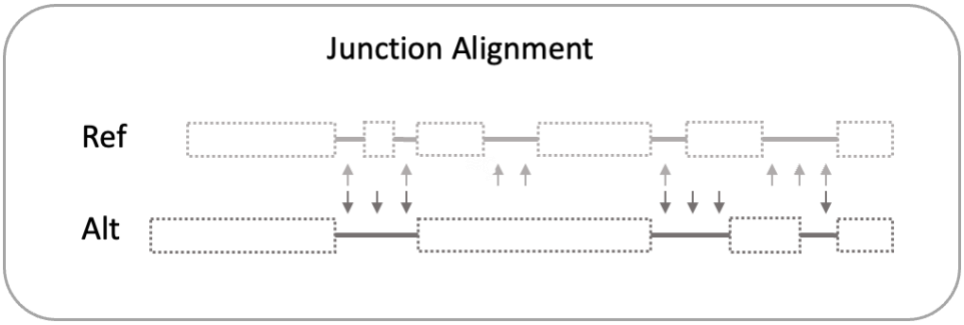
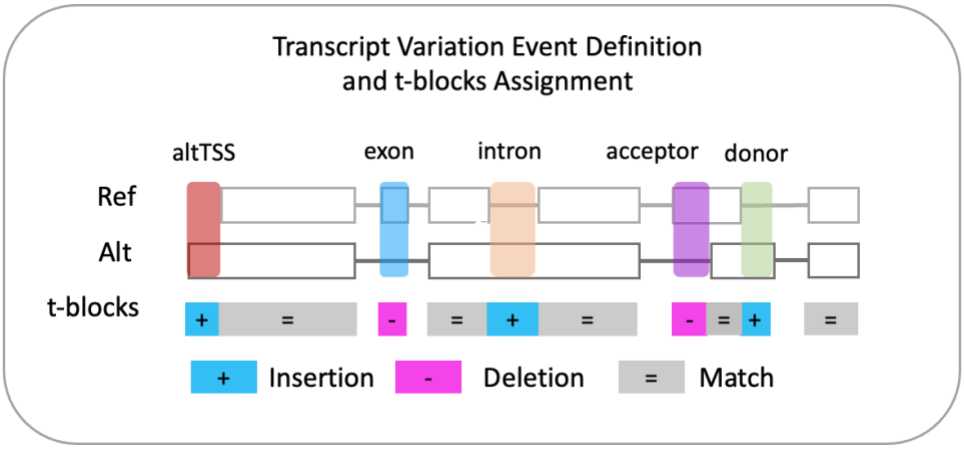
SUPPLEMENTAL FILE 1

BIOSURFER FOR SYSTEMATIC TRACKING OF REGULATORY MECHANISMS LEADING TO PROTEIN ISOFORM DIVERSITY

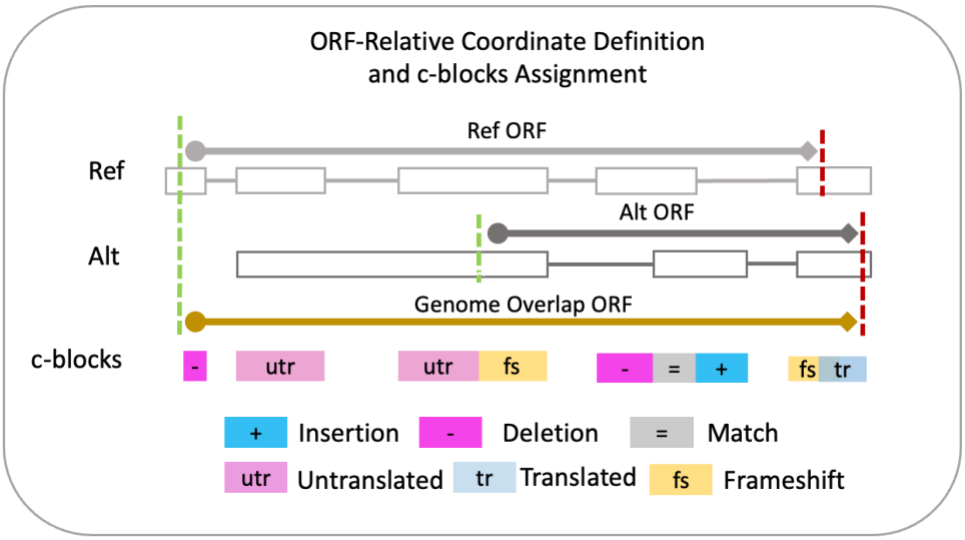
How does Biosurfer multi-layered alignment work? A conceptual workflow:



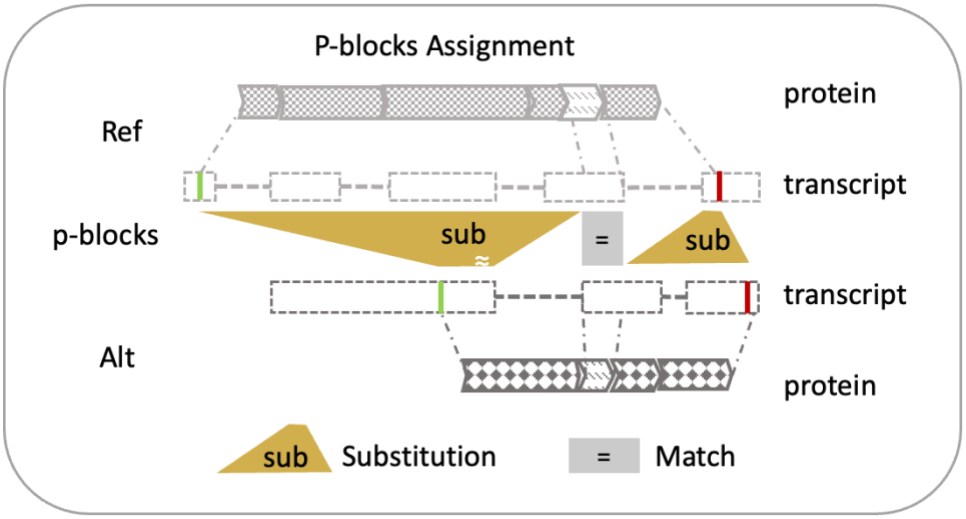
- **Identify Splicing Junctions:** Locate the points where splicing occurs, creating a connection map between the reference and alternative sequences.
- **Match Overlaps:** Connect junctions from the reference sequence to those in the alternative sequence when they share common segments.



- **Define Transcript Variation Events**: Catalog the types of changes observed, such as variations in splicing or differences in the starting points of transcription.
- **Assign t-Blocks**: Group the identified variations into t-blocks, defining each based on the nature of the sequence change - whether it is a Match (unaltered sequence), Deletion (sequence absence), or Insertion (additional sequence).



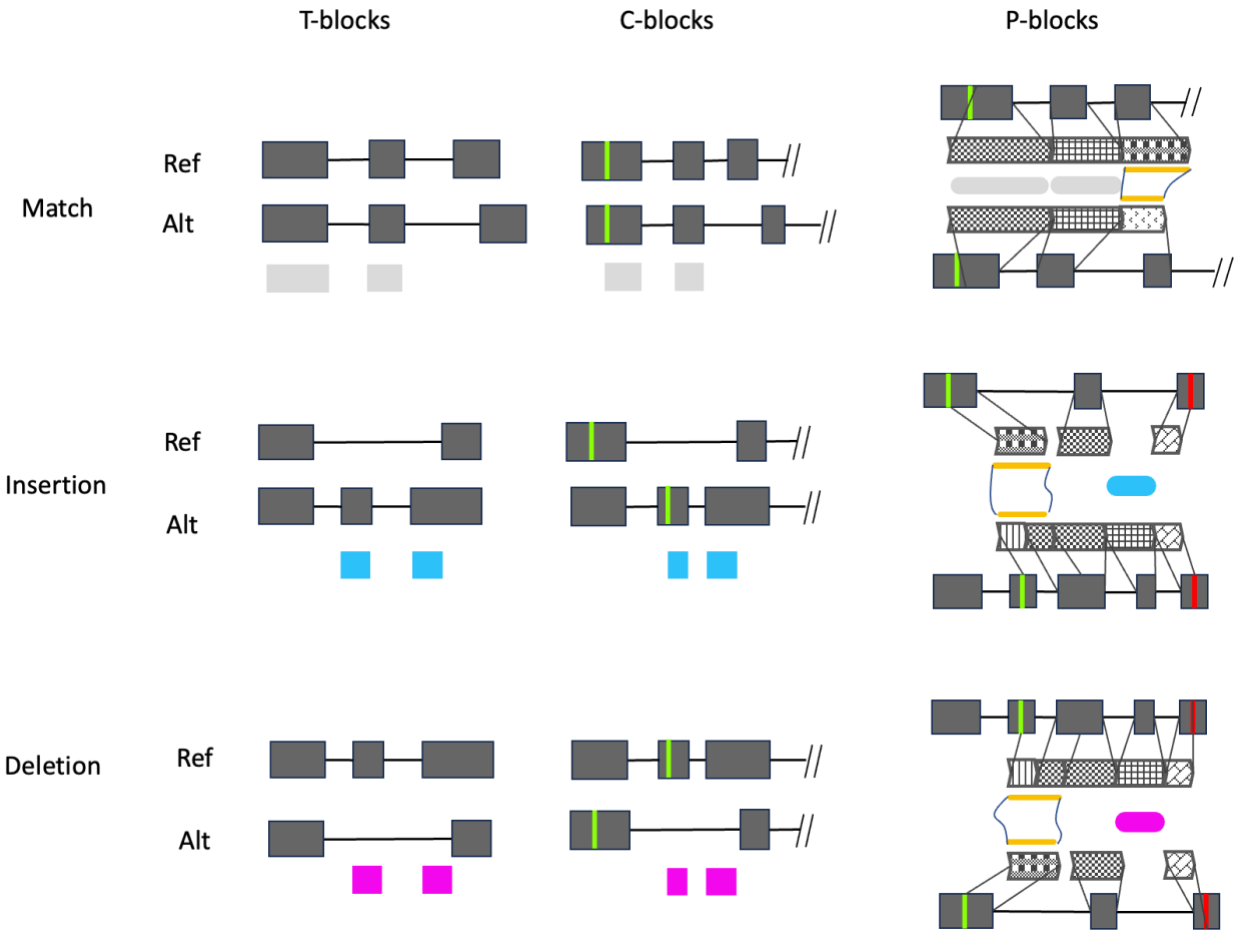
- **Define ORF Coordinates:** Create a zero-based coordinate system centered around the open reading frame (ORF) of a gene, with the start codon marked as the origin. This system provides a consistent framework to measure and compare the genetic sequences of different transcripts.
- **Assign c-Blocks:** Organize the genetic sequences into c-blocks based on their alignment with the ORF coordinates into Match, Deletion, or Insertion. Categorize contiguous codon sequences based on whether they are translated or untranslated, and examine each c-block for frameshifts, which occur when insertions or deletions change the reading frame of the codons.



- **Assign p-Blocks:** Group together sections of amino acids to identify which parts of the proteins are identical and which have variations. This creates a clear picture of how protein isoforms compare, highlighting the differences that could affect their functions.

How does Biosurfer multi-layered alignment work? A list of toy examples:

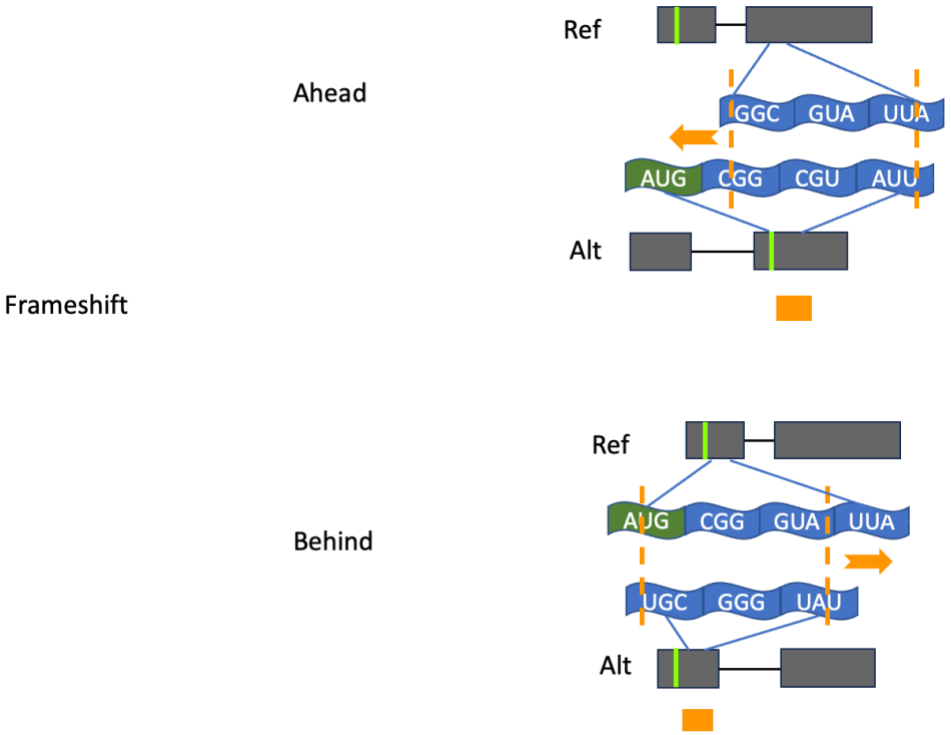
Simple Categories



For simple events, t-blocks, c-blocks and p-blocks all have the categories in Match, Insertion, or Deletion, designated by colored blocks. In P-blocks, substitution arises from a combination of insertion/deletion/frameshift events found at the c-block level, which is marked in yellow bars. Note that the yellow substitution p-blocks in the third column are only shown for context. They are not the simple events we are displaying here.

Translational Status and Frameshift (c-blocks Only)

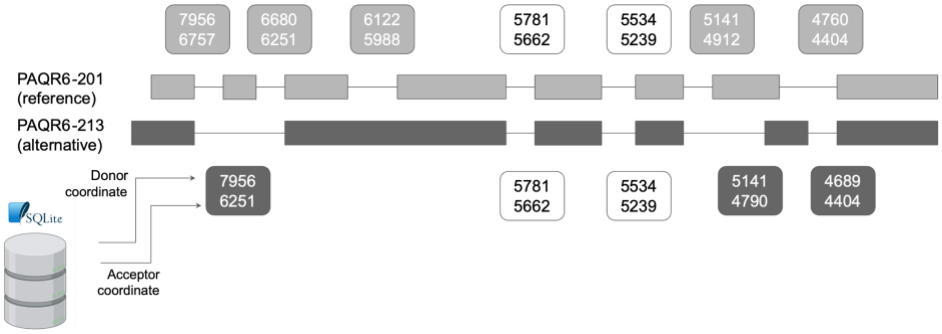
In c-blocks, translational status is assigned based on the ORF-relative coordinates system. Untranslated is when codon is not translated in the alternative isoform. Translated is when codon is not translated in the reference isoform.



Frameshift is a status of paired codon when the translational frame is one nucleotide ahead or behind between alternative and reference isoforms. Ahead is when codons in the alternative isoform are positioned one nucleotide downstream. Behind is when codons in the alternative isoform are positioned one nucleotide upstream.

How does Biosurfer multi-layered alignment work – a detailed, real example:

Illustration of the hybrid alignment steps for PAQR6 gene.

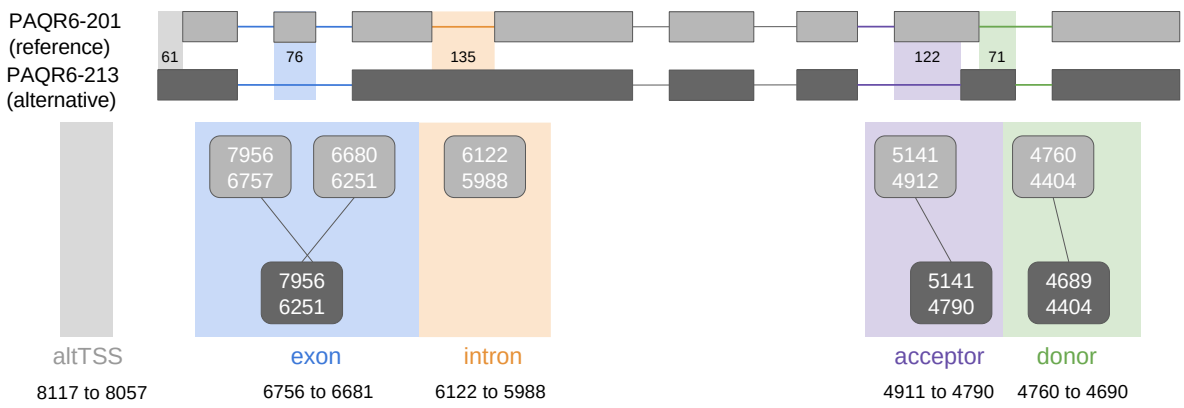


Step 1: Splicing junctions between the reference and alternative isoform are defined.

After inputting a GTF file to instantiate a Biosurfer SQLite database, information about splice junction coordinates is retrieved from the SQLite database. The junctions serve as the nodes of a bipartite graph that will make up the transcript-centric layer. The range of each node represents the span of the junction or the intronic region. Coordinates 7,956-6,757 (on chromosome 1 in this case) represents the first node of the reference transcript PAQR6-201.

Step 2: Construction of the bipartite graph for the reference and alternative isoform junctions.

A junction node from the reference is connected to a junction node from the alternative isoform when they have overlapping ranges (i.e., when there are one or more nucleotides in common between two junctions/introns). Node pairs are removed if they have perfectly overlapping boundaries (e.g., 5,781-5,662 was removed because these junctions are found in both the reference and alternative isoform). Thus, the remaining junction nodes form connected components, which correspond to transcript variation events that are the basis for classification in the next step.

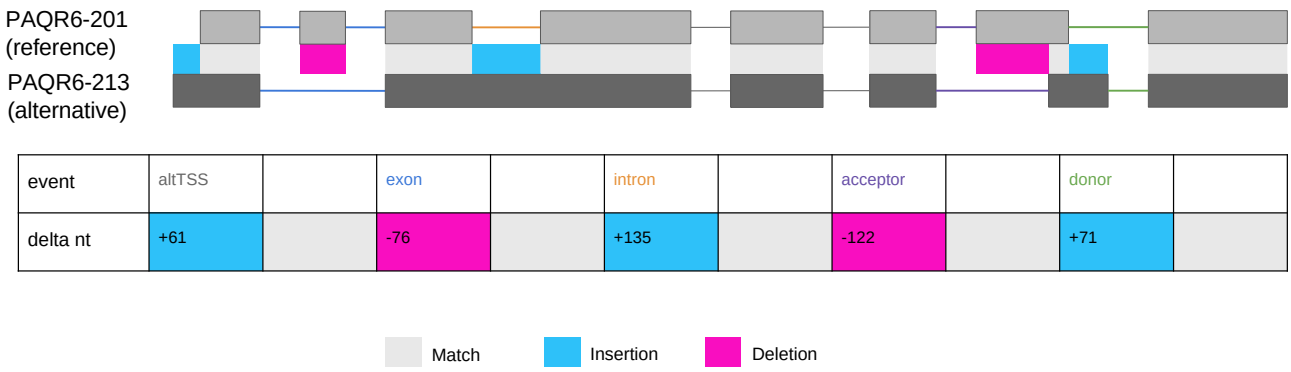


	event	altTSS	exon	intron	acceptor	donor
Genomic coordinates	coordinates	[8117, 8057]	[6756, 6681]	[6122, 5988]	[4911, 4790]	[4760, 4690]
	delta nt	+61	-76	+135	-122	+71
Transcript coordinates	PAQR6-201		[100, 176)		[734, 856)	
	PAQR6-213	[0, 61)		[289, 424)		[883, 954)

Step 3: Classification of transcript variation events.

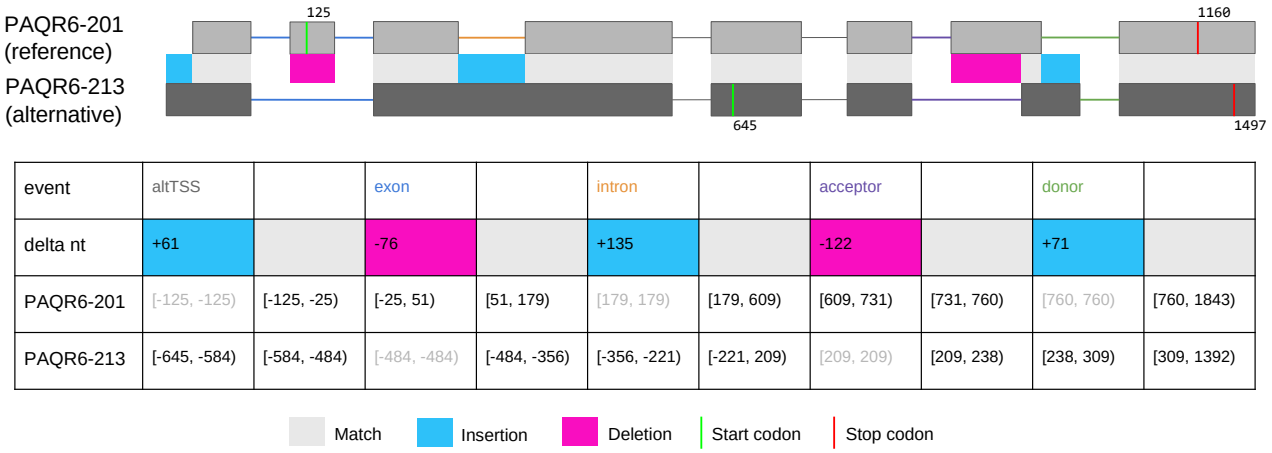
Next, transcript variation events are defined, which includes alternative splicing events (exon skipping, acceptor, etc.) as well as variations occurring at the ends of the transcript, such as alternative transcriptional start sites (altTSS, has coordinates 8,117-8,057). In the example above, there are five transcriptional events (altTSS, exon, intron, acceptor, donor). Note that linked splicing events, such as an included exon that is immediately followed by an alternative donor, are considered a single “compound” transcriptional event. For user-friendly interpretation of the positions of the transcript events, the splicing event coordinates are converted to 0-based coordinates that are relative to the transcript.

Each of these transcript variation events are considered altered transcript regions (or *t-block*, see **Step 4**), which is a contiguous chain of nucleotide sequence differences between the reference and alternative transcripts.



Step 4: Assignment of transcript alignment blocks or t-blocks.

Once the t-blocks are defined, based on the identity of the transcript variation events in **Step 3**, t-blocks are categorized as either a Match, Deletion, or Insertion.



event	altTSS		exon		intron		acceptor		donor	
delta nt	+61		-76		+135		-122		+71	
PAQR6-201	[-125, -125)	[-125, -25)	[-25, 51)	[51, 179)	[179, 179)	[179, 609)	[609, 731)	[731, 760)	[760, 760)	[760, 1843)
PAQR6-213	[-645, -584)	[-584, -484)	[-484, -484)	[-484, -356)	[-356, -221)	[-221, 209)	[209, 209)	[209, 238)	[238, 309)	[309, 1392)

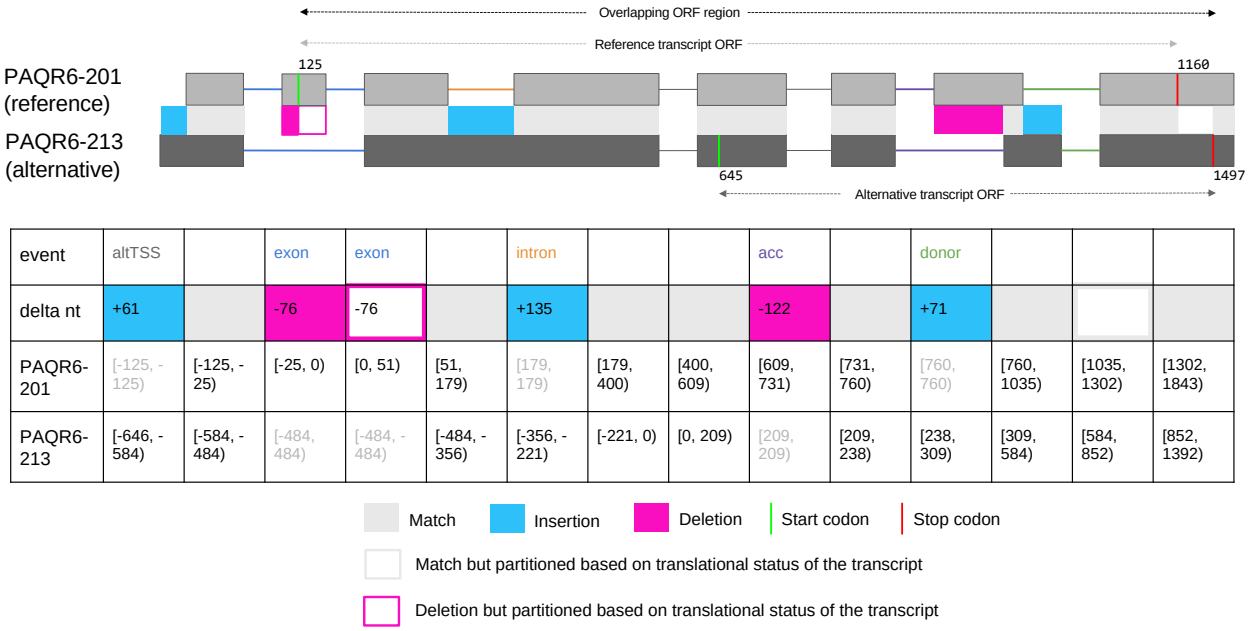
Match Insertion Deletion Start codon Stop codon

Step 5: ORF-relative coordinate system is defined based on the transcript coordinates.

In preparation for comparison of open reading frames (ORF) between isoforms, the first step is to compute 0-based coordinates based on the range of the ORF. In these 0-based ORF-relative coordinates, the first adenine of the AUG is considered index 0, and the last index is the last nucleotide of the stop codon (e.g., A of TGA).

Note that the ranges correspond to the 5’ UTR are negative, because the ORF ranges represent a relative coordinate system in which the 0 index is the reference point. Also note that the ranges near the 3’ UTR are not negative numbers.

Note: Step 6-9 represent intermediate stages required to fully classify all codon blocks (c-blocks).



event	altTSS		exon	exon		intron			acc		donor			
delta nt	+61		-76	-76		+135			-122		+71			
PAQR6-201	[-125, -125)	[-125, -25)	[-25, 0)	[0, 51)	[51, 179)	[179, 179)	[179, 400)	[400, 609)	[609, 731)	[731, 760)	[760, 760)	[760, 1035)	[1035, 1302)	[1302, 1843)
PAQR6-213	[-646, -584)	[-584, -484)	[-484, 484)	[-484, -484)	[-484, -356)	[-356, -221)	[-221, 0)	[0, 209)	[209, 209)	[209, 238)	[238, 309)	[309, 584)	[584, 852)	[852, 1392)

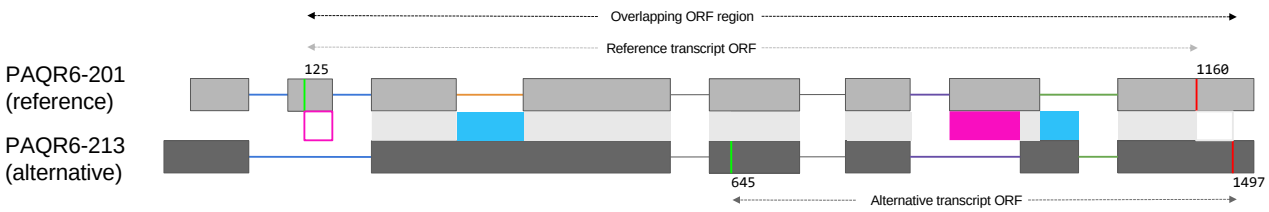
Match Insertion Deletion Start codon Stop codon

Match but partitioned based on translational status of the transcript

Deletion but partitioned based on translational status of the transcript

Step 6: Creation of temporary "translational status" t-blocks, in preparation for latter codon comparisons.

T-blocks can either be Match, Insertion, or Deletion segments, which is a classification based on sole comparison of transcribed regions. However, with the definition of an ORF, i.e., the range of the transcript that is translated into protein, these regions can be further sub-segmented and sub-classified, based on the translational status of the reference or alternative isoform region. In order to track this information, this step involves the creation of temporary objects—translation status-annotated t-blocks. For example, the ORF-relative coordinates [-25, 50) correspond to the second exon of the reference isoform. This exon contains the start codon; therefore, these regions are sub-segmented into the untranslated region [-25, 0) and the translated region [0, -51).



event	altTSS		exon	exon		intron			acc		donor			
delta nt	+61		-76	-76		+135			-122		+71			
PAQR6-201	[-125, -125)	[-125, -25)	[-25, 0)	[0, 51)	[51, 179)	[179, 179)	[179, 400)	[400, 609)	[609, 731)	[731, 760)	[760, 760)	[760, 1035)	[1035, 1302)	[1302, 1843)
PAQR6-213	[-646, -584)	[-584, -484)	[-484, 484)	[-484, -484)	[-484, -356)	[-356, -221)	[-221, 0)	[0, 209)	[209, 209)	[209, 238)	[238, 309)	[309, 584)	[584, 852)	[852, 1392)

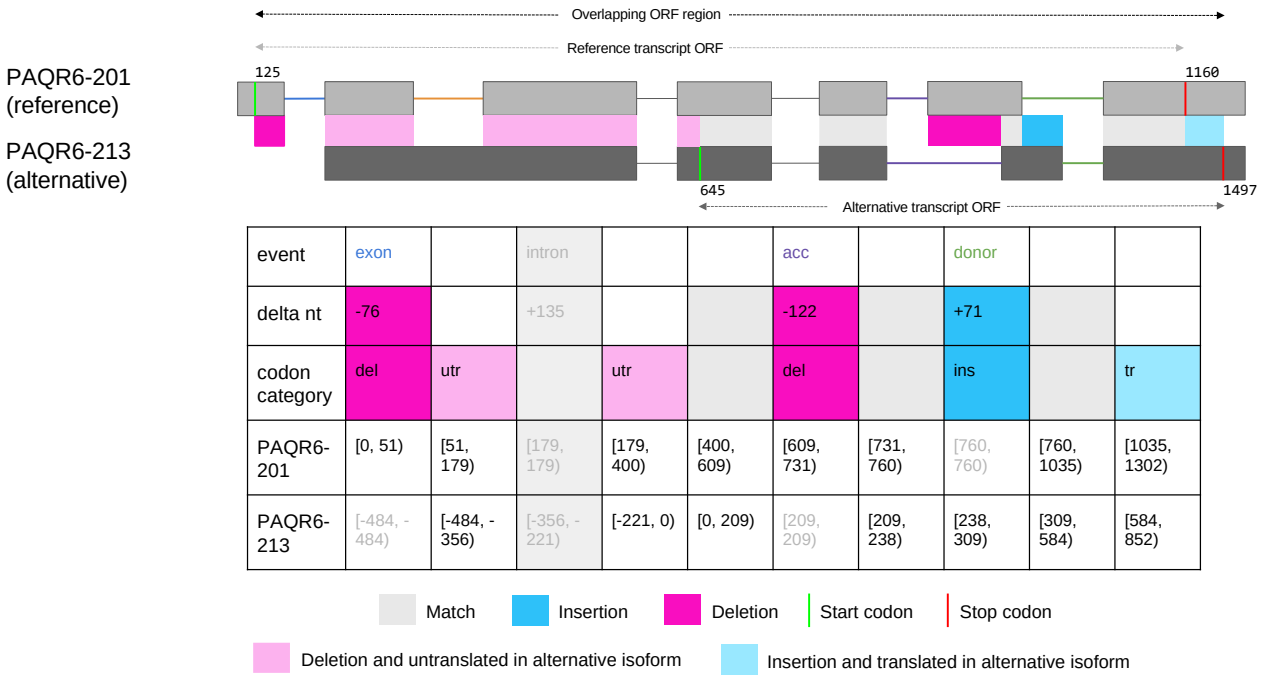
Match Insertion Deletion Start codon Stop codon

Match and partitioned based on translational status of the transcript

Deletion and partitioned based on translational status of the transcript

Step 7: Removing from consideration the t-blocks and translational status t-block objects that do not overlap either ORFs of reference or alternative isoform.

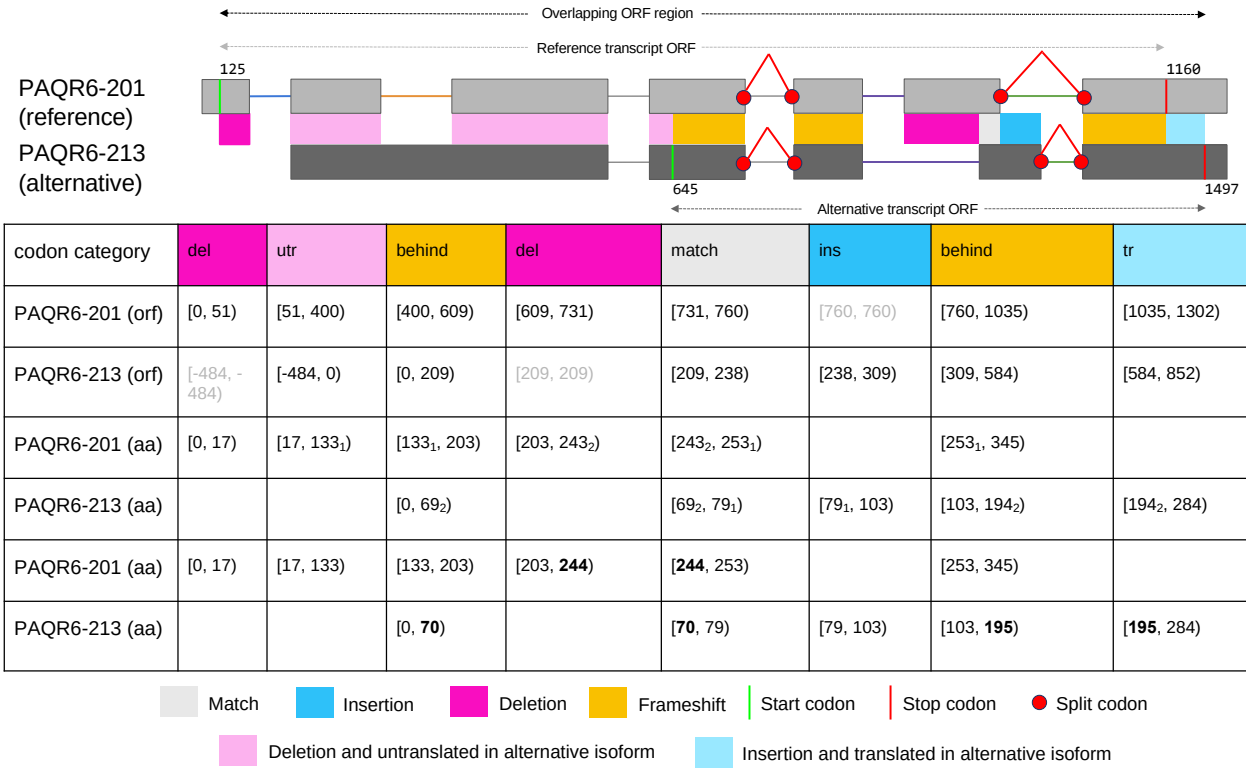
To proceed to the next step of classifying coding regions, the translational status t-blocks that are not within the bounds of a coding region (columns 1-3 and column 14 in the table above), are not considered in subsequent steps.



Step 8: Assignment of codon alignment blocks or c-blocks.

Based on the information of t-blocks and translational status t-blocks defined in **Step 6** and **Step 7**, c-blocks are assigned and categorized as Deletion, Insertion, Match, Untranslated, or Translated. A c-block is categorized as Untranslated if it contains a region that is coding in the reference isoform but remains untranslated in the alternative isoform. Conversely, a c-block is categorized as Translated if a region is transcribed but not translated in the reference (e.g., in the schematic above is the region between 1035 and 1302 of PAQR6-201, which is the 3' UTR of the reference (PAQR6-201), but is translated in the alternative isoform (PAQR6-213).

Note that these c-blocks represent contiguous ranges of codons that represent the same "type". In this step (Step 8), not all c-block types are shown. C-blocks also include individual codons that are split across junctions, which are handled in **Step 9**. In other words, c-blocks can include multiple, contiguous paired codons or a single paired codon.



codon category	del	utr	behind	del	match	ins	behind	tr
PAQR6-201 (orf)	[0, 51)	[51, 400)	[400, 609)	[609, 731)	[731, 760)	[760, 760)	[760, 1035)	[1035, 1302)
PAQR6-213 (orf)	[-484, -484)	[-484, 0)	[0, 209)	[209, 209)	[209, 238)	[238, 309)	[309, 584)	[584, 852)
PAQR6-201 (aa)	[0, 17)	[17, 133_1)	[133_1, 203)	[203, 243_2)	[243_2, 253_1)		[253_1, 345)	
PAQR6-213 (aa)			[0, 69_2)		[69_2, 79_1)	[79_1, 103)	[103, 194_2)	[194_2, 284)
PAQR6-201 (aa)	[0, 17)	[17, 133)	[133, 203)	[203, **244**)	[**244**, 253)		[253, 345)	
PAQR6-213 (aa)			[0, **70**)		[**70**, 79)	[79, 103)	[103, **195**)	[**195**, 284)

Match Insertion Deletion Frameshift Start codon Stop codon Split codon

Deletion and untranslated in alternative isoform Insertion and translated in alternative isoform

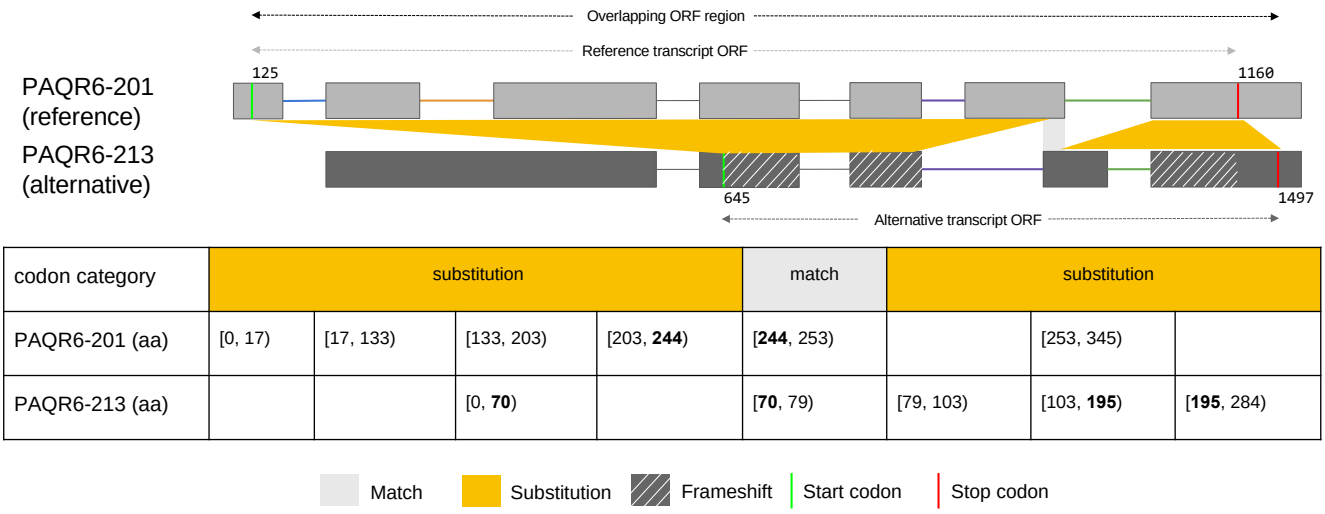
Step 9: Calling of the frameshift and split codon c-blocks (e.g., edge, complex).

Of the remaining contiguous range c-blocks, the relative frame of translation is determined. “Ahead” c-blocks are cases in which the codon is one phase ahead of the reference codon, in the c-block range. “Behind” c-blocks (two examples above) are cases in which the codon is one phase behind. Frameshifts are considered to occur when the length of the insertion or deletion t-block is not divisible by 3.

In **Step 8** and after assigning frameshift c-blocks, contiguous range c-blocks have been defined, but the single codons that reside on the edges of junctions are not yet categorized.

In **Step 8** and **Step 9**, the c-block layer is built up conceptually in a bottom-up manner in which codons that overlap by 2 or 3 of the 3 nucleotides are “paired” and then categorized. Paired codons that involve junction-spanning codons can have different relationships to each other topological (i.e., genomic coordinate) patterns or the identity of the nucleotide triplet or encoded amino acid. We defined all possible codon pairing patterns in an exact manner, finding distinct codon pair patterns, described in **Table 1**.

Note that these split codon c-block classifications relate to but are distinct from the concept of the codon phase, the number of nucleotides of a junction-spanning codon that resides in 3’ end of the spliced exon, which can be phase 0, 1, or 2.



codon category	substitution				match	substitution		
PAQR6-201 (aa)	[0, 17)	[17, 133)	[133, 203)	[203, **244**)	[**244**, 253)		[253, 345)	
PAQR6-213 (aa)			[0, **70**)		[**70**, 79)	[79, 103)	[103, **195**)	[**195**, 284)

Step 10: Assignment of protein alignment blocks or p-blocks.

To classify differences that are most relevant to protein functional differences, we defined p-blocks in a manner that is geared towards a protein-centric viewpoint. P-blocks arise from a global view of the transcript/codon-constrained alignment of AA and sub-segmenting the sequence into regions that match and do not match (c, Deletion, Substitution p-blocks) between the respective isoforms.

For the first pass of assigning p-blocks, the ordered chain of c-blocks is used to define groups of amino acids that correspond to each c-block, i.e., the precursor to p-block formation. Next, for all Match c-blocks, which represent exactly aligning regions of the two protein isoforms, a Match p-block is created. Then, for all consecutive non-match c-blocks (Insertion, Deletion, Ahead frameshift, Behind frameshift, Untranslated, Translated) are grouped in preparation for assignment of their p-block status. P-blocks that correspond to a single c-block that represents a single Insertion or Deletion will prompt creation of an Insertion or Deletion p-block, respectively*. The remaining groups of c-blocks represent all cases in which there is a “substitution” of one polypeptide for another. At this point, Substitution protein regions should map to a group of c-blocks that have at least one frameshift, or both an insertion and deletion.

After this first pass of assigning p-blocks, what is remaining are any cases of split codons (edge or complex, see **Table 1**). For these split codons, it is next determined if the accompanying AA pair is matched in identity—if so, then the split codon is then grouped with the Match p-block. Alternatively, for split codons that correspond with an AA pair that does not match in identity, it (the c-block) will be grouped with the adjacent non-match p-block (Insertion, Deletion, or Subsitution).

*Note: This allows for the possibility of deletions and insertion that contain ragged codons and thus do not represent pure deletion or insertion from the perspective of amino acid sequence.