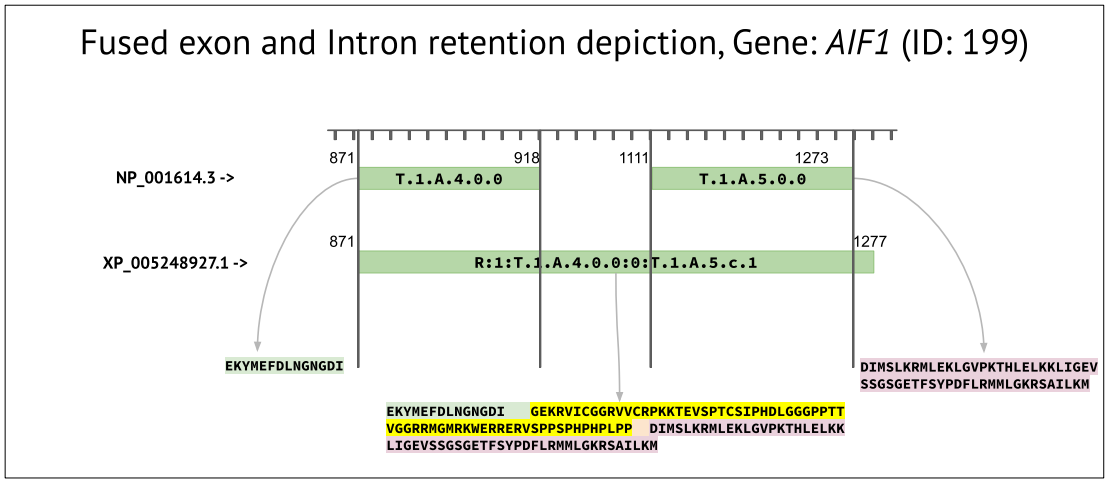


Figure S2. Depiction of Fused exon instance for two separate exons at ordinal position 4 and 5 in human gene *AIF1*. ENACT depicts such cases as 'intron retention' (exon fusion) between exon instances 4 and 5 in row 1 from Isf NP_001614.3. This exon is larger than exon 4 and 5 as it has retained the intron region between them. Vertical grey lines depict the splice boundaries for exons 4 and 5. The amino acid sequence, for instance, 4, has been written with a light green background, and exon 5 has been written with pink background. The fused exon sequence matches instances 4 and 5, where unique additions of amino acids are written with yellow background.

S3: Comprehensive illustration of ENACT algorithm

This section provides a detailed and comprehensive demonstration of the ENACT algorithm using human *RUNX1T1* gene in Figure S3, focusing on annotating exons (multiple overlapping instances) located at the 5' region. This 5' region (N-terminal) undergoes complex splicing patterns as noted by RefSeq (https://www.ncbi.nlm.nih.gov/gene/862) and is also observed as a fusion in cancers [4]. Through Figure S3, we show how ENACT's handling of complex splicing patterns.

Gene *RUNX1T1* and comprehensive capture of exon variant relationships

A. Figure S3-A shows isoforms of gene *RUNX1T1*, highlighting variable exons and their variants at the 5' end of the gene (rectangular block, exon 4 to 8). The RISO isoform identifier is colored in purple, and its constituent exons, which compose the initial reference set of exons ($RSOEx_{RISO}$), are colored in light blue. Exons at ordinal positions 1-2 and 5-7 will not form part of $RSOEx_{RISO}$ initially, and their variants are added in subsequent algorithm steps. Before this, exons overlapping with RSOEx (exons 4', 8' and 8''), referred to as *OlEx,* are filtered and maintained under $Exon_{variant}$ for processing

to derive splice site relationships (splice site variants of n/c/b, or intron retention) to their *RSOEx* counterparts. Figure S3-B depicts how splice site overlap is computed based on genomic coordinates to determine whether only 5', 3', or both 3' and 5' splice sites vary in $Exon_{variant}$ compared to its corresponding overlapping exons in *RSOEx*. These variations are annotated as n, c, or b within the $Exon_{variant}$ set.

B. For exons not initially included in *RSOEx* or filtered into $Exon_{variant}$, the algorithm identifies them for appropriate inclusion in *RSOEx* and $Exon_{variant}$ sets based on their ordinal position relevance and length criterion. Exons in that non-overlapping exons set (*NolEx*) may exist as single genomic coordinate instances or have multiple splice site varying instances (*NolEx-B*). The *NolEx-B* set undergoes iterative processing (explained in Figure S3-C) to identify one exon as representative for *RSOEx* updation. At the same time, other overlapping entities are marked as splice site variants and updated to the $Exon_{varinat}$ set.

To identify representative exons among *NolEx-B*, the algorithm prioritizes overlap evaluation from smaller-sized exons and identifies all overlapping (based on genomic coordinates) exonic instances. Then, one representative ($Qualifier_{exon}$) is chosen using 'lmin' criteria in 'SelectExon' procedure. Other exon entities overlapping (based on genomic coordinates) to $Qualifier_{exon}$ are added to $Exon_{variant}$ set. For example,

- Iter1: shows the handling of exons at position 6. Exon-6' was selected, and its overlapping set (genomically overlapped exons (*GoEx*)) was constructed (having exon 6 and 5-6). Exon 6 from this *GoEx* is selected as representative and updated to *RSOEx*, while others are added to $Exon_{variant}$ set. These instances are now removed from *NolEx-B*.
- Iter2: Exon-5 (a single instance) is added to *RSOEx* and removed from *NolEx-B*.
- Iter3: the last set of overlapping exons (exon-7) is similarly processed. Instance 7 is chosen as representative, and the corresponding *GoEx* is moved to the $Exon_{variant}$ set.

This process continues till the *NolEx-B* list is exhausted. Subsequently, updated *RSOEx* are sorted based on their genomic coordinates and assigned ordinal positions.

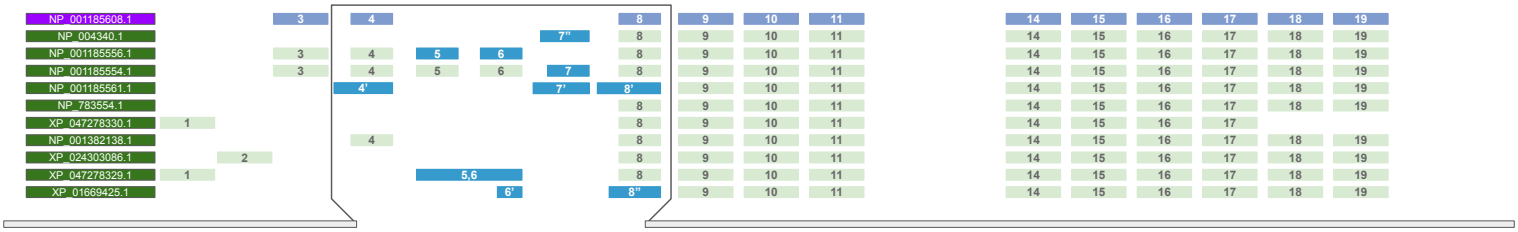
The selection of reference isoform (RISO) plays an important role here. The greater the number of exons in *RSOEx* (populated initially from RISO), the smaller the computational cost for running the 'selectExon' procedure to update the *RSOEx* later. Therefore, we preferred to select RISO, which has the highest number of coding exons.

C. Figure S3-D shows the relationships between updated *RSOEx* and $Exon_{variant}$ sets, illustrating how comprehensive Block-III attributes are notated as n/c/b splice variants for $Exon_{variant}$ set from corresponding *RSOEx*, and track of each unique such variation is accounted for ordinal position. Exons overlapping to two *RSOEx* instances (e.g., exon 5-6) based on genomic coordinates are annotated as intron retention events. (also specified in Figure S2).

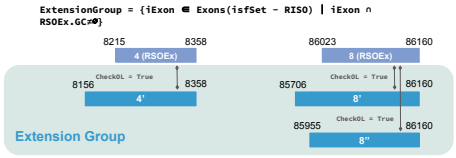
In the next step, prevalence is assessed based on the exon's occurrence in isoforms. Exons present in all isoforms are annotated constitutive (G), those in some isoforms as alternate (A), and those in all isoforms but with splice site variations as constitutive-like/facultative (F) (Figure S3-E).

ENACT procedural depiction to annotate N-Ter exon cluster of *RUNX1T1* gene

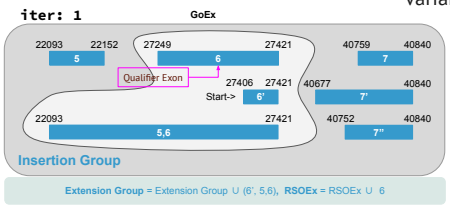
A Exon architecture *RunX1T1*

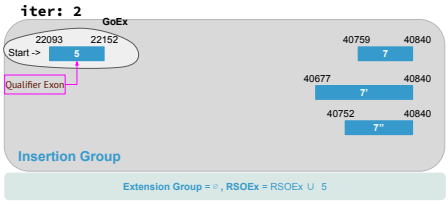


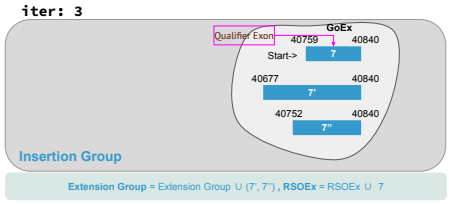
B Filtering Exon$_{variants}$ for RSOEx$_{RISO}$



C Updating RSOEx and Exon$_{variant}$



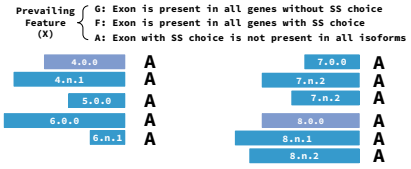




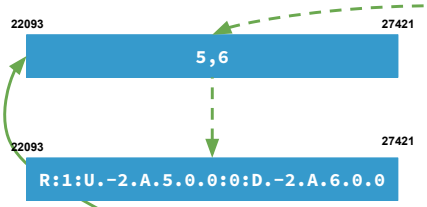

```
Nolex = {iExon ∈ Exons(NRExons_isfSet - RSOEx_RISO) | iExon ∩ RSOEx.GC = ∅}
Nolex.lengthSort()
while Nolex ≠ ∅;
    start = Nolex[0]
    GoEx = {iExon ∈ Nolex[1:] | start.GC ∩ iExon.GC ≠ ∅}
    if len(GoEx)==0: #NolexA case
            RSOEx = RSOEx ∪ start
    else: #Nolex-B case & 'selectExon' procedure
            GoEx.lengthSort()
            exons_grt_30 = {s∈GoEx|len(s)>29}
            if exons_grt_30 = ∅: #no element in GoEx with length longer 29 nt
                Qualifier_exon = argmax.length(GoEx)
            else:
                Qualifier_exon = argmin.length(exons_grt_30)
            RSOEx = RSOEx ∪ Qualifier_exon

    InsertionGroup = InsertionGroup - start - GoEx
    ExtensionGroup = ExtensionGroup ∪ (GoEx - Qualifier_exon)
    RSOEx.GC = update(RSOEx.GC)
```

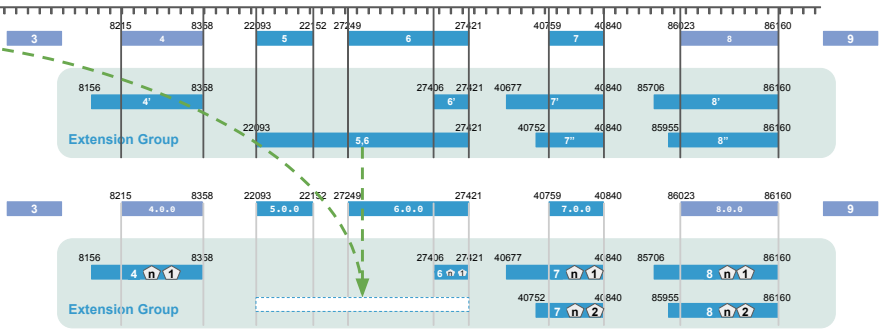
E Prevailing feature (Block-II)



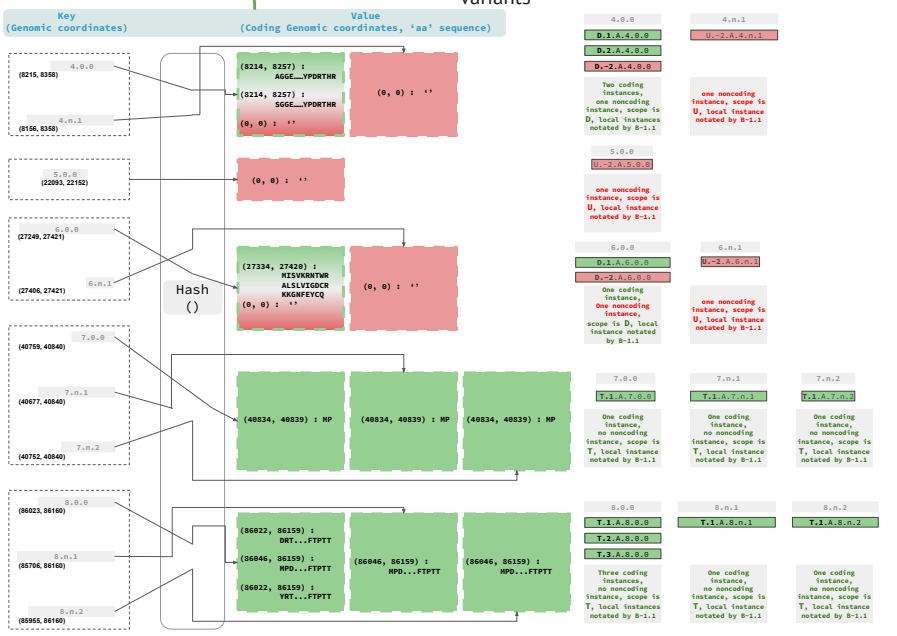
F Intron retention annotation



D Depicting splice site relationship between Exon$_{variant}$ and RSOEx



G Coding potential map of Exon$_{variants}$ and RSOEx



H Representation of cluster in ENACTdb

NP_001185608.1 ->	D.1.A.4.0.0	5	6	7	T.1.A.8.0.0
NP_004340.1 ->	4	5	6	T.1.A.7.n.2	T.1.A.8.0.0
NP_001185556.1 ->	D.-2.A.4.0.0	U.-2.A.5.0.0	D.1.A.6.0.0	7	T.1.A.8.0.0
NP_001185554.1 ->	D.-2.A.4.0.0	U.-2.A.5.0.0	D.-2.A.6.0.0	T.1.A.7.0.0	T.1.A.8.0.0
NP_001185561.1 ->	U.-2.A.4.n.1	5	6	T.1.A.7.n.1	T.1.A.8.0.0
NP_783554.1 ->	4	5	6	7	T.1.A.8.n.1
XP_047278330.1 ->	4	5	6	7	T.3.A.8.0.0
NP_001382138.1 ->	D.2.A.4.0.0	5	6	7	T.1.A.8.0.0
XP_024303086.1 ->	4	5	6	7	T.2.A.8.0.0
XP_047278329.1 ->	4	R:1:U.-2.A.5.0.0:0:D.-2.A.6.0.0		7	T.1.A.8.0.0
XP_016869425.1 ->	4	5	U.-2.A.6.n.1	7	T.1.A.8.n.2

Figure S3: ENACT algorithm steps showing human *RUNX1T1* gene annotation. This figure illustrates ENACT algorithm decision steps for exons 4-8 in the 5' region of the *RUNX1T1* gene. A) Selected isoforms with their NCBI protein identifiers show several exonic variations. RISO (purple, with maximal number of coding exons) includes exons in light blue, while other exons (from other isoforms) are in light green. Panels from B-D depict step-by-step details. Genomic coordinates for relevant exon(s) are shown to infer overlapping segments. Exons from isoforms other than RISO are compared with $RSOEx_{RISO}$ (initial RISO exons) to find genomically overlapped entities (processing step discussed in B panel) and non-overlapping (*NolEx*) set (panel C details various stages of detailing them). B) Overlapping exons to $RSOEx_{RISO}$ are first extracted and added to the $Exon_{variant}$ set. C) *NolEx* set may have singleton non-overlapping exons (*NolEx-A*) or multiple exons overlapping among themselves (*NolEx-B*). In this example, we only have the *NolEx-B* set. This set is processed iteratively to select representative exon for *RSOEx* till *NolEx-B* is empty. In the iter-1 step, the smallest exon-6' is chosen to identify overlapping cases (genomically overlapped exons, (*GoEx*)), which are exon 6 and 5-6. Using the *selectExon* module (pseudocode shown inset), a representative exon-6 is selected for *RSOEx*. Following the same approach, the Iter-2 step shows steps in choosing exon-5 and iter 3 for exon-7 to select a representative entity for *RSOEx*. After each iter and selection of *RSOEx* representative entity, remaining exons from GoEx are added to $Exon_{variant}$ set, as shown in cyan colored section immediately below iter box D) Shows relationships between *RSOEx* and $Exon_{Variant}$ set and illustrates n/c/b splice variant annotation for $Exon_{variant}$ group. Exon-5-6 overlapping with two *RSOEx* instances is called intron retention case annotation discussed in the F panel. E) After noting splice site variants, the prevalence feature is determined from the occurrence of exons in listed isoforms and based on annotated as constitutive/alternate/constitutive-like. F) Shows annotation step of intron retention (IR). G) Association of translational attributes of exons 4-8 are shown where coding genomic coordinate and 'aa' sequence are sourced for every exon instance from all its isoform occurrences. Hash maps associate coding genomic coordinates and amino acid sequence contributions with genomic coordinates as unique keys. This collectively defines exons' noncoding/coding and dual status, and all attributes are consolidated in EUIDs. H) Isoform composing these exons are shown with updated EUIDs.

D. Figure S3-G: The penultimate step involves determining each exon's amino acid coding potential and status, yielding block-I protein-coding scope establishment and variability assessment. Each exon and its variants (*RSOEx* + $Exon_{variant}$) are assessed for amino acid coding status as coding/noncoding or dual. For example:
    a. Exon-6: Its reference instance (6.0.0) occurs in two isoforms, where it is part of CDS (has coding genomic coordinates) in one transcript and UTR in the other, so it is assigned as dual ('D') global scope in Block-I. Conversely, $Exon_{variant}$ of exon-6 that undergoes 5' splice site

alteration (6.n.1) is noncoding in its occurring isoforms and is assigned 'U' global scope in Block-I.

Exons contributing varying amino acid sequences in different isoforms are assigned unique numeric codes as Block-1 local scope. For example,

- Coding and noncoding instances at reference exon 6 are differentiated with block 1 local scope values '1' and '-2'.
- Similarly, exon instances 4.0.0 and 8.0.0 differ in coding contributions and yield at least 2 different 'aa' sequences. ENACT uses Block-I code >1 to distinguish them and auto increments when encountering every such unique instance.

At this stage of assigning protein-coding scope attribute (Block-1), additional redundancy expectation (different 'aa' sequence and coding genomic coordinates in unchanged genomic coordinates) was circumvented by notating local scope values. These are mapped using the hash map to genomic coordinate, representing *RSOEx* and $Exon_{variant}$ (See Figure S3-G). More relationship depiction at the ordinal position and transition between CDS and UTR through splice site variations has been discussed comprehensively in Figures 3 and 6 of the main manuscript.

Through this example, we demonstrate the robustness of our algorithm in capturing complex exon relationships. We also show how ENACT handles redundant annotations, where it tracks different splicing variations (at each ordinal position) and intron retention while encompassing protein-coding potential.

S4: Overview of ENACT Database resource

Using our nomenclature, we have annotated exons of five widely studied model organisms, *viz. Caenorhabditis elegans, Drosophila melanogaster, Danio rerio, Mus musculus,* and *Homo sapiens* and documented them in the ENACT resource database (ENACTdb, [3]). Database is publicly available at https://www.iscbglab.in/enactdb/. Table S1 summarizes the number of annotated exons/transcripts of genes encoded in five organisms available in ENACTdb.

Table S1: Summary of gene/transcript/exon statistics in five model organisms.

Organism	Number of protein-coding	Number of transcripts	Number of exons

	genes		
C. elegans (Ce)	19,972	28,534	1,25,054
D. melanogaster (Dm)	13,972	30,755	65,958
D. rerio (Dr)	26,374	48,821	2,68,035
M. musculus (Mm)	22,134	92,400	2,32,520
H. sapiens (Hs)	20,443	1,30,739	2,41,910

References

1. Tung KF, Pan CY, Chen CH, Lin WC: **Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset.** *Sci Rep* 2020, **10:**16245.
2. Lonn ME, Hudemann C, Berndt C, Cherkasov V, Capani F, Holmgren A, Lillig CH: **Expression pattern of human glutaredoxin 2 isoforms: identification and characterization of two testis/cancer cell-specific isoforms.** *Antioxid Redox Signal* 2008, **10:**547-557.
3. Verma P, Thakur D, Pandit SB: **Exon Nomenclature and Classification of Transcripts database (ENACTdb): A resource for analyzing alternative splicing mediated proteome diversity.** *Bioinformatics Advances* 2024.
4. Grinev VV, Barneh F, Ilyushonak IM, Nakjang S, Smink J, van Oort A, Clough R, Seyani M, McNeill H, Reza M, et al: **RUNX1/RUNX1T1 mediates alternative splicing and reorganises the transcriptional landscape in leukemia.** *Nat Commun* 2021, **12:**520.