Method

Biosurfer for systematic tracking of regulatory mechanisms leading to protein isoform diversity

Mayank Murali,[1] Jamie Saquing,[2] Senbao Lu,[3,4] Ziyang Gao,[3,4] Emily F. Watts,[2] Ben Jordan,[2] Zachary Peters Wakefield,[5,6] Ana Fiszbein,[5,6] David R. Cooper,[2] Peter J. Castaldi,[7,8] Dmitry Korkin,[3,4] and Gloria M. Sheynkman[2,9,10,11]

[1]*Broad Institute of MIT and Harvard University, Cambridge, Massachusetts 02142, USA;* [2]*Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, Virginia 22903, USA;* [3]*Bioinformatics and Computational Biology Program, Worcester Polytechnic Institute, Worcester, Massachusetts 01609, USA;* [4]*Computer Science Department, Worcester Polytechnic Institute, Worcester, Massachusetts 01609, USA;* [5]*Bioinformatics Program, Boston University, Boston, Massachusetts 02215, USA;* [6]*Department of Biology, Boston University, Boston, Massachusetts 02215, USA;* [7]*Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA;* [8]*Division of General Medicine and Primary Care, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA;* [9]*Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia 22903, USA;* [10]*Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia 22903, USA;* [11]*UVA Cancer Center, University of Virginia, Charlottesville, Virginia 22903, USA*

Long-read RNA-seq has shed light on transcriptomic complexity, but questions remain about the functionality of downstream protein products. We introduce Biosurfer, a computational approach for comparing protein isoforms, while systematically tracking the transcriptional, splicing, and translational variations that underlie differences in the sequences of the protein products. Using Biosurfer, we analyzed the differences in 35,082 pairs of GENCODE annotated protein isoforms, finding a majority (70%) of variable N-termini are due to the alternative transcription start sites, while only 9% arise from 5′ UTR alternative splicing (AS). Biosurfer's detailed tracking of nucleotide-to-residue relationships helps reveal an uncommonly tracked source of single amino acid residue changes arising from the codon splits at junctions. For 17% of internal sequence changes, such split codon patterns lead to single residue differences, termed "ragged codons." Of variable C-termini, 72% involve splice- or intron retention-induced reading frameshifts. We systematically characterize an unusual pattern of reading frame changes, in which the first frameshift is closely followed by a distinct second frameshift that restores the original frame, which we term a "snapback" frameshift. We analyze the long-read RNA-seq-predicted proteome of a human cell line and find similar trends as compared to our GENCODE analysis, with the exception of a higher proportion of transcripts predicted to undergo nonsense-mediated decay. Biosurfer's comprehensive characterization of long-read RNA-seq data sets should accelerate insights of the functional role of protein isoforms, providing mechanistic explanation of the origins of the proteomic diversity driven by the AS. Biosurfer is available as a Python package.

[Supplemental material is available for this article.]

Through mechanisms of alternative transcription, splicing, and polyadenylation, nearly every human gene can produce multiple protein products, with ~20K genes giving rise to at least 180K annotated isoforms (Frankish et al. 2023). The pathway from a gene to a protein is marked by several regulatory mechanisms that are highly tuned across development and cell states, with disruption of this regulation producing aberrant isoforms that lead to pathophysiological states such as cancer and cardiovascular disease (Cooper et al. 2009). Hence, approaches are needed to systematically characterize the upstream regulatory causes and functional impacts of such protein isoform sequence changes.

The transcript and, by extension, protein isoform diversity may now be globally characterized at great depth for individual samples (Glinos et al. 2022; Reese et al. 2023). Transcript diversity can be readily characterized by long-read RNA-seq, which employs single-molecule sequencing of individual cDNA or RNA molecules to determine the sequence across the entire length of spliced transcripts (Sharon et al. 2013; Workman et al. 2019; Tian et al. 2021; Joglekar et al. 2023; Pardo-Palacios et al. 2024), with platforms from Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) being most commonly used (Clarke et al. 2009; Eid et al. 2009). Long-read RNA-seq captures long-range connectivity between multiple exons of a transcript and can reveal complex splice patterns unattainable by short-read sequencing (De Paoli-Iseppi et al. 2021), including dependencies across distal splicing events (Anvar et al. 2018) and alternative 5′ and 3′ transcript usage. Given this readily characterized complexity afforded by long-read sequencing, a natural question is the extent to which such variations of the transcriptome lead to functional effects of the proteome. Toward this goal, a necessary step is defining the

Corresponding author: gs9yr@virginia.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279317.124.

35:1012–1024 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/25; www.genome.org

potential proteome. Both our group and others have reported methods in which long-read-derived transcript sequences serve as templates for predicting full-length protein isoform sequences, thereby providing a global snapshot of potential protein isoforms expressed in a particular biological condition (Miller et al. 2022; Veiga et al. 2022; Abood et al. 2024).

The complexity of alternative splicing (AS)—which for practical purposes in this manuscript, we define here as all transcriptional variations, including alternative transcription start sites (TSS) and transcription termination sites (TTS)—can be observed and characterized at different levels: RNA transcript, open reading frames (ORFs), and finally protein sequences (Reixachs-Solé and Eyras 2022). The complex interplay between the changes occurring to AS variants and their ORF and protein products is hard to characterize and quantify. Changes in mRNA sequence may lead to nonlinear or traditionally untracked variations. For example, a subtle splicing event could lead to a reading frame shift and thus to more marked changes to the C terminus of the protein than the originating small change at the RNA level would suggest. Or, AS could occur at codon boundaries, leading to altered amino acid (AA) identities of codons that technically overlap in genome space but are differentially "split" across exon–exon junctions. And complex interplay may also be observed between transcriptional variations and ORF choice, as alternative 5′ transcription or AS could lead to different availability of initiator codons, delimiting start codon choice co-translationally.

We argue that rather than being an esoteric exercise, the ability to characterize all potential interplay of RNA–ORF–protein variation is critical for fully elucidating the transcript and proteomic diversity encoded within long-read RNA-seq data sets. As long-read RNA-seq approaches are increasingly adopted in large-scale studies of hundreds of samples (Glinos et al. 2022; Reese et al. 2023) and are maturing into stable tools being adopted by the community (Pardo-Palacios et al. 2021), extracting all sources of biological molecular diversity is critical. Such interplay cannot be characterized by comparison of sequences using just one modality, such as transcript-focused annotation tools like SQANTI or Matt (Tardaguila et al. 2018; Gohr and Irimia 2019) (see Supplemental Text for a more extensive discussion of transcriptome analysis tools), or conventional protein sequence alignment tools like ClustalW (Chenna et al. 2003). Indeed, more recently, multimodal comparisons have been reported. For example, ORFanage is an approach for large-scale annotation of ORFs across predicted transcripts in the CHESS database, the main focus being optimal selection of ORFs based on protein alignments (Varabyou et al. 2023a,b). Other tools geared toward the phylogenetics community have developed frame-aware alignment in which the alignment scoring system includes penalties for frame-shift-inducing gaps (Ranwez et al. 2011; Evans and Loose 2015; Jammali et al. 2022). However, these tools do not comprehensively elucidate the interplay between transcript and protein variation.

To characterize how AS impacts the protein sequence, several bioinformatic tools and databases have been developed, such as VastDB, ASPicDB, ExonOntology, and DIGGER (Martelli et al. 2011; Tapial et al. 2017; Tranchevent et al. 2017; Louadi et al. 2021). Tools such as tappAS and IsoTV enable annotation of how protein isoform sequences differ and, in addition, infer their potential functional differences (de la Fuente et al. 2020; Annaldasula et al. 2021). For example, tappAS is a Java application for quantifying differential isoform usage but also to functionally annotate such isoforms, using the module IsoAnnot (de la Fuente et al. 2020). IsoAnnot maps protein features (e.g., Pfam domains) across isoforms of a gene, and determines how splicing leads to partial or full removal of protein features, indicating potential changes to molecular functions. Despite the existence of these tools, of need is the ability to systematically capture all possible effects to protein isoforms, with the accompanying information of the underlying RNA–protein relationships.

To fill this gap, we developed Biosurfer, a tool that enables one to compare related protein isoforms to each other while retaining the context of the upstream regulatory variations, which makes it unique among the landscape of published transcript and protein isoform analysis tools. Biosurfer is unique in that it is a computational pipeline that compares related protein isoform sequences and, in the process, tracks simultaneously the changes at all three levels, to understand the impact of AS on transcriptome, ORFeome, and proteome diversity. Biosurfer computes details not immediately apparent from genome annotation files or manual inspection in genome browsers, such as how between isoforms of the same gene the frame of translation and codon topology influences AA sequence identity changes. To accomplish this multilayered comparison, a genome is used as a "scaffold" to exactly position all nucleotides, codons, and AA elements, and associate local and context-dependent attributes with each element. The resulting data structure of three distinct yet interlinked layers inform on the upstream biological mechanisms leading to protein sequence changes. To demonstrate the utility of Biosurfer, we globally characterize variation across protein isoforms in the reference human annotation (GENCODE) and protein isoforms predicted from long-read RNA-seq of a human stem cell line. This characterization includes comprehensive elaboration of all sources of N-terminal, internal, and C-terminal protein variations observed, highlighting mechanisms of RNA–protein interplay.

Methods

Biosurfer package for isoform analysis and visualization

Biosurfer is a computational pipeline that performs a multilayered comparison between a pair of protein isoforms, in which differences at three different levels: RNA (nucleotides, nt), ORF (codon), and protein (AA residues, AA) are simultaneously tracked. The developed data structure enables not only the comparison of AAs, but the tracking of frameshifts, patterns of codon splitting at junctions, and the attendant upstream nucleotide and codon differences that explain AA changes. Such tracking aids in the systematic annotation of explanatory mechanism(s) underlying AA residue changes, such as whether a substitution of a stretch of AA residues is due to AS or a frameshift (Supplemental Fig. S1).

The Biosurfer pipeline is organized into three stages (Fig. 1). First, an SQLite database is populated with detailed information on each protein isoform at the associated transcript, ORF, and protein product levels. For each isoform, the required inputs are (i) a transcript FASTA, (ii) a protein FASTA, and (iii) a matching GTF with both Exon and CDS features. The inputs can either be extracted from the reference annotations (e.g., GENCODE [Harrow et al. 2012]), or they can be user-defined (e.g., predicted protein isoform sequences from long-read RNA-seq data [Miller et al. 2022] by using ORF callers such as CPAT [Wang et al. 2013] or Transdecoder [Haas et al. 2013]). Second, multilayered isoform-level alignments are generated, and each alignment is represented as three key data structures: t-blocks, c-blocks, and p-blocks corresponding to the view of the aligned isoforms at the transcript, codon sequence, and protein sequence levels, respectively (Supplemental Fig. S2).

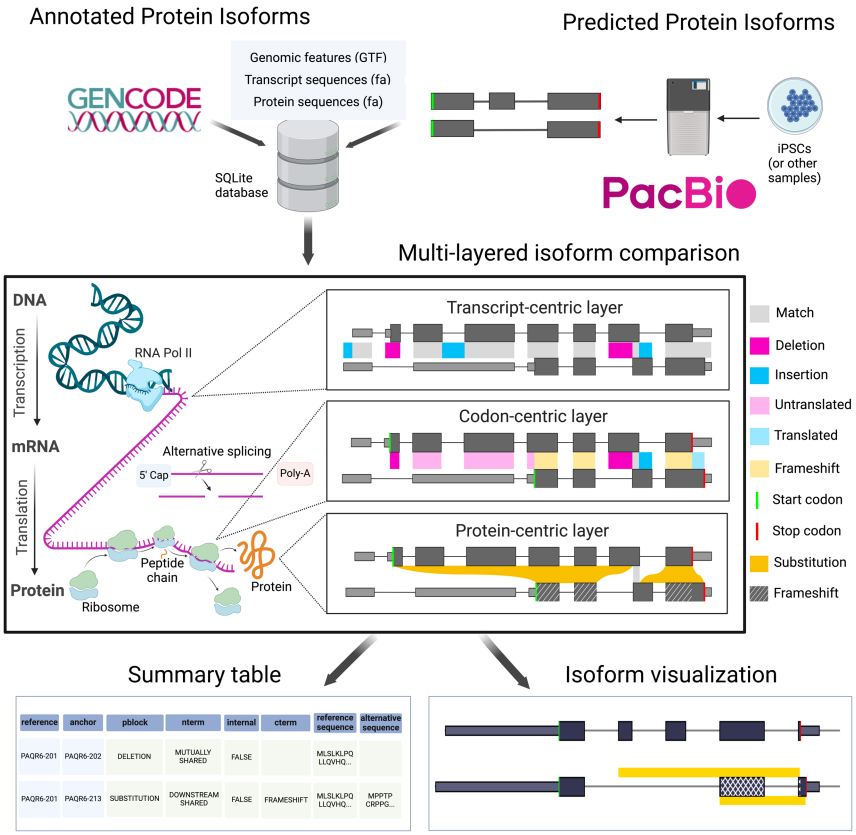


Figure 1. Biosurfer for analysis and visualization of protein isoform sequence differences. Biosurfer analyzes protein isoforms from reference annotations (e.g., GENCODE) or proteins predicted from long-read RNA-seq data. Analysis initiates with the creation of an SQLite database populated with isoform-relevant elements. Biosurfer performs a multilayered comparison of transcript-, codon-, and protein-level differences between pairs of protein isoforms. Variable regions, as well as their associated annotations, are output in tabular format and visualization files, which include protein-relevant details such as the frame of translation. Note that the terms "match," "deletion," etc., represent very different comparisons depending on the biological layer. (GTF) Gene transfer format; (iPSCs) induced pluripotent stem cells.

Third, all information is summarized in tabular format, to facilitate analysis of the transcription-level and codon-level mechanisms driving the proteomic diversity. In addition, a visual representation of such mechanisms integrated with the protein-level view of the isoforms can be output as PNG files.

Data structures for presenting transcript-, ORF-, and protein-level isoform alignments in Biosurfer

Biosurfer compares transcript-, ORF-, and protein-level sequences for each gene. One isoform is selected as reference and the other isoforms are denoted as alternative. The choice of which transcript (and associated isoform) is the "reference" is entirely up to the control of the user; however, the suggested default is the APPRIS principal isoform. When aligning each alternative isoform to the reference, the matched regions are referred to as matched *blocks*. The remaining regions on each sequence that are not matched blocks are referred to as unmatched blocks. See the Supplemental File for the step-by-step schema of the comparison process, which is summarized below.

Transcript blocks (t-blocks)

T-blocks represent subsegments of the transcript sequence that are shared or unique to the reference or alternative transcript. Transcript-level differences are determined by analyzing the transcript-to-genome coordinates alignment (GTF file) provided as an input by the user. Specifically, Biosurfer defines the aligned exonic regions that are shared or unique to each transcript. The resulting ranges, called *t-blocks*, are categorized as match, deletion, or insertion t-blocks. Deletion or insertion t-blocks are further annotated with the associated biological mechanism leading to the transcript nucleotide change, e.g., alternative transcriptional start site, splicing event, or polyadenylation. The AS events are then further classified into four basic types: retained intron, alternative donor, alternative exon, or alternative acceptor.

Codon blocks (c-blocks)

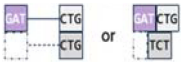
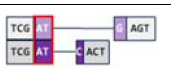
C-blocks represent a codon-centric layer defined through the comparison of the protein-coding regions of transcripts, i.e., ORF, between two isoforms. The c-block data structure is the most complex, but critical, layer in Biosurfer that connects information between the transcript and protein layers.

For ultimate granularity and precision, ORFs are compared based on the alignment of codons that overlap in the genome space, in which one codon in the reference isoform is compared with another codon in the alternative isoform (see Supplemental Methods for additional details). First, codons across the two isoforms are "paired" based on their mutual positions and base overlap: in a basic scenario, the two codons are identical, and their positions match (Table 1; Supplemental Fig. S3A); in a more complex scenario, the codons are split and only partially overlapped (1 or 2 bases, see Table 1; Supplemental Fig. S3B,C). In all other cases, the codons will be designated as "unpaired" (Table 1; Supplemental Fig. S3B). To keep track of unpaired codons within the data structure, they are linked to a "placeholder" codon, which serves to maintain the consistency of the c-block structures. Once aligned, each single codon pair is categorized with respect to multiple attributes that explain their relationship. Overall, the paired and unpaired codons are classified into nine categories based on their translation status, frameshift status, and codon topology (Table 1).

Protein blocks (p-blocks)

P-blocks represent a protein-centric layer defined through t- and c-block-guided comparisons of two protein isoforms, thus focusing solely on the relationships between the AA residues of the respective protein isoforms. This decision was motivated by the fact that in most cases, functional differences exhibited by an alternative protein isoform are attributable to differences in protein rather than nucleotide sequences between the alternative and reference isoforms, and these differences could be conceptually decoupled from the knowledge of the upstream (ORF- or transcript-level) residue-altering mechanisms.

Table 1. Categories of overlapping codon pairs that are the basis for codon blocks (c-blocks)

Codon pair status	General category	C-block category	Codon topology	Description	Positional identity	Schema
Paired	Simple	Match	Both: contiguous or split	Exact match in genomic position and nucleotide triplet identity	3/3 overlap	
Unpaired: empty "placeholder" codon		Deletion	Both contiguous	Codon missing in the alternative isoform	0/3 or 1/3 overlap	
		Insertion		Codon missing in the reference isoform	0/3 or 1/3 overlap	
Unpaired: "placeholder" codon that represents an untranslated position	Translational status	Untranslated	Both: contiguous or split	Nucleotide triplet not translated in the alternative isoform	0/3 or 1/3 overlap	
		Translated		Nucleotide triplet not translated in the reference isoform	0/3 or 1/3 overlap	
Paired	Frameshift	Ahead	Both contiguous	Alternative isoform codon one position downstream	2/3 overlap	
		Behind		Alternative isoform codon one position upstream	2/3 overlap	
	Split codon patterns	Edge[a]	Both split	Codons overlapping at the junction	Different	
		Complex[a]	Both split or 1 of the 2 codons split	Codons overlapping at the junction, and are frameshifted	Different	

[a]A subset of these paired codons can result in AA differences, or "ragged" codon pairs.

The *p-blocks* for a pair of isoforms are determined by translating c-blocks defined at the previous level. The resulting data structure represents subsegments of contiguous stretches of AA residues from each protein sequence. Each subsegment could be as short as a small translated portion of an exon (even a single residue, in the case of NAGNAG splicing), but it can also span multiple exons of a gene. Because the actual matching happened at the two previous levels (t-blocks and c-blocks), no alignment is required at the p-block level. The resulting protein subsegments can be either (1) fully matched (100% sequence identity, across all residues in the subsegment), or (2) mismatched, where a subsegment in one protein sequence will be aligned against a gapped region in the other sequence. Thus, a p-block represents either a matched pair of protein subsequences or a subsequence matched against a gapped region. P-blocks are then classified as match, insertion, deletion, or substitution changes. Substitution p-blocks must arise from a combination of insertion/deletion/frameshift events found at the c-block level.

Overall, these p-block changes are agnostic to the upstream mechanisms, e.g., at the transcript or ORF levels; however, at the same time, the corresponding upstream mechanisms can be retrieved and analyzed within Biosurfer, unlike with traditional protein aligners.

Criteria for classifying N-terminal, internal, and C-terminal alterations

To systematically categorize protein sequence alterations as N-terminal, internal, or C-terminal, we apply the following criteria:

An alteration is considered *N-terminal* if there are sequence differences at the N terminus of the protein, corresponding to one or more exons (with the first exon either fully or partially mapping to the N-terminal region). These differences must be due to an insertion, deletion, or substitution p-block. Importantly, there must be a match p-block downstream from these N-terminal differences, indicating that the protein sequences are identical beyond the altered N-terminal region.

An alteration is classified as *internal* when there is an insertion, deletion, or substitution p-block located within the protein sequence that does not include the N-terminal or C-terminal regions. This internal p-block must be flanked on both sides by match p-blocks, signifying that the sequences upstream and downstream from the alteration are identical.

An alteration is deemed C-terminal if there are sequence differences at the C terminus of the protein, involving one or more exons (with the last exon either fully or partially mapping to the C-terminal region). These differences must be due to an insertion, deletion, or substitution p-block. Additionally, a match p-block must be present upstream of these C-terminal differences, indicating that the protein sequences are identical before the altered C-terminal region.

Analysis of GENCODE isoforms using Biosurfer

The principal and alternative isoforms were defined using the APPRIS annotation of genes (Rodriguez et al. 2013). First, the set of APPRIS isoforms is identified for each gene from the input genome data (GENCODE v42, basic annotation [Frankish et al. 2021]) by extracting the key transcript features associated with the protein isoform, such as "transcript_id," "transcript_name," and the associated APPRIS "tag" (Rodriguez et al. 2013). Second, the transcript's APPRIS status is determined as "principal," "alternative," or "none," based on first rank-ordering of transcripts based on APPRIS tag and setting the transcript with the highest

APPRIS value as "principal" and all other transcripts as "alternative." We did not consider further genes with only one annotated isoform. Subsequently, we compiled a structured data set for each transcript, encompassing identifiers, gene associations, strand orientation, and APPRIS status.

Long-read RNA-seq analysis of WTC-II cell line

Total RNA from WTC-11 cells was extracted using the RNeasy Kit (Qiagen) and analyzed on an Agilent Bioanalyzer. We observed an RNA concentration of 30 ng/μL with the RNA Integrity Score (RIN) of 9.9. As described previously (Mehlferber et al. 2022), cDNA was synthesized from the extracted RNA and the Iso-Seq Express Kit SMRTbell Express Template prep kit 2.0 (Pacific Biosciences) was used on a Sequel II system to obtain long-read sequence information and output circular consensus (CCS) reads. Data are available at the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRR18130587 and previously published (de Souza et al. 2023).

We analyzed the WTC-11 data with a proteogenomics Nextflow pipeline we previously developed (Mehlferber et al. 2022; Miller et al. 2022). The output CCS reads from long-read sequencing were processed with Iso-Seq3 and SQANTI3 (version 1.3) for transcript classification and quality assessment. The CPAT (Wang et al. 2013) algorithm was used to predict ORFs, which were grouped into protein isoforms.

The Biosurfer code is publicly available (see Software availability).

Results

Characterization of altered protein regions in the human annotation (GENCODE)

Here, we demonstrate the utility of Biosurfer through genome-wide analysis of protein isoforms in the GENCODE annotation (basic annotation, version 42 [Frankish et al. 2021]). We analyzed 35,083 reference-alternative protein isoform pairs across 11,815 genes (Fig. 2A). Each pair consists of the reference protein isoform for a gene—the APPRIS principal isoform (Rodriguez et al. 2013)—and an alternative protein isoform. The number of isoform pairs corresponds to the number of "alternative" isoforms for a gene (Supplemental Table S1). Genes with only one annotated isoform (8166 of 19,981) were excluded from the analysis (Supplemental Table S2).

Globally, we found a total of 44,326 altered protein regions with an average of 1.3 altered regions per isoform. Note that we are using the term altered protein region, in this section of the manuscript, to refer to p-blocks that encode a change in protein sequence (i.e., a deletion, insertion, or substitution p-block, as in a nonmatch p-block) (see Methods). Altered protein regions are contiguous regions of altered protein sequence relative to the "reference" (i.e., APPRIS principal) protein, which includes p-block insertions, deletions, and substitutions. A majority (79%, 27,672 of 35,083) of isoform pairs contain a single altered region (Fig. 2B). Still, 7411 alternative isoforms contain two or more discontinuous altered regions. The average number of altered regions per transcript increases with the number of exons and the length of gene, ORF, CDS, and protein for both reference and alternative isoform (Supplemental Fig. S4). Some isoform pairs exhibiting the highest number of altered regions include *DNAH14*, in which 14 regions are found for proteins of *DNAH14* (Reference: DNAH14-220, Alternative: DNAH14-211), potentially explained by its large number of exons (86 exons for DNAH14-220).

Among the 44,326 altered protein regions (Fig. 2C), the median number of affected AAs (lost or gained due to a protein insertion, deletion, or substitution) is 49 AA, with the first and third quartiles containing 21 AA and 128 AA, respectively. Among the altered regions, 14% (6189) are insertions, 47% (20,780) are deletions, and 39% (17,357) are substitutions (Fig. 2D). Full annotations for these altered regions at the protein and codon-level, representing the Biosurfer output for protein-blocks (p-blocks) and codon-blocks (c-blocks) can be found in Supplemental Tables S3 and S4.

The lengths of p-block insertions tend to be shorter than the length of deletions ($P < 2.2 \times 10^{-16}$, Mann–Whitney U test) or substitutions ($P < 2.2 \times 10^{-16}$, Mann–Whitney U test) (Fig. 2E–H). Since the deletion or insertion status of a protein region is dependent on which isoform is denoted as the reference, this trend may reflect the tendency that longer isoforms are more likely to be defined as the "reference" isoform, which may be biologically driven or influenced by genome annotation guidelines (Rodriguez et al. 2013; O'Leary et al. 2016; The UniProt Consortium 2017; Frankish et al. 2021; Varabyou et al. 2023a). We found a similar trend for p-block substitutions; the lengths for substituted regions tended to be longer for the sequence affected in the reference (Fig. 2G) versus the alternative (Fig. 2H) isoform; although, for 3263 (18% of cases), the affected regions in alternative isoforms can be longer (Fig. 2I).

Analysis of N-terminal protein variations

Variations in the N-termini are found in 28% (12,504 of 44,326) of all possible reference-alternative isoform pairs, corresponding to 5872 genes (Supplemental Table S5). We examined for the N-terminal variations the explanatory mechanism, including alternative TSSs, AS, and alternative translation initiation sites (TISs).

All cases of variable N-termini involve two initiation codons (AUGs), one upstream and one downstream, relative to the genome. A major category we first observed are those in which the N terminus is different due to start codons that are mutually exclusively present across the two transcripts (Fig. 3A). Specifically, the start codon present in the reference isoform is absent from the alternative isoform, and vice versa. These "mutually exclusive start codons" or MXS were observed for 3123 reference-alternative isoform pairs (Fig. 3A). MXS may arise either from an alternative TSS or from AS in the 5′ UTR. We found that nearly all (99%, 3097 of 3123) cases are caused by alternative TSS usage (Fig. 3A, hatched region of the bar), with only a small, but nonzero, fraction (1%, 26 of 3123) of MXS cases arise from splicing of the 5′ UTR, in which splicing regulation is influencing N-terminal usage. An example of TSS-driven MXS for *PRKACA* is shown in Figure 3B (pair, PRKACA-201 and PRKACA-202).

The second category is when the upstream start codon is transcribed in only one of the two isoforms, but the downstream start codon is present in both transcripts. We refer to this scenario as shared downstream starts (SDSs), of which there were 6878 cases (Fig. 3A). Like with cases of MXS, SDS arises primarily from alternative TSS usage (80% of cases) versus 5′-UTR AS (20% of cases).

MXS and SDS are common patterns underlying alterations of the N-termini or protein, driven by differential availability of initiator codons in the mature transcript. We asked if there may be differences in the length of such N-terminal alterations for SDS versus MXS events. Measuring the differential length of the affected N-terminal regions between reference-alternative isoform pairs, we found that, on average, SDS tends to affect a greater proportion

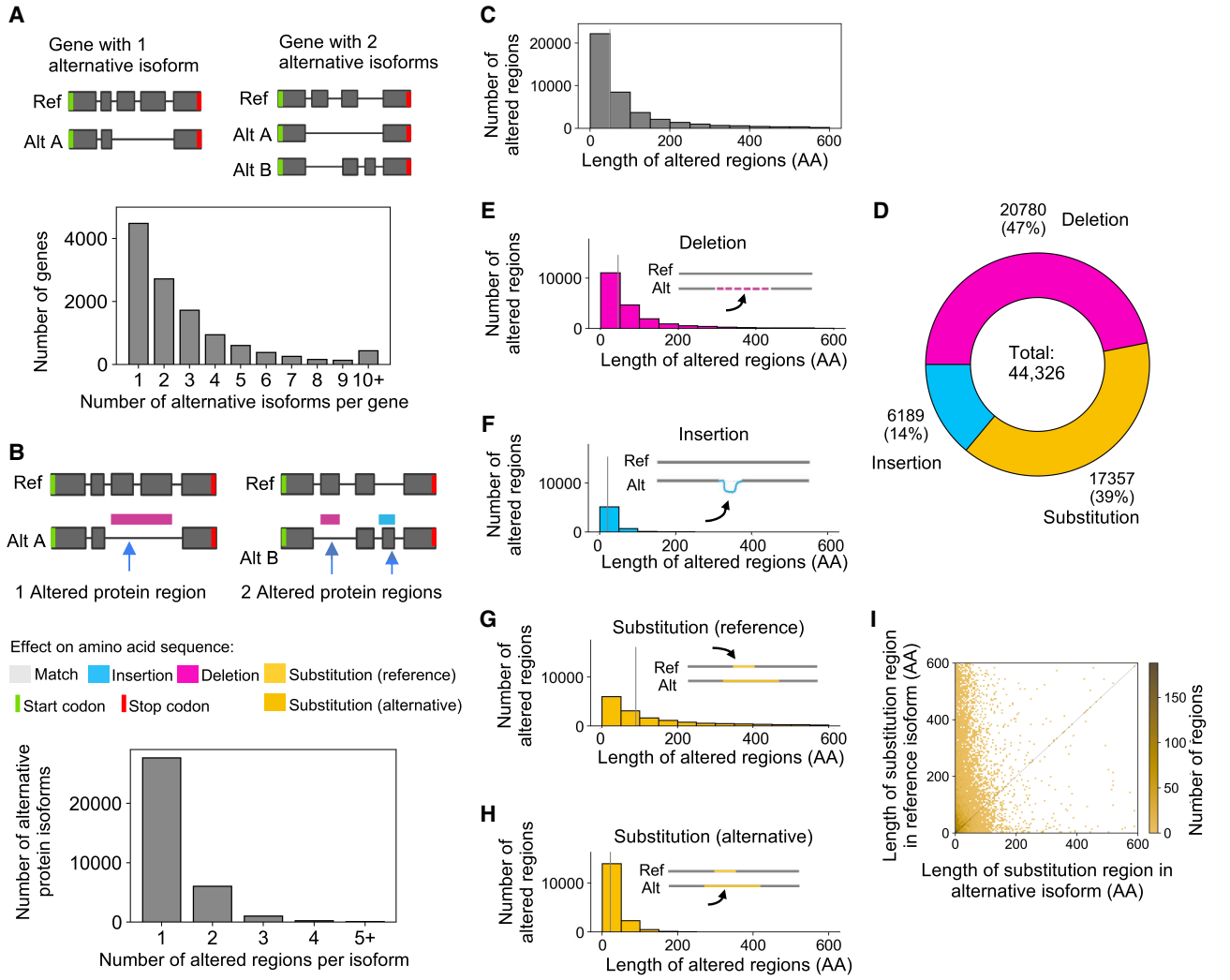


Figure 2. Characterization of altered protein regions (Biosurfer p-blocks) across the GENCODE-annotated human proteome. (*A*) Schematic of genes with one or two alternative protein isoforms and distribution of the number of alternative protein isoforms per gene. (*B*) Schematic of the altered protein regions (here, highlighted in pink and blue), displayed relative to the underlying transcript structures. The proteome-wide distribution of the number of affected protein regions observed per alternative isoform. (*C*) Distribution of the length of altered protein regions across the annotated proteome. Differences >600 AAs are not included (2347 cases, 5.3% of the data). These altered protein regions include cases of (1) deleted regions, (2) inserted regions, and (3) the region (in the reference isoform) in which one polypeptide subsegment is substituted for another. In other words, this distribution (*C*) represents an aggregation of the distributions shown in *E*–*G*. (*D*) Fraction of altered protein regions affected by insertions, deletions, and substitutions. (*E*–*H*) Distribution of the lengths of altered protein regions for (*E*) deletions, (*F*) insertions, (*G*) substituted region in the reference isoform, and (*H*) substituted region in the alternative isoform. (*I*) Comparison of the lengths of altered regions in the reference versus alternative isoforms for substitutions.

of protein length, as compared to MXS (Fig. 3C; Supplemental Fig. S5; $P = 2.8 \times 10^{-148}$, Mann–Whitney *U* test). The larger differences in length driven by SDS could be explained by cases in which transcription is initiated from internal sites of the gene, giving rise to an ORF that corresponds to a subsequence of the ORF in the other isoform, theoretically producing a truncated C-terminal-containing subsequence of the full-length protein.

Recently, a mechanism related to SDS was described in which internal exons (not the 5′ most exon, i.e., first exon of a transcript) in one transcript can be immediately downstream from a DNA element of novel promoter activity and thus serve as the first transcribed exon in other isoforms (Fiszbein et al. 2022). Such so-called hybrid exons thus can operate as both sites of transcription initiation and AS, in effect, swapping their roles depending on the regulatory context. Of the cases of SDS, we observed that ~22% correspond to these hybrid exon swaps (Fig. 3D; Supplemental Table S6). The functional consequences of hybrid exon usage are not well understood; however, one potential function could be the production of a protein with a truncated N terminus, which could remove signal peptide sequences or binding domains (Kelemen et al. 2013).

In addition to MXS and SDS, wherein start codon availability is controlled through differential transcription, we also observed many cases in which both upstream and downstream start codons co-occur in one or both transcripts of a pair. In these cases, the choice of start codon may be influenced by cotranslational regulation, e.g., ribosome initiates translation at alternative initiation sites (altTIS).

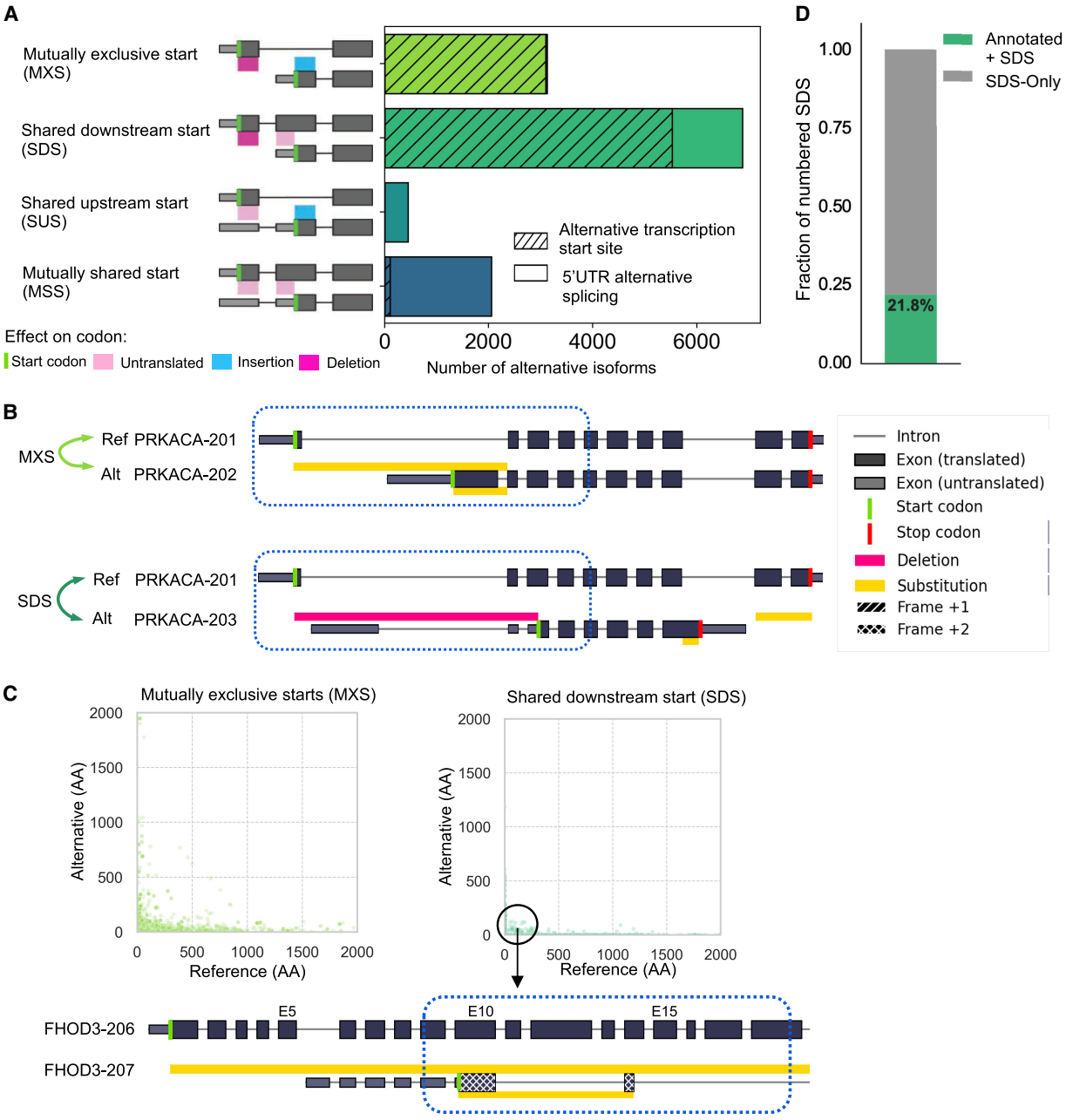


Figure 3. Analysis of mechanisms underlying variable N-terminal proteins across the GENCODE-annotated human proteome. (*A*) Distribution of the types of alternative N-terminal regions, classified based on the presence and translational status of the start codon. Hatches denote the fraction of alternative N-terminal regions associated with alternative TSSs, as opposed to 5′ UTR AS. (*B*) Biosurfer output of altered N-terminal regions for *PRKACA* gene that undergo MXS (light green arrows) and SDS (dark green arrows). In the example of MXS, the yellow bars above and below PRKACA-202 (alternate) transcript indicate the N-terminal ranges that differ between the reference and alternative isoform. The yellow bar above PRKACA-202 shows the N-terminal protein sequence that is specific to the reference (PRKACA-201) and the bar below the transcript indicates the N-terminal region specific to the alternative isoform. The Biosurfer bars span the intronic lines between exons, but intronic regions do not contribute to the protein sequence differences. In the example of SDS, the pink Biosurfer bar above the transcript of PRKACA-203 represents the range of transcript sequence that is translated in the reference, but not translated in the alternative isoform. (*C*) Scatterplot of the length of affected N-terminal variation in the reference versus alternative, faceted by mutually exclusive starts (MXS) or shared downstream start (SDS) status. An interesting case of an SDS leading to unique N-terminal sequence in the reference is caused by the usage of a different frame at the initiation of translation, with an example shown for isoforms of the gene *FHOD3*. (*D*) Fraction of SDSs caused by hybrid exon swaps.

For 2054 cases, we found altTIS, which we classified as instances of a mutually shared start (MSS) codon. We also found 449 cases in which the upstream start codon is present in both isoforms, but the downstream start codon is only present in one isoform.

In such cases of shared upstream start (SUS) (Fig. 3A), the explanatory cotranslational mechanism is not as clear, based on the ribosomal scanning model of translation initiation (Kozak 1978). The upstream start codon present in both isoforms would need

to be bypassed by the ribosome only in the alternative isoform. Therefore, some of the SUS annotations may need to be validated or may be erroneous ORF calls, as early ORF prediction workflows attribute higher scores to longer ORFs (Wang et al. 2013; Varabyou et al. 2023a), under-annotating ORFs that utilize the upstream (annotated) start codon but is much shorter than the reference due to a reading frame shift (Wang et al. 2013).

Analysis of internal protein variations

The internal regions of the protein isoforms account for 43% (19,263 of 44,326) of possible reference-alternative isoform pairs, corresponding to 6673 genes (Supplemental Table S7). A large majority of these regions (80%, 15,444 of 19,263) are caused by single simple splicing events: exon skipping, alternative acceptor, alternative donor, or an in-frame retained intron. As expected, exon skipping events are most numerous making up 9505 of cases. In terms of the general effect on protein sequence, most altered regions (70%, 13,453 of 19,263) lead to a deletion or removal of AA residues (Fig. 4A), and, again, in such cases, exon skipping is most common (51%, 6908 of 13,453).

Going beyond simple splicing events, we observed that 19% of variable internal regions (3701 of 19,263) were associated with multiple events, which we refer to as compound, or linked, splicing events. We found that 56% (2099 of 3701) of compound events involve multiexon skipping, the rest being combinations of alternative donor/acceptor sites with exon skipping/inclusion (Fig. 4B).

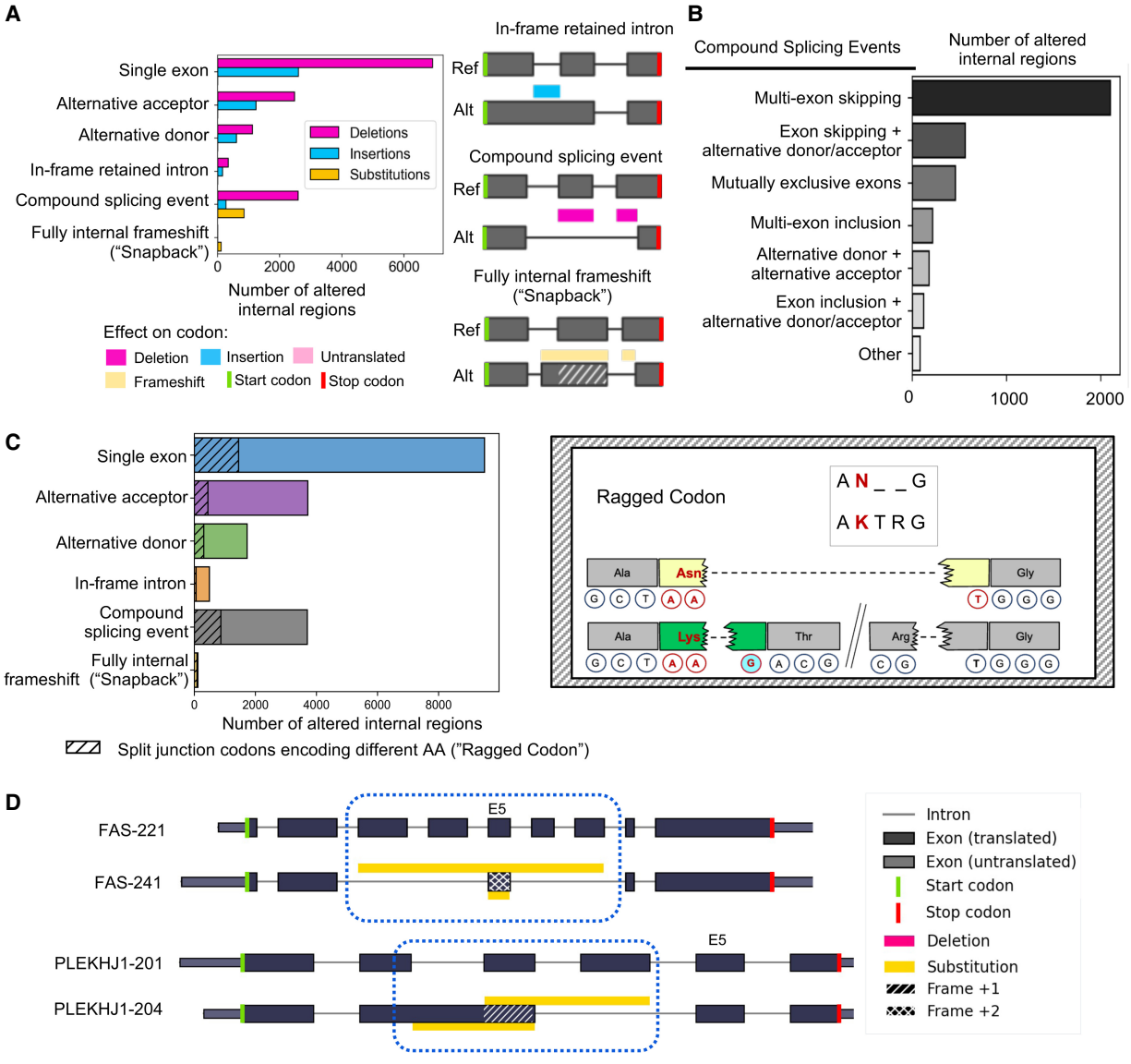


Figure 4. Analysis of internal protein-altered regions across proteins in GENCODE. (*A*) Frequencies of the categories of the splicing mechanism underlying internal protein sequence changes, split by their protein-coding impact (deletion, insertion, and substitution). All sequence regions involve a reference-alternative isoform pair. (*B*) Frequency of compound splicing events across the altered internal regions. (*C*) The proportion of each internal protein region type for which there exists a split codon pattern near its boundaries that would cause a single AA difference, or "ragged" codon. (*D*) Examples of successive frameshifting ("snapback" frameshift) that leads to an affected protein region that is wholly internal to the protein, for genes *FAS* and *PLEKHJ1*.

Complex nucleotide to AA relationships that affect internal protein sequences

Previous studies of the impact of splicing on proteins have typically focused on cases in which differential splicing of transcript regions directly corresponds to changes in protein sequence (Reixachs-Solé and Eyras 2022). However, in many instances, there is not a simple one-to-one relationship between nucleotides in a transcript and the corresponding AA identities in the encoded protein. Such complexity alters the protein sequence in nonintuitive ways. Using the detailed codon tracking afforded by Biosurfer, we systematically characterized the protein-level impact of variations not commonly described: codons that span junctions and unusual reading frame shifts.

To characterize differentially split codons, we examined all paired codons (see c-block section in Methods) that are split across junctions and determined the identity of the associated AAs (Supplemental Fig. S3). Across all internal altered protein regions, we found 17% (3213 of 19,263) of regions that are flanked by one or more split codon pairs encoding different AA residues (Fig. 4C; also see Table 1). These so-called ragged codons affect a single residue and are always adjacent to an altered protein region. Split codons are enriched by splice event type (Chi-square test: P-value = 5.54×10^{-99}), and we found that ragged codons are more frequently found in protein insertions compared to deletions or substitutions (Chi-square test: P-value = 4.11×10^{-106}).

Notably, we found an uncommonly characterized pattern of successive reading frame shifts that exclusively affects the internal residues of a protein. In these cases, the alternative isoform's reading frame is shifted due to one splicing event, but then shifts back into register of the reference frame due to a second, independent splicing event. We refer to these events as "snapback" frameshifts, as there is a return back to the original reading frame. Snapback frameshifts have been previously observed, such as in the gene *HSF4*, which produces internally frameshifted isoforms that have been demonstrated to exert different regulatory effects (Tanabe et al. 1999), but the snapback phenomenon generally speaking has not been systematically described. Within GENCODE, we found 118 examples of such snapback isoforms across 95 genes, including *FAS* and *PLEKHJ1* (Fig. 4D; Supplemental Table S8). What is potentially interesting about these cases is that the same underlying genomic sequence encodes different AA residues, and genetic mutations could lead to two different residue changes depending on the isoform, although the functional significance of such variations is unknown.

Analysis of C-terminal variations

C-terminal changes make up 28% (12,241 of 44,326) of reference-alternative pairs, corresponding to 6138 genes (Supplemental Table S9).

To break down the sources of C-termini variability, we found it useful to distinguish the most direct preceding cause of altered C-termini. In principle, all C-terminal changes must arise from an upstream splicing event that influences the termination codon used (notwithstanding posttranslational cleavage events). However, such changes could be further classified. The altered C terminus could arise from alternative terminal (i.e., last) exons, each harboring a different stop codon, so that the splicing event directly influences the stop codon availability (i.e., direct splice-driven events). In other instances, C-terminal changes could arise from a somewhat indirect relationship to the splice event, such as when a splicing event causes a translational frameshift, in effect, "revealing" in the other frame a new stop codon that is now decoded by the ribosome (i.e., frameshift-driven events) (Supplemental Table S9).

In direct splice-driven events, the stop codon availability is dictated by the actively transcribed regions that contain the stop codon. We find this scenario for 72% (8877 of 12,241) of all C-terminal variations (Fig. 5A). These variations can be further classified based on the pattern of splicing at the C terminus. The first pattern involves an exon extension into the intron region, introducing a premature stop codon. These exon extensions introduce termination, or "EXIT," make up 3498 (39.4% of 8877) cases (Fig. 5B). The second pattern involves usage of alternative terminal coding exons or "ATE," making up 3301 (37.1% of 8877) of cases (Fig. 5B). Overall, EXIT and ATE changes lead to a shorter C-termini in the alternative isoform (distribution shown in Fig. 5C).

EXIT versus ATE reflects how different "modes" of spliceosome regulation could lead to distinct C-terminal consequences. In EXIT, the reference-containing donor splice site fails to be spliced in the alternative isoform, leading to partial or full intron retention. An example of EXIT is shown in the right panel of Figure 5C for the pair, TDRD12-206 and TDRD102-202. In ATE, on the other hand, the spliceosome catalyzes splicing at one of two splice site acceptor sites, influencing terminal exon identity and thus stop codon used. A well-known pattern of ATE is poison exons, a mechanism by which the inclusion of an exon leads to a premature termination codon that either elicits nonsense-mediated decay or generates a truncated protein product (Carvill and Mefford 2020). Poison exons are evolutionarily conserved and likely play a role in downregulating gene expression (Lareau et al. 2007). We found 550 (6.2% of total 8877) cases of potential poison exons. Other ATE patterns include one in which the alternative last exon of the alternative isoform resides in the UTR region of the reference isoform, suggesting that such sequences in the 3′ UTR could dually code both transcript and protein functional elements. We found 819 (9% of 8877) cases of such alternative last exon in UTR (ALE in UTR) (Fig. 5B). A third ATE variation, referred to as a cut-out splice terminal exon (COSTE), the 5′ end of the last exon is shared, but the alternative isoform utilizes a splice site that skips over the remaining portion of the last exon in the reference, thereby creating a different last exon not found in the original reference. We identified 273 cases of this pattern (3% of 8877) (Fig. 5B).

Frameshift-driven events influence stop codon usage somewhat indirectly through shifts in the translational reading frame. In such cases, a splice-induced reading frame shift causes all downstream codons to be read in a different frame and stop codons are "revealed," or decoded, by the ribosome. We found 3364 (28% of 12,241) cases of frameshift-induced C-terminal changes (Fig. 5A, examples are shown in Fig. 5D, pairs, GIPC3-201 and GIPC3-202, along with FANCM-201 and FANCM-231). Generally speaking, frameshifts lead to a substantial shortening of the C-terminal region in the alternative isoform; however, we found 549 cases (16% of 3364) in which the C-terminal region is longer in the alternative versus the reference isoform (Fig. 5D).

We also observed across all frameshift-driven events a depletion of isoform pairs in which a large portion of the reference isoform (e.g., 2000 AA or longer) is truncated due to a frameshift in the alternative isoform (Fig. 5D, left-hand scatterplot), a trend not observed in an experimentally predicted proteome (see section below and Supplemental Fig. S15B). This global depletion likely represents gene annotation decisions, as dramatically truncating frameshift events that contain untranslated exon junctions downstream from the stop codon (in the 3′ UTR) would lead to predicted

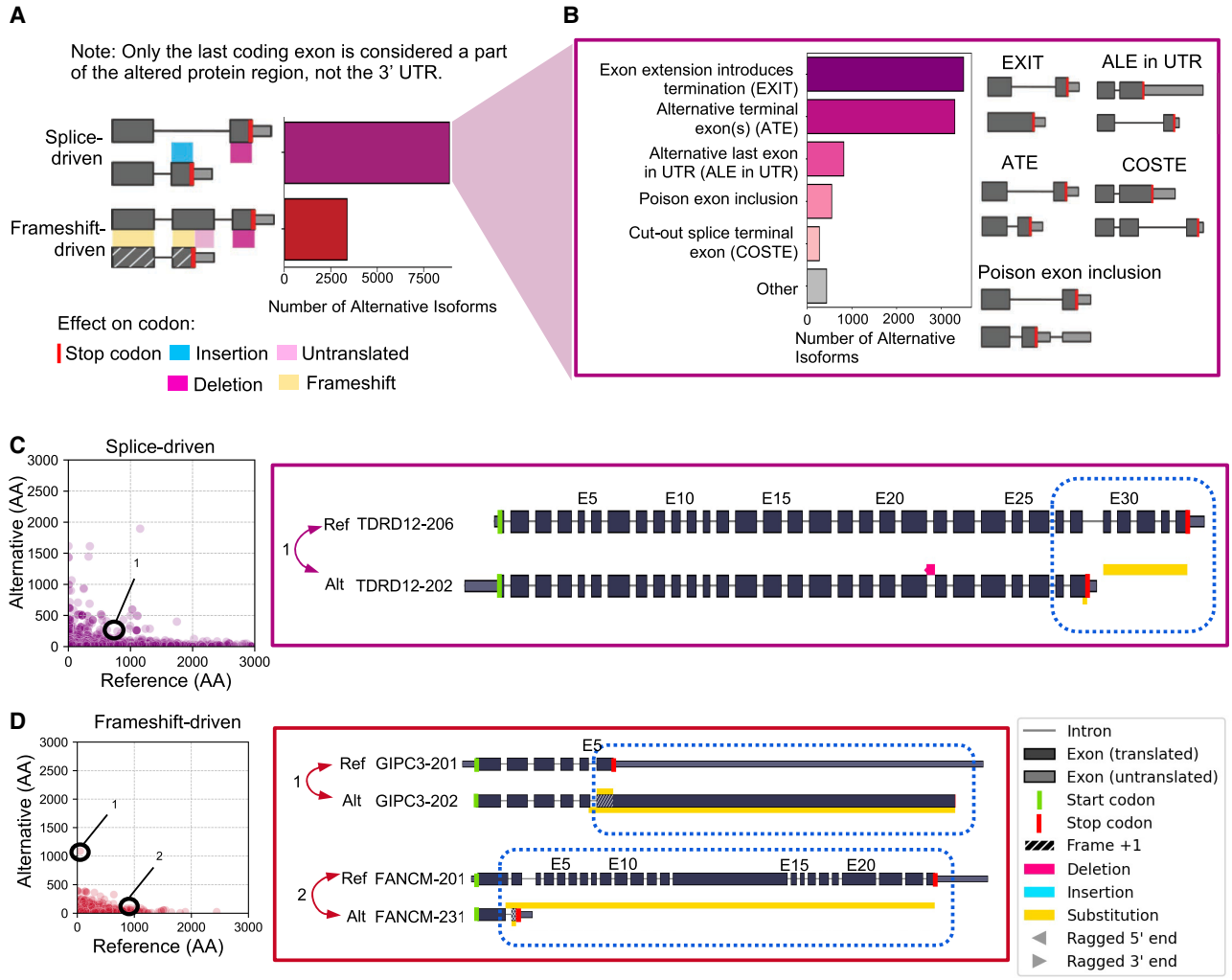


Figure 5. Analysis of alternative C-terminal protein sequences in GENCODE v42. (*A*) Frequency of alternative C-terminal categories based on splicing or frameshifts being the primary driving factor. (*B*) Distribution of the frequencies for various splice-driven patterns. (C) Scatterplot of the length of splice-driven C-terminal variation in the reference versus alternative. An example of this category is observed in the *TDRD12* gene. *TDRD12* undergoes splice-driven alteration causing an alternative terminal exon in TDRD12-202 to harbor the stop codon (*D*) Scatterplot of the length of frameshift-driven C-terminal variation in the reference versus alternative. Biosurfer plot examples of frameshift-driven category illustrated in *GIPC43* and *FANCM* genes.

NMD, and these transcripts would be filtered out or reassigned an NMD biotype (Harrow et al. 2012). Supplemental Figure S6 shows differential lengths between reference and alternative isoforms for the various C-terminal variation categories described in this section.

Functional enrichments based on variation type

We further examined genes that exhibited N-terminal, internal, and C-terminal variation to determine if such genes might exhibit certain biological properties. We performed GO enrichment analysis using Enrichr for N-terminal, internal, and C-terminal variable genes, results in Supplemental Table S10 (Kuleshov et al. 2016). We found that C-terminal variable genes may be enriched in apoptotic processes. Internal variable genes seem to be enriched for phosphorylation and may reflect the fact that AS occurs in intrinsically disordered regions that more likely harbor phosphorylation sites (Buljan et al. 2012). N-terminal variable genes may be enriched in transcriptional regulation. We found a negative correlation between variable N-term genes and the presence of a signal peptide (Supplemental Table S11, chi-squared test, *P*-value $< 2.2 \times 10^{-16}$) (The UniProt Consortium 2017).

Cases of nonoverlapping or entirely distinct ORFs

For the remaining reference-alternative pairs, 0.7% (318 of 44,326), that were not classified as N-terminal, internal, or C-terminal variations, we found such cases represent ORFs that do not overlap in genome space (example in Supplemental Fig. S7A) or are completely unrelated in polypeptide sequence due to a combination of nonoverlapping and at least one frameshifted region (example in Supplemental Fig. S7B). In such cases of unrelated ORFs, the entire isoform is considered to be a p-block substitution and output in Biosurfer tables.

Characterization of altered protein regions in the mouse genome

To demonstrate the wide applicability of Biosurfer and to investigate whether those patterns found in the human genome

described above are conserved in other species, we applied the same analysis pipeline to a mouse genome (GENCODE M35) (Harrow et al. 2012). We analyzed 16,398 reference-alternative protein isoform pairs across 21,639 genes. We found 20,425 altered protein regions with an average of 1.2 altered regions per isoform. In total, 13,198 (80% of 16,398) isoform pairs contain a single altered region, while the rest pairs have two or more discontinuous altered regions. The characteristics and distribution of variation in N-terminal, internal, and C-terminal regions are very similar between mouse and human genome and are illustrated in Supplemental Figures S8–S11, and data may be found in Supplemental Tables S12–S15.

Characterization of altered protein regions across a long-read-predicted proteome

The process of defining the reference proteome heavily draws from sources of experimental evidence such as deeply sequenced cell and tissue types, and the protein isoform sequences represent an aggregate model of the human proteome. Therefore, to characterize potential isoform-related protein variability in a specific biological condition, we employed a "long-read proteogenomics" pipeline (Miller et al. 2022), generating a proteome predicted from long-read RNA-seq transcript sequences collected from a karyotypically normal human stem cell line (WTC-11). Using Biosurfer, we characterized the landscape of protein isoforms in the WTC-11 proteome and found 44,916 protein isoform pairs across 10,117 genes (Supplemental Table S16, p-block and c-block outputs in Supplemental Tables S18, S19). We found that 19,570 of the 44,916 alternative isoforms were known and 23,884 were novel, i.e., without a match to GENCODE using the SQANTI Protein comparison tool (Supplemental Table S17). Assigning the GENCODE APPRIS as the "reference," we defined 53,915 altered protein regions (Supplemental Fig. S12). Compared to the GENCODE analysis, similar trends were observed for N-terminal (Supplemental Fig. S13) and internal region variation (Supplemental Fig. S14, snapback isoforms in Supplemental Table S20). Besides these similar trends, the experimental proteome returned a higher number of C-terminal variations that involved intron retention events (EXIT event type, see Fig. 5B) as compared to GENCODE isoforms, matching earlier findings from EST and cDNA data (Supplemental Fig. S15; Modrek et al. 2001; Nakao et al. 2005).

Discussion

To study the functional impact of alternatively spliced protein isoforms, it is critical to track precise differences in protein isoform sequences and link such variations to the upstream explanatory mechanisms. However, it is challenging to systematically characterize the full interplay between genomic and proteomic variations, which hinders discoveries of novel biological variations represented in long-read RNA-seq data. We developed Biosurfer, a computational approach, available as a Python package, that systematically extracts protein isoform sequence variations while maintaining the explicit links to their underlying transcriptional and posttranscriptional mechanisms.

To demonstrate the utility of Biosurfer, we characterized protein isoform differences across an annotated (GENCODE) and long-read RNA-seq predicted proteome. Using Biosurfer's interlinked transcript, codon, and protein data structures, we determined the upstream mechanisms explaining isoform alterations, uncovering notable complexity. First, we confirmed past observations of alternative transcription underlying most N-terminal variations (Reyes and Huber 2018), and most internal protein sequence differences arising from single splicing events, such as exon skipping (Wang et al. 2008). However, our study goes beyond these known trends. Biosurfer can track in detail the patterns of codons flanking altered protein regions, and we systematically enumerated such split codon patterns that change the encoded AA residue identity and thus contribute to the variation of AA residues. We also used Biosurfer to systematically annotate an unusual frameshift pattern, which involves successive reading frame shifts that lead to a change in protein regions that is entirely internal to the protein, referred to here as "snapback" frameshifts. And last, C-terminal differences are primarily splice-driven or frameshift-driven, and highly truncated alternative isoforms from frameshifts are underrepresented in GENCODE annotations but not in an experimentally proteome predicted from long-read RNA-seq data.

Related to the comparison of proteins, we had to designate a "reference" isoform, although the biological role of most isoforms is unknown (Yang et al. 2016; Reixachs-Solé and Eyras 2022), and thus representative isoforms are chosen depending on the assumptions and goals of the research community (The UniProt Consortium 2017; Pozo et al. 2022).

Biosurfer analyses rely on user-defined protein isoforms. Only canonical start and stop codons are assumed, unless noncanonical sites are annotated in a reference proteome (Mudge et al. 2022) or the user. Determination of the biologically relevant ORF remains an ongoing challenge. Many ORF callers like transdecoder, CPAT, GMST, and others predict ORFs, relying on heuristics and common features of translation, which may not be the rule in every case. Currently, the prediction of proteins from deep coverage long-read RNA-seq data rely on heuristics, such as prioritizing ORFs from an alternative isoform that share the same start AUG codon with the reference, or selection of the most 5′ proximal AUG (Tang et al. 2020; Miller et al. 2022), whereas others have developed computationally efficient scoring strategies that rank more highly the ORFs with most protein similarity to the reference (Varabyou et al. 2023a,b). To provide more reliable ORF annotations, experimental approaches like Ribo-seq demarcate novel coding regions, including sites of noncanonical translation, which might be information that could be incorporated in proteogenomic workflows (Mudge et al. 2022; Leblanc et al. 2024). Furthermore, newly reported ORF annotations could also reveal cases of short ORFs, such as upstream ORFs (uORFs), that are specific to one but not another transcript and might explain certain patterns we observed such as the SUS pattern. Note that Biosurfer currently assumes a single ORF per transcript and does not handle multiple ORFs within the same transcript (i.e., bicistronic transcripts). Despite these limitations, Biosurfer could still be used as a general purpose ORF comparison tool, and one could compare one ORF of a transcript, including uORFs, to an ORF of another transcript. Overall, more options for ORF comparisons should be developed as ORF annotations mature.

Our first version of Biosurfer proposes a new framework for a detailed comparison of protein isoforms, a first step toward inferring their functional consequences in health and disease.

Software availability

The Biosurfer code is publicly available at GitHub (https://github.com/sheynkman-lab/biosurfer). The scripts necessary to reproduce the results presented in this manuscript can be accessed in a separate GitHub repository (https://github.com/sheynkman-

lab/biosurfer_analysis). All necessary code files, including the Biosurfer source code and analysis scripts, are included as Supplemental Code files to ensure accessibility. All necessary input, intermediate, and final output files from the Biosurfer analysis are available at Zenodo (https://doi.org/10.5281/zenodo.13243233) and as Supplemental Data.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by the National Library of Medicine (R01-LM014017) to G.M.S. and D.K.

Author contributions: G.M.S. conceived the project, performed analysis, and wrote and revised the manuscript. D.K. supervised the project and wrote and edited the manuscript. M.M. wrote the Biosurfer code base, conducted analysis, and wrote the manuscript. J.S. wrote the Biosurfer code base and conducted the analysis. B.J. conducted the PacBio analysis. Z.P.W. and A.F. performed the analysis on hybrid exons. E.F.W. conducted analysis. D.R.C. and P.J.C. supervised and wrote the manuscript. Z.G. participated in visualization, analysis, and data presentation of the work. S.L. conducted the analysis, tested the Biosurfer program, and wrote the manuscript.

References

Abood A, Mesner LD, Jeffery ED, Murali M, Lehe MD, Saquing J, Farber CR, Sheynkman GM. 2024. Long-read proteogenomics to connect disease-associated sQTLs to the protein isoform effectors of disease. *Am J Hum Genet* **111:** 1914–1931. doi:10.1016/j.ajhg.2024.07.003

Annaldasula S, Gajos M, Mayer A. 2021. IsoTV: processing and visualizing functional features of translated transcript isoforms. *Bioinformatics* **37:** 3070–3072. doi:10.1093/bioinformatics/btab103

Anvar SY, Allard G, Tseng E, Sheynkman GM, de Klerk E, Vermaat M, Yin RH, Johansson HE, Ariyurek Y, den Dunnen JT, et al. 2018. Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol* **19:** 46. doi:10.1186/s13059-018-1418-0

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

Carvill GL, Mefford HC. 2020. Poison exons in neurodevelopment and disease. *Curr Opin Genet Dev* **65:** 98–102. doi:10.1016/j.gde.2020.05.030

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. 2003. Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res* **31:** 3497–3500. doi:10.1093/nar/gkg500

Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4:** 265–270. doi:10.1038/nnano.2009.12

Cooper TA, Wan L, Dreyfuss G. 2009. RNA and disease. *Cell* **136:** 777–793. doi:10.1016/j.cell.2009.02.011

de la Fuente L, Arzalluz-Luque Á, Tardáguila M, Del Risco H, Martí C, Tarazona S, Salguero P, Scott R, Lerma A, Alastrue-Agudo A, et al. 2020. tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biol* **21:** 119. doi:10.1186/s13059-020-02028-w

De Paoli-Iseppi R, Gleeson J, Clark MB. 2021. Isoform age - splice isoform profiling using long-read technologies. *Front Mol Biosci* **8:** 711733. doi:10.3389/fmolb.2021.711733

de Souza VBC, Jordan BT, Tseng E, Nelson EA, Hirschi KK, Sheynkman G, Robinson MD. 2023. Transformation of alignment files improves performance of variant callers for long-read RNA sequencing data. *Genome Biol* **24:** 91. doi:10.1186/s13059-023-02923-y

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323:** 133–138. doi:10.1126/science.1162986

Evans T, Loose M. 2015. Alignwise: a tool for identifying protein-coding sequence and correcting frame-shifts. *BMC Bioinformatics* **16:** 376. doi:10.1186/s12859-015-0813-8

Fiszbein A, McGurk M, Calvo-Roitberg E, Kim G, Burge CB, Pai AA. 2022. Widespread occurrence of hybrid internal-terminal exons in human transcriptomes. *Sci Adv* **8:** eabk1752. doi:10.1126/sciadv.abk1752

Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, et al. 2021. GENCODE 2021. *Nucleic Acids Res* **49:** D916–D923. doi:10.1093/nar/gkaa1087

Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, Sisu C, Wright JC, Arnan C, Barnes I, et al. 2023. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* **51:** D942–D949. doi:10.1093/nar/gkac1071

Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, Brown KL, Garimella K, et al. 2022. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608:** 353–359. doi:10.1038/s41586-022-05035-y

Gohr A, Irimia M. 2019. *Matt*: Unix tools for alternative splicing analysis. *Bioinformatics* **35:** 130–132. doi:10.1093/bioinformatics/bty606

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8:** 1494–1512. doi:10.1038/nprot.2013.084

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22:** 1760–1774. doi:10.1101/gr.135350.111

Jammali S, Djossou A, Ouédraogo W-YDD, Nevers Y, Chegrane I, Ouangraoua A. 2022. From pairwise to multiple spliced alignment. *Bioinform Adv* **2:** vbab044. doi:10.1093/bioadv/vbab044

Joglekar A, Hu W, Zhang B, Narykov O, Diekhans M, Balacco J, Ndhlovu LC, Milner TA, Fedrigo O, Jarvis ED, et al. 2023. Single-cell long-read mRNA isoform regulation is pervasive across mammalian brain regions, cell types, and development. bioRxiv doi:10.1101/2023.04.02.535281

Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S. 2013. Function of alternative splicing. *Gene* **514:** 1–30. doi:10.1016/j.gene.2012.07.083

Kozak M. 1978. How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell* **15:** 1109–1123. doi:10.1016/0092-8674(78)90039-9

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44(**W1**):** W90–W97. doi:10.1093/nar/gkw377

Lareau LF, Inada M, Green RE, Wengrod JC, Brenner SE. 2007. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446:** 926–929. doi:10.1038/nature05676

Leblanc S, Yala F, Provencher N, Lucier J-F, Levesque M, Lapointe X, Jacques J-F, Fournier I, Salzet M, Ouangraoua A, et al. 2024. Openprot 2.0 builds a path to the functional characterization of alternative proteins. *Nucleic Acids Res* **52:** D522–D528. doi:10.1093/nar/gkad1050

Louadi Z, Yuan K, Gress A, Tsoy O, Kalinina OV, Baumbach J, Kacprowski T, List M. 2021. DIGGER: exploring the functional role of alternative splicing in protein interactions. *Nucleic Acids Res* **49:** D309–D318. doi:10.1093/nar/gkaa768

Martelli PL, D'Antonio M, Bonizzoni P, Castrignanò T, D'Erchia AM, D'Onorio De Meo P, Fariselli P, Finelli M, Licciulli F, Mangiulli M, et al. 2011. ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. *Nucleic Acids Res* **39:** D80–D85. doi:10.1093/nar/gkq1073

Mehlferber MM, Jeffery ED, Saquing J, Jordan BT, Sheynkman L, Murali M, Genet G, Acharya BR, Hirschi KK, Sheynkman GM. 2022. Characterization of protein isoform diversity in human umbilical vein endothelial cells via long-read proteogenomics. *RNA Biol* **19:** 1228–1243. doi:10.1080/15476286.2022.2141938

Miller RM, Jordan BT, Mehlferber MM, Jeffery ED, Chatzipantsiou C, Kaur S, Millikin RJ, Dai Y, Tiberi S, Castaldi PJ, et al. 2022. Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biol* **23:** 69. doi:10.1186/s13059-022-02624-y

Modrek B, Resch A, Grasso C, Lee C. 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res* **29:** 2850–2859. doi:10.1093/nar/29.13.2850

Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, Gonzalez JM, Magrane M, Martinez TF, Schulz JF, et al. 2022. Standardized annotation of translated open reading frames. *Nat Biotechnol* **40:** 994–999. doi:10.1038/s41587-022-01369-0

Nakao M, Barrero RA, Mukai Y, Motono C, Suwa M, Nakai K. 2005. Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals. *Nucleic Acids Res* **33:** 2355–2363. doi:10.1093/nar/gki520

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44:** D733–D745. doi:10.1093/nar/gkv1189

Pardo-Palacios F, Reese F, Carbonell-Sala S, Diekhans M, Liang C, Wang D, Williams B, Adams M, Behera A, Lagarde J, et al. 2024. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods* **21:** 1349–1363. doi:10.1038/s41592-024-02298-3

Pozo F, Rodriguez JM, Martínez Gómez L, Vázquez J, Tress ML. 2022. APPRIS principal isoforms and MANE Select transcripts define reference splice variants. *Bioinformatics* **38:** ii89–ii94. doi:10.1093/bioinformatics/btac473

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: multiple alignment of coding SEquences accounting for frameshifts and stop codons. *PLoS One* **6:** e22594. doi:10.1371/journal.pone.0022594

Reese F, Williams B, Balderrama-Gutierrez G, Wyman D, Çelik MH, Rebboah E, Rezaie N, Trout D, Razavi-Mohseni M, Jiang Y, et al. 2023. The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. bioRxiv doi:10.1101/2023.05.15.540865

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

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

Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J-J, Lopez G, Valencia A, Tress ML. 2013. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res* **41:** D110–D117. doi:10.1093/nar/gks1058

Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31:** 1009–1014. doi:10.1038/nbt.2705

Tanabe M, Sasai N, Nagata K, Liu XD, Liu PC, Thiele DJ, Nakai A. 1999. The mammalian *HSF4* gene generates both an activator and a repressor of heat shock genes by alternative splicing. *J Biol Chem* **274:** 27845–27856. doi:10.1074/jbc.274.39.27845

Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN. 2020. Full-length transcript characterization of *SF3B1* mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* **11:** 1438. doi:10.1038/s41467-020-15171-6

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

Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, et al. 2018. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* **28:** 396–411. doi:10.1101/gr.222976.117

Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, Kariyawasam H, Du MRM, Schuster J, Wang C, et al. 2021. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol* **22:** 310. doi:10.1186/s13059-021-02525-6

Tranchevent L-C, Aubé F, Dulaurier L, Benoit-Pilven C, Rey A, Poret A, Chautard E, Mortada H, Desmet F-O, Chakrama FZ, et al. 2017. Identification of protein features encoded by alternative exons using exon ontology. *Genome Res* **27:** 1087–1097. doi:10.1101/gr.212696.116

The UniProt Consortium. 2017. Uniprot: the universal protein knowledgebase. *Nucleic Acids Res* **45:** D158–D169. doi:10.1093/nar/gkw1099

Varabyou A, Erdogdu B, Salzberg SL, Pertea M. 2023a. Investigating open reading frames in known and novel transcripts using ORFanage. *Nat Comput Sci* **3:** 700–708. doi:10.1038/s43588-023-00496-1

Varabyou A, Sommer MJ, Erdogdu B, Shinder I, Minkin I, Chao K-H, Park S, Heinz J, Pockrandt C, Shumate A, et al. 2023b. CHESS 3: an improved, comprehensive catalog of human genes and transcripts based on large-scale expression data, phylogenetic analysis, and protein structure. *Genome Biol* **24:** 249. doi:10.1186/s13059-023-03088-4

Veiga DFT, Nesta A, Zhao Y, Deslattes Mays A, Huynh R, Rossi R, Wu T-C, Palucka K, Anczukow O, Beck CR, et al. 2022. A comprehensive long-read isoform analysis platform and sequencing resource for breast cancer. *Sci Adv* **8:** eabg6711. doi:10.1126/sciadv.abg6711

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476. doi:10.1038/nature07509

Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. 2013. CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41:** e74. doi:10.1093/nar/gkt006

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al. 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* **16:** 1297–1305. doi:10.1038/s41592-019-0617-2

Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, Sun S, Yang F, Shen YA, Murray RR, et al. 2016. Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* **164:** 805–817. doi:10.1016/j.cell.2016.01.029

Received March 15, 2024; accepted in revised form January 6, 2025.

Biosurfer for systematic tracking of regulatory mechanisms leading to protein isoform diversity

Mayank Murali, Jamie Saquing, Senbao Lu, et al.

Genome Res. 2025 35: 1012-1024 originally published online March 14, 2025
Access the most recent version at doi:10.1101/gr.279317.124