Perplexity as a Metric for Isoform Diversity in the Human Transcriptome

Megan D. Schertzer[1,2*†], Stella H. Park[1*], Jiayu Su[3], Gloria M. Sheynkman[4,5,6†], David A. Knowles[1,2,3,7†]

[1] New York Genome Center, New York, NY
[2] Department of Computer Science, Columbia University, New York, NY
[3] Department of Systems Biology, Columbia University, New York, NY
[4] Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA
[5] Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA
[6] UVA Comprehensive Cancer Center, University of Virginia, Charlottesville, VA
[7] Data Science Institute, Columbia University, New York, NY
* Equal contribution.
† Corresponding authors.

Abstract

Long-read sequencing (LRS) has revealed a far greater diversity of RNA isoforms than earlier technologies, increasing the critical need to determine which, and how many, isoforms per gene are biologically meaningful. To define the space of relevant isoforms from LRS, many existing analysis pipelines rely on arbitrary expression cutoffs, but a single threshold cannot accommodate the broad variability in isoform complexity across genes, cell-types, and disease states captured by LRS. To address this, we propose using *perplexity*–an interpretable measure derived from entropy–that quantifies the effective number of isoforms per gene based on the full, unfiltered isoform ratio distribution. Calculating perplexity for 124 ENCODE4 PacBio LRS datasets spanning 55 human cell types, we show that it provides intuitive assessments of isoform diversity and captures uncertainty across genes with varying complexity. Perplexity can be calculated at multiple gene regulatory levels–from transcript to protein–to compare how isoform diversity is reduced across stages of gene expression. On average, genes have an ORF-level perplexity of 2.1, indicating production of two distinct protein isoforms. We extended this analysis to evaluate expression variation across tissues and identified 4,593 ORFs across 3,102 genes with moderate to extreme tissue-specificity. We propose perplexity as a consistent, quantitative metric for interpreting isoform diversity across genes, cell types, and disease states. All results are compiled into a community resource to enable cross-study comparisons of novel isoforms.

Introduction

It has been known for decades that the majority of human genes can produce multiple isoforms (Pan et al. 2008; E. T. Wang et al. 2008). Early catalogs were built using cDNA cloning and Sanger sequencing, largely through expressed sequence tags (ESTs) (Gerhard et al. 2004; Harrow et al. 2006). Short-read RNA-sequencing later enabled high-throughput discovery of alternative splicing events across cell types and conditions (Harrow et al. 2012; Frankish et al. 2019; ENCODE Project Consortium 2012). However, short reads typically span only a single exon or splice junction, enabling event-level analysis (Y. I. Li et al. 2017; Shen et al. 2014), but making it difficult to reconstruct full-length transcripts. As a result, isoform-level quantification has relied on probabilistic read assignment to reference transcripts (Trapnell et al. 2010; Bray et al. 2016; Patro et al. 2017)–an error-prone process when isoforms are highly similar, gene structures are complex, or reads originate from unannotated isoforms (Sarantopoulou et al. 2021; Zhang et al. 2017; Steijger et al. 2013). Consequently, low-abundance isoforms in short-read datasets represent low-confidence structures and are routinely filtered out using strict expression cutoffs.

Long-read RNA-sequencing (LRS) overcomes these limitations by capturing full-length RNA molecules, providing direct evidence of isoform structures and enabling detection of combinations of splicing and transcriptional events (Sharon et al. 2013; Garalde et al. 2018; Workman et al. 2019; Al'Khafaji et al. 2023). Thus, LRS is fundamentally reshaping our understanding of the human transcriptome–not only by revealing vast numbers of novel isoforms (Monzó, Liu, and Conesa 2025; Pardo-Palacios, Wang, et al. 2024; Reese et al. 2023; Glinos et al. 2022; Leung et al. 2021; Humphrey et al. 2025), but also by challenging assumptions about the interpretation of low-abundance isoforms. LRS shifts the primary uncertainty from whether these isoforms are "real" to questions of their biological relevance.

Accordingly, filtering choices during analysis–such as transcript per million (TPM) thresholds, minimum isoform usage cutoffs, and sample recurrence requirements–can significantly impact estimates of isoform diversity. Yet, there is currently no consensus on how to handle low-abundance isoforms in LRS data. Reported estimates of transcriptome complexity can therefore vary by orders of magnitude even within the same cell type. This inconsistency complicates biological interpretation, undermines cross-study comparisons, and limits the broader utility of isoform-level analyses in gene annotation, regulation, and function.

To address this, we propose a systematic framework for quantifying isoform diversity using *perplexity*–an interpretable, entropy-based metric that reflects the effective number of isoforms per gene based on the full distribution of isoform expression. Unlike methods that rely on fixed cutoffs, perplexity considers all detected isoforms and scales naturally with isoform complexity. We apply this framework to 124 ENCODE4 PacBio LRS datasets across 55 human cell types (Reese et al. 2023), capturing isoform diversity at four levels: (1) gene potential–the number of detected isoforms, (2) gene perplexity, (3) protein-coding transcript perplexity, and (4) ORF (protein) perplexity. Although gene potential is highly variable–from 1 to 316 isoforms per gene–the average ORF perplexity is 2.1, indicating that a typical human gene produces two

distinct proteins. We further evaluate ORF perplexity across individual samples to assess variation across tissues.

All results are compiled into a community resource that includes reproducible isoform IDs based on hashed exon coordinates, which can be easily updated as new LRS data become available. Together, this study provides an unbiased and generalizable framework for quantifying isoform diversity across complex genes, cell types, and disease contexts.

Results

Perplexity Captures RNA Isoform Diversity of Protein-coding Genes in ENCODE4 LRS Data

Long-read RNA-sequencing (LRS) captures full-length RNA molecules and provides the most comprehensive view of isoform diversity to date. We analyzed 124 ENCODE4 PacBio LRS datasets across 55 cell and tissue types (Reese et al. 2023) using a custom pipeline (Supplemental Figure 1A). Isoforms were defined based on shared splice junctions–logic adapted from *FLAIR collapse* (Tang et al. 2020)–irrespective of precise read start or end site variation, resulting in 199,406 distinct RNA isoforms across 19,226 genes. We categorized 126,880 isoforms (63%) as protein-coding (L. Wang et al. 2013), 26,304 (13%) as nonsense-mediated decay (NMD), 38,831 (19%) as retained intron (RI), and 7,391 (4%) without an open reading frame >30 amino acids (noORF). Notably, 73% of protein-coding isoforms are novel based on GENCODE v46 basic annotations, consistent with the high number of novel isoforms identified in prior LRS analyses (Monzó, Liu, and Conesa 2025; Pardo-Palacios, Wang, et al. 2024; Reese et al. 2023; Glinos et al. 2022). To focus on protein isoform diversity, we filtered for genes with at least two protein-coding isoforms, excluding 6,568 genes (3,882 with 0 and 2,686 with 1 detected isoforms) (Supplemental Figure 1B-C), and retaining 185,160 isoforms across 12,658 genes (Supplemental Figures 1D-E, Supplemental Table S1).

To define an upper bound of a gene's isoform diversity, we calculated a metric we term *gene potential*, the total number of detected isoforms across LRS samples. In the ENCODE4 LRS data, gene potential averaged 14.6 isoforms (median = 9), including an average of 9.8 protein-coding isoforms (Supplemental Figure 1D). Many isoforms are detected at low abundance (Supplemental Figure 1E), and 46.5% of novel protein-coding isoforms contain novel splice sites–defined as NNC with SQANTI3 (Supplemental Figure 1F) (Pardo-Palacios, Arzalluz-Luque, et al. 2024).

Gene potential varies widely, with some genes expressing hundreds of isoforms; for example, we detect 316 isoforms for *GNAS*, 255 for *COL1A2*, and 249 for *DES*. While few would claim that all of these isoforms are functional, there is no consensus on which or how many isoforms per gene are biologically relevant. Most pipelines rely on user-defined expression cutoffs to filter low-abundance isoforms, but these thresholds are arbitrary and introduce interpretation biases that vary across studies.

To systematically quantify isoform diversity without relying on fixed expression cutoffs, we propose a framework that applies Shannon entropy and its derived metric, perplexity, to isoform expression distributions across protein-coding genes. Perplexity provides a single, interpretable value that captures both the number and balance of detected isoforms for a gene, reflecting the *effective number of isoforms*. While entropy (H) has been applied to transcriptomics data (Martínez and Reyes-Valdés 2008; MacArthur and Lemischka 2013; Gandrillon et al. 2021; Sterne-Weiler et al. 2018; Trapnell et al. 2010; Ritchie et al. 2008; Jones et al. 2024), its direct transformation to perplexity (2^H) offers a more intuitive readout in units of isoforms.

For example, a gene expressing two isoforms equally has an entropy of 1 bit and a perplexity of 2 effective isoforms (Gene A), while a gene expressing eight equally has 3 bits and a perplexity of 8 (Gene C; Figure 1A). In LRS datasets, most genes fall between these idealized cases, expressing isoforms at unequal proportions. Gene B, for instance, expresses 7 isoforms with ratios: [0.4, 0.35, 0.1, 0.07, 0.04, 0.03, 0.01], resulting in an entropy of 2.06 bits and a perplexity of 4.18, illustrating how lowly expressed isoforms–rather than being filtered out–still contribute to the perplexity-based metric of isoform diversity. Even for genes with exceptionally high gene potential, such as *GNAS* with 316 detected isoforms, perplexity (15.9 in this case) reflects contributions from the full distribution of isoform ratios, most of which are low-abundance.

While the average gene potential is 14.6, the average gene perplexity is 3.4. Direct comparison of gene potential and gene perplexity reveals wide variation across genes (Figure 1B). For some genes, such as *CRK* and *EPHA7*, potential and perplexity are nearly equal, indicating that isoforms are expressed at approximately uniform levels. Other genes, like *MBD1* and *GNAS*, show a moderate reduction from potential to perplexity, reflecting broad but uneven isoform expression. Finally, genes like *DDB1* and *RPS4X* exhibit a stark disparity between potential and perplexity, suggesting that aside from a few highly expressed isoforms, most are present at very low levels–potentially reflecting high transcriptional and splicing noise (Figure 1C).

Previous studies using ESTs or short read RNA-seq have reported that gene expression and exon count are positively correlated with isoform diversity (Melamud and Moult 2009; Pickrell et al. 2010; Tapial et al. 2017). This relationship is expected: highly expressed genes are more deeply sampled and genes with more exons have more opportunities for variable splice junction usage, both increasing the likelihood of detecting multiple isoforms. Indeed, in the ENCODE4 LRS data, gene potential showed a modest positive correlation with gene expression (R = 0.36, Figure 1D), and a correlation, albeit weaker, with exon count (R = 0.16, data not shown). In contrast, perplexity is uncoupled from gene expression, showing only a very weak *negative* correlation (R = -0.05, Figure 1E). Perplexity showed a similar association with exon count (R = 0.21; data not shown), but this relationship is confounded by technical bias for shorter read lengths–and thus shorter genes. Overall, expression and exon count explain only 4.5% of the variance in gene perplexity, compared to 15.8% for gene potential. This discrepancy suggests that the correlation between gene potential and expression is primarily driven by technical factors–such as increased detection power in highly expressed genes–rather than reflecting genuine differences in isoform regulatory complexity.

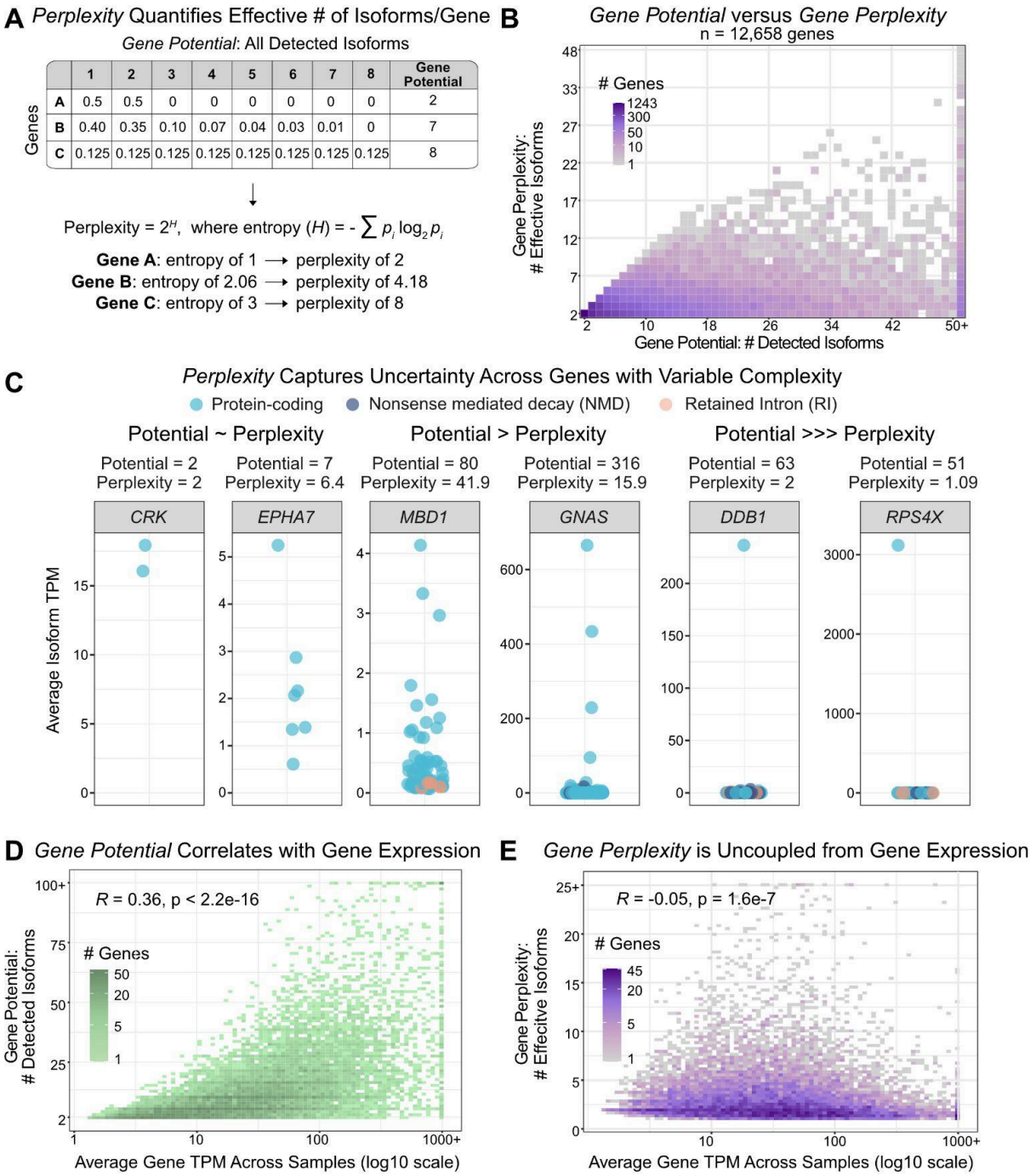

Figure 1. Perplexity Captures RNA Isoform Diversity of Protein-coding Genes in ENCODE4 LRS Data. (A) Illustrative gene examples showing how the distribution of isoform ratios is related to gene potential, entropy and perplexity. (B) Gene potential compared to gene perplexity for genes with >1 protein-coding isoforms. Color intensity corresponds to the number of genes at each point. (C) Six example genes from ENCODE4 LRS data plotting average isoform TPM. Dots are colored based on the transcript biotype: protein-coding, NMD, and RI. (D-E) Average gene expression versus (D) gene potential and (E) gene perplexity for genes with

multiple annotated protein-coding isoforms. Samples with a gene expression of 0 are excluded from the average. Color intensity corresponds to the number of genes at each point. Pearson's *R* is shown.

Perplexity to Evaluate Protein Isoform Diversity

In its broadest formulation, gene perplexity was calculated using all detected transcripts, including protein-coding, NMD, RI, and noORF biotypes (Figure 1). However, perplexity can also be recalculated on subsets of transcripts to capture different layers of gene regulation. For example, excluding non-coding transcripts (i.e., NMD, RI, noORF) and re-scaling expression ratios among the remaining protein-coding transcripts yields *protein-coding (pc) transcript perplexity*, which reflects diversity among transcripts capable of being translated. To focus more specifically on protein-level diversity, transcripts that encode the same open reading frame (ORF) can be collapsed to give *ORF perplexity*, reflecting instead the effective number of distinct protein products produced by each gene (Figure 2A).

As expected, isoform diversity decreases at each regulatory level: from an average gene perplexity (all transcripts) of 3.4 to a perplexity of 2.7 for pc transcripts and 2.1 for ORFs (Figure 2B). The final value suggests that, on average, each gene produces approximately two distinct proteins that are reliably expressed across samples. On a gene-by-gene basis, however, the degree of change at each level varies greatly–in fact, genes cluster into four groups based on these changes (Figure C-E). *UTR Diverse* genes show similar gene and pc transcript perplexity but low ORF perplexity, indicating that most RNA isoforms are protein-coding and share the same ORF, differing primarily in untranslated regions (top row). *Non-coding Dominant* genes exhibit a sharp decrease from gene to pc transcript perplexity, suggesting that most of their RNA diversity comes from non-coding transcripts, such as NMD, RI, and noORF (2nd row). *Protein Dominant* genes show minimal change across all three levels, indicating that most of their isoforms are protein-coding and produce distinct proteins (3rd row). *Hybrid* genes fall between these extremes with moderate reductions in perplexity across levels, suggesting a mix of functional and redundant isoform diversity (4th row). Interestingly, despite these specific changes across levels, all four groups converge on a similar average ORF perplexity between 1.9-2.2 (Figure 2D).

We found that patterns of perplexity sub-types vary by functional gene categories (Figure 2F). Genes encoding regulatory proteins, such as chromatin regulators, RNA binding proteins, and transcription factors show the highest perplexity at all levels (Lambourne et al. 2024; Talavera, Orozco, and de la Cruz 2009), with ORF perplexities of 2.9, 2.8, and 2.3, respectively. In contrast, housekeeping genes and chaperones have high average gene potential but the lowest ORF perplexities, likely reflecting pervasive low-level transcription rather than functional diversification. These trends highlight that even though housekeeping genes may express a far higher number of distinct transcripts overall, regulatory gene classes exhibit greater proteomic diversity.

Perplexity as a Flexible Threshold to Classify Effective Isoforms

While perplexity offers key advantages over fixed TPM cutoffs–incorporating all detected isoforms, minimizing arbitrary thresholding, and enabling consistency across studies–practically speaking, downstream applications still require the removal of noisy isoforms. To address this, we adapted perplexity as a flexible, information-based threshold to distinguish potentially *effective* from *ineffective* isoforms. In the example shown for a representative gene in Figure 2A, the gene perplexity of 6.59 is rounded to 7, and then the seven most expressed isoforms are deemed effective. Unlike fixed TPM cutoffs, this perplexity-derived threshold adapts to varying gene complexity.

Applied transcriptome-wide, this strategy classifies 42,611 isoforms (23%) as effective (mean 29.2 TPM) and 142,549 as ineffective (mean 0.62 TPM; Supplemental Figure 2A). Notably, 98% of detected canonical isoforms and 49% of other annotated isoforms are effective, compared to just 13% of novel isoforms (Supplemental Figure 2B). This same classification strategy can be applied consistently across gene-(as done here), pc transcript-, and ORF-level perplexities.

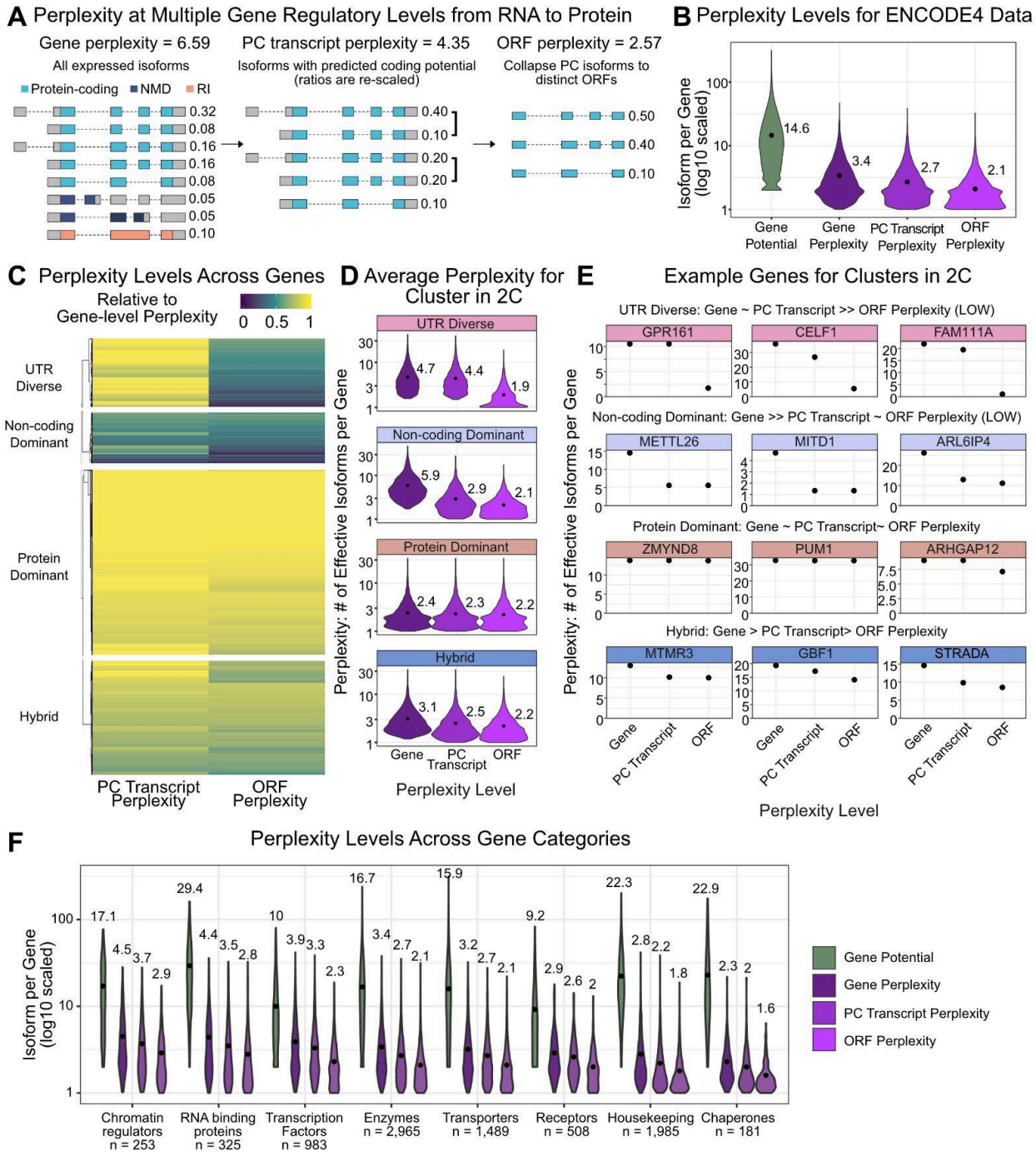

Figure 2. Perplexity to Evaluate Protein Isoform Diversity. (A) Isoform diversity of each gene can be evaluated at three levels: gene, protein-coding (pc) transcript, and ORF. Exon colors correspond to the isoform category. Isoform ratios are given on the right. (B) Number of isoforms per gene across all genes for gene potential and each perplexity level. Numbers next to the violins are the average isoforms per gene for each metric. (C) Genes on the y-axis are hierarchically clustered by Euclidean distance, based on the relative drop in perplexity from gene-level to pc transcript and ORF levels. Four groups are labelled based on changes across

perplexity levels. (D) Average perplexity across the three levels for each gene cluster in C. Numbers give the average perplexity for each level in each cluster. (E) Three example genes from each cluster in C are plotted. For each gene, gene-level perplexity is plotted alongside pc transcript and ORF perplexity. (F) Distributions of gene potential and gene, pc transcript, and ORF perplexity across gene categories. The average per gene in each category is labelled above each violin. The x-axis is ordered by the average ORF perplexity. The number of genes in each category are given on the x-axis.

ORF Perplexity per Sample to Assess Expression Breadth and Variability

In Figures 1 and 2, perplexity was calculated using isoform ratios averaged across all ENCODE samples. While this captures the overall complexity of isoform usage, it does not reflect how isoform dominance or variability may differ across tissues, which are indicators of functional relevance.

To address this, we first determined how consistently ORFs were expressed across tissues. We classified ORFs as effective or ineffective on a per-sample basis, and calculated the percentage of samples in which each ORF is effective, a metric we refer to as *expression breadth* (Figure 3A). For example, as shown in Figure 3A, ORF 1 is effective in all 4 samples (100% expression breadth), while ORFs 5 and 7 are only effective in a single sample (25% expression breadth). As expected, ORFs deemed effective in the aggregate gene-level perplexity analysis (Figure 2) tend to have higher expression breadth than ineffective ORFs–averaging 63% vs. 3%, respectively (Figure 3B).

While expression breadth tracks ORF presence across tissues, it does not account for how much its relative expression varies. For instance, an ORF that shifts from a usage ratio of 0.1 to 0.9 is more variable, and potentially more tissue-specific, than one that remains consistently between 0.8 and 0.9. To assess such *expression variability*, we quantified the standard deviation of ORF usage ratios across samples (Figure 3A). We found that most ORFs classified as ineffective in the aggregate gene-level perplexity analysis showed minimal variability (mean SD = 0.03; Figure 3B), consistent with these ORFs exhibiting uniform background expression.

To explore the relationship between expression breadth and expression variability, we plotted percent-effective tissues against the standard deviation of ORF usage and assigned each ORF to one of four quadrants (Figure 3C). Quadrant I, *Broad Switching,* includes ORFs with both high breadth and high variability. Quadrant II, *Universal*, contains ORFs with similarly high breadth but low variability. Together, these two quadrants account for the vast majority of canonical ORFs (37% and 57%, respectively; Figure 3D; Supplemental Figure 3B), underscoring their widespread expression signatures. Quadrant III, *Background*, is defined by both low breadth and low variability and contains the majority of ineffective ORFs (99%). Quadrant IV, *Tissue-specific*, contains ORFs with low breadth but high variability, which are potential markers of context-specific function. We find that a large proportion of novel and annotated (not canonical) ORFs classified as effective fall into Quadrant IV, 26% and 21%, respectively. In contrast, canonical (4%) and ineffective (1%) ORFs are rarely classified as

tissue-specific (Figure 3D; Supplemental Figure 3B). Overall, tissue-specific ORFs are most strongly represented in brain, heart, and embryonic tissues, although resolution is limited in many other tissues with only one or two samples represented in ENCODE4 (Supplemental Figure 3C). Receptors, transcription factors, and chromatin regulators have the highest proportion of *tissue-specific* ORFs, while housekeeping genes have the lowest (Supplemental Figure 3D).

This quadrant scheme separates ORFs into four expression behaviors: broad switching, universal stability, consistently low expression, and tissue-specific usage. The *CSDE1* gene, a RNA-binding and translation regulator, illustrates this full space, with at least one ORF in each category (Figure 3D-E). *CSDE1-202* is broadly expressed except in heart and muscle, where *CSDE1-204* is dominant (Figure 3E). Both show high variability across samples, but differ in tissue breadth so that *CSDE1-202* falls into Quadrant I (broad switching) while *CSDE1-204* is in Quadrant IV (tissue-specific). These isoforms differ by a cassette exon included in 204 and skipped in 202 (Figure 3E). In comparison, *CSDE1-203*, the canonical ORF, shows moderate, but consistent expression across samples and falls into Quadrant II (Universal), while *CSDE1-214*, with minimal and uniform expression, belongs to Quadrant III (Background) (Figure 3E). Notably, although isoform-specific functions of *CSDE1* remain poorly characterized, the gene has been implicated in cardiovascular disease, neurodevelopmental and psychiatric disorders, and cancer (Ciocia et al. 2024), highlighting the potential for isoform-specific disease relevance.

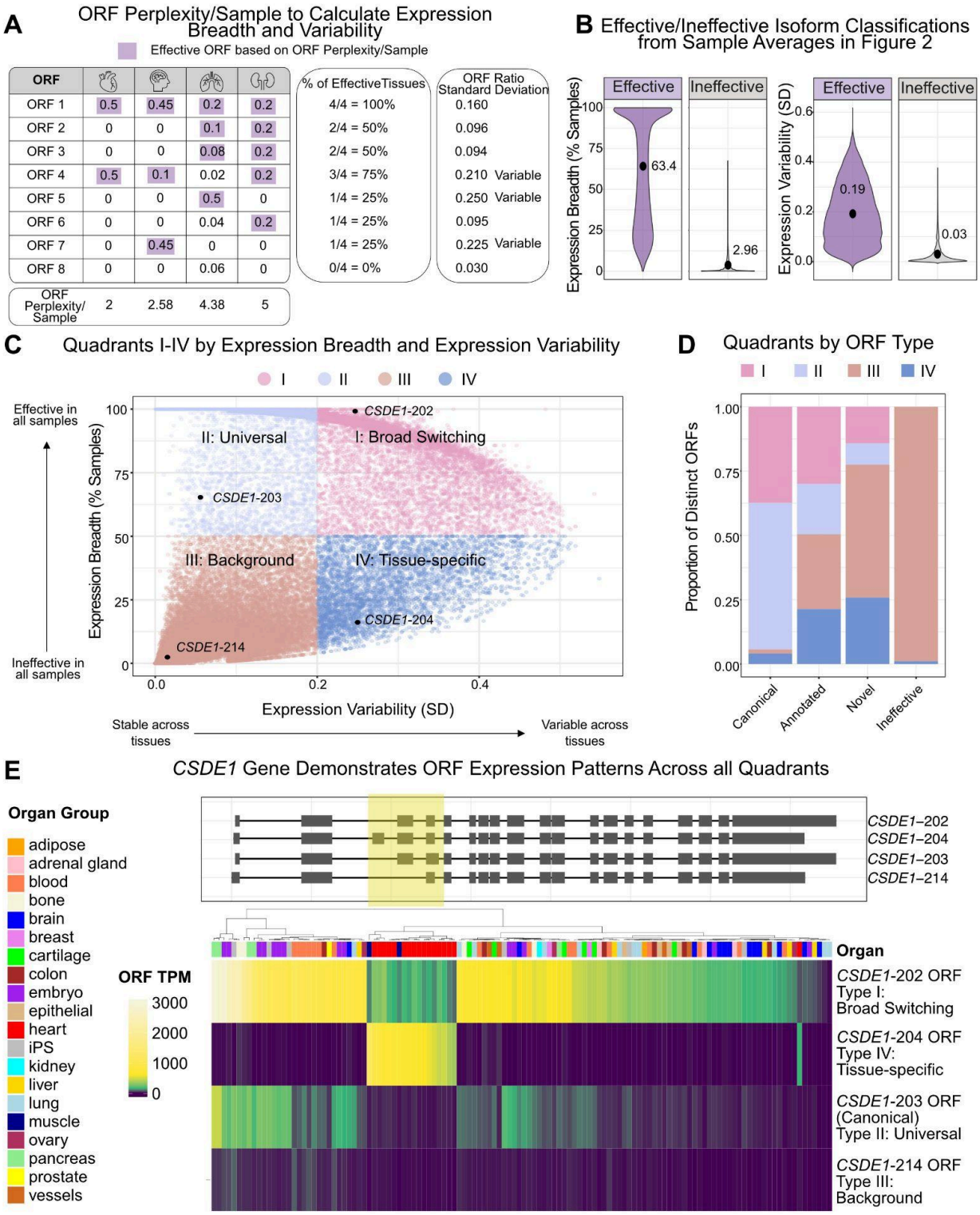

Figure 3. ORF Perplexity per Sample to Assess Expression Breadth and Variability. (A) ORF perplexity calculated per sample to give *Expression Breadth* and *Expression Variability*.

ORFs in purple boxes are those classified as effective in each sample. ORF ratios are only included for samples where the gene is expressed. (B) Effective and ineffective ORF classifications per violin here are calculated from ORF ratios averaged across samples–not per sample. Conversely, y-axes are metrics from sample-specific ratios described in A. The numbers next to each violin are averages. (C) All ORFs are divided into four quadrants based on expression variability (standard deviation; x-axis) versus expression breadth (percent-effective tissues; y-axis) from B. Descriptive labels are given to each quadrant. The *CSDE1* gene has ORFs across all four quadrants–one ORF from each quadrant is marked. (E) Proportion of ORFs from canonical, annotated, novel, and ineffective isoforms in each quadrant from C. (E) Top: Structure of the four *CSDE1* isoforms marked in C. The exons that differ between the isoforms are highlighted in yellow. Bottom: Sample-specific TPM of four *CSDE1* ORFs –each one is marked in a different quadrant in C. Organ types are labelled at the top. The canonical isoform (203) is labelled.

Discussion

The way we define the transcriptome–the full set of RNA products generated by the genome–is a foundational decision that influences nearly every layer of downstream biological interpretation. It shapes our understanding of gene output, guides hypothesis generation, informs the interpretation of disease-associated variants, and is the first step in building system-level models. At its core, a transcriptome is defined by the method used to determine whether a transcript is credible–a seemingly small step that is arguably the most upstream and consequential decision shaping our understanding of functional RNA and protein diversity.

For this critical step, the field currently relies on user-defined thresholds–an ad hoc mix of TPM cutoffs, cross-sample expression support, or isoform abundance ratios–to filter out low-abundance isoforms. These heuristic choices result in widely divergent estimates of transcriptome complexity. As LRS matures, there is now an opportunity to replace arbitrary filters with more systematic, data-driven frameworks.

In this study, we introduced perplexity as a flexible, quantitative metric that is an alternative to arbitrary thresholding for evaluating isoform diversity. Perplexity incorporates the full, unfiltered isoform distribution and adapts to gene-specific structural and expression complexity, providing an interpretable metric of the *effective number of isoforms* that is both rigorous and biologically intuitive.

Perplexity is unique for LRS analysis in that all detected isoforms–regardless of expression level–contribute to the output, rather than being excluded as presumed transcriptional and splicing noise (Melamud and Moult 2009; Wan and Larson 2018). This approach acknowledges that observed molecular diversity is already far more biologically constrained than the theoretical space of all possible isoforms. For example, the *TTN* gene contains 363 exons, which could theoretically produce 4.7×10^{108} isoforms through exon skipping alone–without considering alternative start and ends or other splicing events. Yet only 14 protein-coding

isoforms of *TTN* are annotated in GENCODE, highlighting the remarkable specificity of the transcriptional and splicing machinery (Deveson et al. 2018).

Perplexity is not just one metric, but a generalizable concept for quantifying isoform diversity across multiple levels: gene, transcript, and protein. This captures both the extent and source of a gene's diversity–whether arising from UTRs, non-coding transcripts, or distinct ORFs. These sources of variation suggest biological functions of alternative isoforms, from dosage control via NMD or translational regulation to the production of distinct protein products (Fair et al. 2024; Nilsen and Graveley 2010; Rogalska, Vivori, and Valcárcel 2023). When applied across all samples, perplexity revealed that genes express an average of 3.4 transcript isoforms and 2.1 protein isoforms, offering a data-driven answer to long-standing questions about biologically meaningful isoform diversity.

In this study, we applied perplexity metrics on a comprehensive, cross-tissue LRS dataset. LRS has dramatically expanded our view of the human transcriptome, enabling annotation of hundreds of thousands of novel isoforms. Despite this, the scale of LRS efforts still trails behind short-read studies, and many cell types have not been sequencing so that current references lack many tissue-specific isoforms. In addition, LRS faces two main technical limitations (Reese et al. 2023; Pardo-Palacios, Wang, et al. 2024; Calvo-Roitberg, Daniels, and Pai 2024; Mikheenko et al. 2022). First, incomplete reverse transcription (RT) processivity and PCR amplification bias during library preparation lead to underrepresentation of long transcripts (>12 kb), as observed in (Reese et al. 2023). Second, TSSs can be imprecisely annotated due to read truncation caused by RNA degradation, internal 3' priming, or RT prematurely aborting synthesis before reaching the 5' end (Pardo-Palacios, Arzalluz-Luque, et al. 2024). To minimize overcalling TSS diversity, we collapsed reads to isoforms without requiring identical 5' ends, with the consequence of underestimating 5'UTR variation in our study. Overcoming these technical barriers remains an active area of development with improved RT enzymes, direct RNA-sequencing, 5' cap-capture methods, and validation with orthogonal datasets (CAGE for TSS verification) (Ogami, Oishi, and Hoshino 2023; Pardo-Palacios, Arzalluz-Luque, et al. 2024; Garalde et al. 2018).

While perplexity offers an improved framework to quantify isoform diversity and infer functionality, it remains an expression-based metric. As such, it assumes that highly expressed RNA isoforms are more likely to contribute to gene-level function. However, RNA and protein abundance are only moderately correlated (Gry et al. 2009; Jiang et al. 2020; Schwanhäusser et al. 2011). Low-abundance isoforms can be efficiently translated or result in highly stable proteins (Weatheritt, Sterne-Weiler, and Blencowe 2016; Blencowe 2006), 5'UTR differences can affect translation efficiency, and N-terminal amino acids and internal isoform-specific sequences can influence protein stability (Varshavsky 2011). Therefore, any isoform expression-based proxy for functional impact should be paired with direct detection of protein isoforms, a nascent area of study (R. M. Miller et al. 2022; Mehlferber et al. 2022; Sinitcyn et al. 2023; Kedan et al. 2025), as well as functional validation.

We acknowledge that our perplexity estimates for lowly expressed genes will be slightly downward biased (i.e., we slightly underestimate perplexity) because of limited coverage (G. Miller and Miller 1955). We considered applying existing bias correction approaches, but these have their own limitations: Miller-Madow bias correction requires knowing the true number of possible isoforms (G. Miller and Miller 1955) and the Nemenman-Shafee-Bialek estimator is computationally expensive to calculate requiring numerical integration or Monte Carlo (Nemenman, Shafee, and Bialek 2001). Given the high coverage of the datasets analyzed here, the bias in our estimates will be minimal. This intuition is supported by the slight negative correlation between expression and perplexity in Figure 1E: substantial downward bias for more lowly expressed genes would in contrast result in a *positive* correlation.

We envision perplexity as a broadly applicable framework for evaluating transcript complexity across diverse biological contexts and across inter-laboratory studies. Our full analysis pipeline is available on GitHub, enabling the community to evaluate perplexity across LRS datasets. With minimal modifications, the perplexity framework could be extended to LRS data from other species, and even other isoform-centric data types, including those focused on RNA and protein modifications. Ultimately, by integrating the perplexity-derived metrics introduced here, researchers can begin to systematically prioritize isoforms for functional experimental follow-up.

Methods

Our full LRS pipeline is available as a Bash script on GitHub (https://github.com/daklab/Perplexity_Isoform_MasterTable). The pipeline takes as input raw FASTQ files and a sample-to-tissue mapping file, and outputs a master table containing all of the gene- and isoform-level metrics described below. The master table summarizing ENCODE4 PacBio LRS datasets in this study is in Supplemental Table S3.

Customized long read sequencing analysis pipeline

Download LRS Data from ENCODE

Data was downloaded from ENCODE on September 30th, 2024, in FASTQ format. The dataset was filtered for unperturbed, Homo sapiens, long read RNA-seq on PacBio Sequel II and Sequel platforms. The polyadenylated mRNA sequences were assembled on GRCh38. Based on these filters, 124 samples were used in this study.

Alignment and Filtering of Full Length Reads

Raw reads were aligned with Minimap2 v2.17 (H. Li 2018) using recommended parameters for PacBio long-read sequences: -ax splice:hq -uf. Reads with MAPQ smaller than 60 were filtered out, and only primary aligned reads were kept. BAM output files were converted to BED and PSL files using helper scripts from FLAIR v1.6.1 (Tang et al. 2020): bam2Bed12.py and bed_to_psl.py.

We filtered for full length reads by requiring each read to overlap a GENCODE v46 basic annotated last exon (transcription end sites; TES) and annotated first exon (transcription start sites; TSS). By requiring TES overlap, reads from internal priming were removed. Reads with high confident 3' ends were then overlapped with annotated TSS to remove non-full length reads which included fragments, fragments with retained introns, and reads that aligned fully within the 3'UTR. See the top panel of Supplemental Figure 1A for the number of reads removed at each step.

Collapsing Reads to Isoforms

We wrote a custom script that utilizes the logic of FLAIR collapse (Tang et al. 2020) to collapse full length reads to isoforms based on shared junction coordinates. Only internal junctions were considered in this step–exact read start and ends were not considered. To remove technical artifacts, we required 10 supporting reads across all 124 samples per isoform. Single-exon reads were filtered out. Isoforms with all internal junctions matching with annotated transcripts from GENCODE v46 were assigned with annotated transcript IDs. If more than one annotated transcript matched a collapsed isoform–for two annotated transcripts that only differ in UTR lengths–their transcript IDs were appended with a semicolon. For all isoforms including the novel ones, Ensembl gene IDs from GENCODE v46 were assigned to isoforms using FLAIR helper script: identify_gene_isoform.py. Unannotated genes are assigned with chromosome loci for their gene ID. Python v3.11.5 hash function was used to generate reproducible novel isoform ids based on exon start and end coordinates. TES and TSS of an isoform were defined by the supporting read with the maximum UTR length. Resulting isoforms were output as PSL and GTF files.

Identifying Retained Introns and Open Reading Frames

We ran SQANTI3 v5.1.1 QC (Tardaguila et al. 2018) to identify isoforms with retained introns (RI). All isoforms categorized as "intron_retention" were marked as RI for further analysis.

CPAT v3.0.4 was used to predict open reading frames (ORFs) in the remaining isoforms (L. Wang et al. 2013). To run CPAT, the PSL file was converted to FASTA file format using FLAIR helper script (Tang et al. 2020): psl_to_sequence.py. A prebuilt human logit model and hexamer table were downloaded from CPAT and were used to predict ORFs. A minimum ORF length of 90 nucleotides, 30 amino acids, was required and the top scoring ORF was selected per isoform. We used the faTrans utility v377 from the UCSC Genome Browser tools to translate DNA sequences into amino acid sequences (Karolchik et al. 2004). Isoforms with a predicted ORF were labelled as protein-coding and isoforms with no ORF predicted were labelled as 'noORF'.

At this step, isoforms were overlapped with GENCODE v46 basic annotations. For annotated isoforms, 94% of CPAT ORF predictions matched the annotated ORF. For the isoforms that CPAT got wrong, we used the annotated ORF. In addition, if a gene is not annotated as "protein_coding" in GENCODE v46 basic, we classify all isoforms pertaining to that gene as

"noORF" or "RI". These isoforms get filtered out in Supplemental Figure S1B, since they have protein-coding isoforms equal to 0, and are not considered throughout the paper. Importantly, this step removes pseudogene alignments.

Custom NMD Prediction

A custom script (filter_NMD.py) was used to identify isoforms predicted to undergo nonsense-mediated decay (NMD). The script processes all isoforms with predicted ORFs and classifies an isoform as NMD if it contains a stop codon located upstream of the final splice junction–a premature termination codon (PTC). Additionally, any isoform already annotated as NMD in GENCODE v46 basic were classified as NMD here. Annotated isoforms with GENCODEv46 protein-coding annotation that were predicted to have a PTC were rescued and marked as protein-coding.

Filtering for Protein-coding Diversity

Genes with 0 or 1 protein-coding isoforms were filtered out for our further analysis since we focus on protein-coding genes with isoform diversity. Plots in Figure 1A and Supplemental Figure 1B summarize these isoforms. These isoforms are also contained in the master table.

Master Table Calculations

Gene, Isoform, and ORF IDs: For annotated isoforms, Ensembl (GENCODE v46) gene and isoform ids are included. For novel isoforms, Ensembl gene IDs are assigned based on their loci, by the FLAIR (Tang et al. 2020) helper script previously discussed (identify_gene_isoform.py). Transcript IDs of novel isoforms were generated using hash IDs of internal splicing junctions. This is to ensure that novel isoforms get consistent IDs regardless of the user, time, and environment at the time of read collapsing. This allows novel isoforms to be cross referenced across different studies. ORF IDs are collapsed isoform IDs, separated by a semicolon, based on identical amino acid sequences.

Average gene and isoform TPM: TPM was averaged across all ENCODE LRS samples where the gene is expressed.

Gene Potential (Gene-level Metric): Defined as the total number of detected isoforms per gene, including protein-coding, NMD, RI, and noORF, across all 124 samples.

Gene Perplexity (Gene-level Metric): Represents the *effective number of isoforms* per gene. Perplexity was calculated as 2^H, where $H = -\sum p_i \log_2(p_i)$ is the Shannon entropy and p_i is a probability distribution of isoform i proportions within a gene. For this specific metric, p_i was computed based on the average isoform proportions across all 124 ENCODE samples.

Protein-coding Transcript Potential (Gene-level Metric): Similar to *gene potential*, but restricted to the number of detected protein-coding isoforms per gene.

Protein-coding Transcript Perplexity (Gene-level Metric): Similar to *gene perplexity*, but calculated using only protein-coding isoforms. The p_i values were re-scaled to reflect the expression proportions among protein coding isoforms for each gene.

ORF Potential (Gene-level Metric): Similar to *gene-* and *protein-coding transcript potential*, but based on the number of distinct ORFs detected per gene.

ORF Perplexity (Gene-level Metric): Similar to *gene-* and *protein-coding transcript perplexity*, but isoform proportions were summed for isoforms with the same predicted ORF. The resulting ORF-level proportions were used as p_i to calculate ORF perplexity.

Number of Exons (Gene-level Metric): Calculated as the number of exons from the GENCODE v46 annotated isoform with the greatest number of exons for each gene.

Percent of Effective Tissues (ORF-level Metric): For each sample, ORF perplexity was calculated per gene and rounded to the nearest integer, n. ORFs were then ranked by expression ratio p_i, and the top n ORFs were classified as effective. Percent of effective tissues for a given ORF was calculated by dividing the number of samples in which the ORF was effective by the total number of samples in which the gene was expressed.

ORF ratio standard deviation (ORF-level Metric): Standard deviation of ORF probability across tissues where the ORF is expressed.

Novel Junction (Classification): TRUE for novel isoforms defined as NNC (novel not in catalog) with SQANTI3 versus isoforms that contain new combinations of known splice sites.

Gene Categories (Classification): Genes were assigned to a category based on Protein Atlas (https://www.proteinatlas.org) (Uhlén et al. 2015), ENCODE for RNA-binding proteins (Van Nostrand et al. 2020), and Housekeeping Atlas (Hounkpe et al. 2021). Categories from these resources include RNA-binding proteins, transcription factors, enzymes, receptors, transporters, chaperones, chromatin regulators, actin-binding proteins, and housekeeping genes. Genes without a category here were labelled as ‘Other’. If a gene was in more than one category, each category was included all labelled.

Isoform Effective (Classification): Gene-level perplexity was rounded to n and the n isoforms with the highest average ratios across samples were classified as effective. Effective ORFs per gene were similarly defined from rounded ORF-level perplexity.

Isoform Annotated/Isoform Canonical (Classification): Isoforms with splice junction chains matching transcripts in GENCODE v46 basic annotation were classified as annotated isoforms. Isoforms were further labelled as canonical if any matching transcript ID carried the "Ensembl Canonical" tag. For isoforms whose splice junction chains matched multiple annotated

transcripts, all matching transcript IDs were appended, and the isoform was considered canonical if any of those transcripts were tagged as Ensembl Canonical.

Acknowledgements
M.D.S. was supported by NIH/NIGMS (F32GM142213). S.H.P. was partially supported by a Columbia University Fu Foundation School of Engineering and Applied Science Bonomi Scholarship and Columbia University Scholars Program Summer Enhancement Fellowship. G.M.S. is supported by NIH/NCI (R33CA281919). D.A.K. is supported by Columbia University and NYGC startup funds, the MacMillan Center for the Study of the Non-Coding Cancer Genome, NIH/NCI (R21CA272345), and NSF CAREER DBI2146398. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions
M.D.S. and D.A.K. conceived the study. M.D.S., S.H.P, and J.S performed analyses. M.D.S., S.H.P, G.M.S., and D.A.K. wrote the paper with input from co-authors. M.D.S, G.M.S., and D.A.K. supervised the study.

Competing Interests
G.M.S. is on the scientific advisory board of Quantum-Si Incorporated and holds stock in Quantum-Si Incorporated.

Data Availability
No new sequencing data was generated in this study. All PacBio long read RNA-sequencing data processed in this study was previously published through ENCODE4 and is publicly available at: https://www.encodeproject.org/.

Code Availability
The bash script containing the full analysis pipeline is available on github: https://github.com/daklab/Perplexity_Isoform_MasterTable. The master table generated as a part of this study is included as Supplemental Table S3.

References

Al'Khafaji, Aziz M., Jonathan T. Smith, Kiran V. Garimella, Mehrtash Babadi, Victoria Popic, Moshe Sade-Feldman, Michael Gatzen, et al. 2023. "High-Throughput RNA Isoform Sequencing Using Programmed cDNA Concatenation." *Nature Biotechnology*, June, 1–5.
Blencowe, Benjamin J. 2006. "Alternative Splicing: New Insights from Global Analyses." *Cell* 126 (1): 37–47.
Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (5): 525–27.
Calvo-Roitberg, Ezequiel, Rachel F. Daniels, and Athma A. Pai. 2024. "Challenges in Identifying mRNA Transcript Starts and Ends from Long-Read Sequencing Data." *Genome Research* 34 (11): 1719–34.
Ciocia, Annagiulia, Neus Mestre-Farràs, Ignacio Vicent-Nacht, Tanit Guitart, and Fátima

Gebauer. 2024. “CSDE1: A Versatile Regulator of Gene Expression in Cancer.” *NAR Cancer* 6 (2): zcae014.
Deveson, Ira W., Marion E. Brunck, James Blackburn, Elizabeth Tseng, Ting Hon, Tyson A. Clark, Michael B. Clark, et al. 2018. “Universal Alternative Splicing of Noncoding Exons.” *Cell Systems* 6 (2): 245–55.e5.
ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.
Fair, Benjamin, Carlos F. Buen Abad Najar, Junxing Zhao, Stephanie Lozano, Austin Reilly, Gabriela Mossian, Jonathan P. Staley, Jingxin Wang, and Yang I. Li. 2024. “Global Impact of Unproductive Splicing on Human Gene Expression.” *Nature Genetics* 56 (9): 1851–61.
Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, et al. 2019. “GENCODE Reference Annotation for the Human and Mouse Genomes.” *Nucleic Acids Research* 47 (D1): D766–73.
Gandrillon, Olivier, Mathilde Gaillard, Thibault Espinasse, Nicolas B. Garnier, Charles Dussiau, Olivier Kosmider, and Pierre Sujobert. 2021. “Entropy as a Measure of Variability and Stemness in Single-Cell Transcriptomics.” *Current Opinion in Systems Biology* 27 (100348): 100348.
Garalde, Daniel R., Elizabeth A. Snell, Daniel Jachimowicz, Botond Sipos, Joseph H. Lloyd, Mark Bruce, Nadia Pantic, et al. 2018. “Highly Parallel Direct RNA Sequencing on an Array of Nanopores.” *Nature Methods* 15 (3): 201–6.
Gerhard, Daniela S., Lukas Wagner, Elise A. Feingold, Carolyn M. Shenmen, Lynette H. Grouse, Greg Schuler, Steven L. Klein, et al. 2004. “The Status, Quality, and Expansion of the NIH Full-Length cDNA Project: The Mammalian Gene Collection (MGC).” *Genome Research* 14 (10B): 2121–27.
Glinos, Dafni A., Garrett Garborcauskas, Paul Hoffman, Nava Ehsan, Lihua Jiang, Alper Gokden, Xiaoguang Dai, et al. 2022. “Transcriptome Variation in Human Tissues Revealed by Long-Read Sequencing.” *Nature* 608 (7922): 353–59.
Gry, Marcus, Rebecca Rimini, Sara Strömberg, Anna Asplund, Fredrik Pontén, Mathias Uhlén, and Peter Nilsson. 2009. “Correlations between RNA and Protein Expression Profiles in 23 Human Cell Lines.” *BMC Genomics* 10 (1): 365.
Harrow, Jennifer, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, et al. 2006. “GENCODE: Producing a Reference Annotation for ENCODE.” *Genome Biology* 7 Suppl 1 (Suppl 1): S4.1–9.
Harrow, Jennifer, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, et al. 2012. “GENCODE: The Reference Human Genome Annotation for The ENCODE Project.” *Genome Research* 22 (9): 1760–74.
Hounkpe, Bidossessi Wilfried, Francine Chenou, Franciele de Lima, and Erich Vinicius De Paula. 2021. “HRT Atlas v1.0 Database: Redefining Human and Mouse Housekeeping Genes and Candidate Reference Transcripts by Mining Massive RNA-Seq Datasets.” *Nucleic Acids Research* 49 (D1): D947–55.
Humphrey, Jack, Erica Brophy, Roman Kosoy, Biao Zeng, Elena Coccia, Daniele Mattei, Ashvin Ravi, et al. 2025. “Long-Read RNA Sequencing Atlas of Human Microglia Isoforms Elucidates Disease-Associated Genetic Regulation of Splicing.” *Nature Genetics* 57 (3): 604–15.
Jiang, Lihua, Meng Wang, Shin Lin, Ruiqi Jian, Xiao Li, Joanne Chan, Guanlan Dong, et al. 2020. “A Quantitative Proteome Map of the Human Body.” *Cell* 183 (1): 269–83.e19.
Jones, Emma F., Anisha Haldar, Vishal H. Oza, and Brittany N. Lasseigne. 2024. “Quantifying Transcriptome Diversity: A Review.” *Briefings in Functional Genomics* 23 (2): 83–94.
Karolchik, Donna, Angela S. Hinrichs, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. 2004. “The UCSC Table Browser Data Retrieval Tool.” *Nucleic Acids Research* 32 (Database issue): D493–96.

Kedan, Amir, Henrik Zauber, Meng-Ran Wang, Qionghua Zhu, Liang Fang, Wei Chen, and Matthias Selbach. 2025. “An Integrated Landscape of mRNA and Protein Isoforms.” *bioRxiv*. https://doi.org/10.1101/2025.04.16.648713.

Lambourne, Luke, Kaia Mattioli, Clarissa Santoso, Gloria Sheynkman, Sachi Inukai, Babita Kaundal, Anna Berenson, et al. 2024. “Widespread Variation in Molecular Interactions and Regulatory Properties among Transcription Factor Isoforms.” *bioRxivorg*. https://doi.org/10.1101/2024.03.12.584681.

Leung, Szi Kay, Aaron R. Jeffries, Isabel Castanho, Ben T. Jordan, Karen Moore, Jonathan P. Davies, Emma L. Dempster, et al. 2021. “Full-Length Transcript Sequencing of Human and Mouse Cerebral Cortex Identifies Widespread Isoform Diversity and Alternative Splicing.” *Cell Reports* 37 (7): 110022.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18): 3094–3100.

Li, Yang I., David A. Knowles, Jack Humphrey, Alvaro N. Barbeira, Scott P. Dickinson, Hae Kyung Im, and Jonathan K. Pritchard. 2017. “Annotation-Free Quantification of RNA Splicing Using LeafCutter.” *Nature Genetics* 50 (1): 151–58.

MacArthur, Ben D., and Ihor R. Lemischka. 2013. “Statistical Mechanics of Pluripotency.” *Cell* 154 (3): 484–89.

Martínez, Octavio, and M. Humberto Reyes-Valdés. 2008. “Defining Diversity, Specialization, and Gene Specificity in Transcriptomes through Information Theory.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (28): 9709–14.

Mehlferber, Madison M., Erin D. Jeffery, Jamie Saquing, Ben T. Jordan, Leon Sheynkman, Mayank Murali, Gael Genet, Bipul R. Acharya, Karen K. Hirschi, and Gloria M. Sheynkman. 2022. “Characterization of Protein Isoform Diversity in Human Umbilical Vein Endothelial Cells via Long-Read Proteogenomics.” *RNA Biology* 19 (1): 1228–43.

Melamud, Eugene, and John Moult. 2009. “Stochastic Noise in Splicing Machinery.” *Nucleic Acids Research* 37 (14): 4873–86.

Mikheenko, Alla, Andrey D. Prjibelski, Anoushka Joglekar, and Hagen U. Tilgner. 2022. “Sequencing of Individual Barcoded cDNAs Using Pacific Biosciences and Oxford Nanopore Technologies Reveals Platform-Specific Error Patterns.” *Genome Research* 32 (4): 726–37.

Miller, G., and G. Miller. 1955. “Note on the Bias of Information Estimates.” *Information Theory in Psychology: Problems and Methods*. https://cir.nii.ac.jp/crid/1573387449968140928.

Miller, Rachel M., Ben T. Jordan, Madison M. Mehlferber, Erin D. Jeffery, Christina Chatzipantsiou, Simi Kaur, Robert J. Millikin, et al. 2022. “Enhanced Protein Isoform Characterization through Long-Read Proteogenomics.” *Genome Biology* 23 (1): 69.

Monzó, Carolina, Tianyuan Liu, and Ana Conesa. 2025. “Transcriptomics in the Era of Long-Read Sequencing.” *Nature Reviews. Genetics*, March, 1–21.

Nemenman, I., Fariel Shafee, and W. Bialek. 2001. “Entropy and Inference, Revisited.” Edited by T. Dietterich, S. Becker, and Z. Ghahramani. *Neural Information Processing Systems* 14 (January):471–78.

Nilsen, Timothy W., and Brenton R. Graveley. 2010. “Expansion of the Eukaryotic Proteome by Alternative Splicing.” *Nature* 463 (7280): 457–63.

Ogami, Koichi, Yuka Oishi, and Shin-Ichi Hoshino. 2023. “Protocol for Analyzing Intact mRNA poly(A) Tail Length Using Nanopore Direct RNA Sequencing.” *STAR Protocols* 4 (2): 102340.

Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. 2008. “Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing.” *Nature Genetics* 40 (12): 1413–15.

Pardo-Palacios, Francisco J., Angeles Arzalluz-Luque, Liudmyla Kondratova, Pedro Salguero, Jorge Mestre-Tomás, Rocío Amorín, Eva Estevan-Morió, et al. 2024. “SQANTI3: Curation

of Long-Read Transcriptomes for Accurate Identification of Known and Novel Isoforms." *Nature Methods* 21 (5): 793–97.
Pardo-Palacios, Francisco J., Dingjie Wang, Fairlie Reese, Mark Diekhans, Sílvia Carbonell-Sala, Brian Williams, Jane E. Loveland, et al. 2024. "Systematic Assessment of Long-Read RNA-Seq Methods for Transcript Identification and Quantification." *Nature Methods* 21 (7): 1349–63.
Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19.
Pickrell, Joseph K., Athma A. Pai, Yoav Gilad, and Jonathan K. Pritchard. 2010. "Noisy Splicing Drives mRNA Isoform Diversity in Human Cells." *PLoS Genetics* 6 (12): e1001236.
Reese, Fairlie, Brian Williams, Gabriela Balderrama-Gutierrez, Dana Wyman, Muhammed Hasan Çelik, Elisabeth Rebboah, Narges Rezaie, et al. 2023. "The ENCODE4 Long-Read RNA-Seq Collection Reveals Distinct Classes of Transcript Structure Diversity." *bioRxiv : The Preprint Server for Biology*, May. https://doi.org/10.1101/2023.05.15.540865.
Ritchie, William, Samuel Granjeaud, Denis Puthier, and Daniel Gautheret. 2008. "Entropy Measures Quantify Global Splicing Disorders in Cancer." *PLoS Computational Biology* 4 (3): e1000011.
Rogalska, Malgorzata Ewa, Claudia Vivori, and Juan Valcárcel. 2023. "Regulation of Pre-mRNA Splicing: Roles in Physiology and Disease, and Therapeutic Prospects." *Nature Reviews. Genetics* 24 (4): 251–69.
Sarantopoulou, Dimitra, Thomas G. Brooks, Soumyashant Nayak, Antonijo Mrčela, Nicholas F. Lahens, and Gregory R. Grant. 2021. "Comparative Evaluation of Full-Length Isoform Quantification from RNA-Seq." *BMC Bioinformatics* 22 (1): 266.
Schwanhäusser, Björn, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. 2011. "Global Quantification of Mammalian Gene Expression Control." *Nature* 473 (7347): 337–42.
Sharon, Donald, Hagen Tilgner, Fabian Grubert, and Michael Snyder. 2013. "A Single-Molecule Long-Read Survey of the Human Transcriptome." *Nature Biotechnology* 31 (11): 1009–14.
Shen, Shihao, Juw Won Park, Zhi-Xiang Lu, Lan Lin, Michael D. Henry, Ying Nian Wu, Qing Zhou, and Yi Xing. 2014. "rMATS: Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data." *Proceedings of the National Academy of Sciences of the United States of America* 111 (51): E5593–5601.
Sinitcyn, Pavel, Alicia L. Richards, Robert J. Weatheritt, Dain R. Brademan, Harald Marx, Evgenia Shishkova, Jesse G. Meyer, et al. 2023. "Global Detection of Human Variants and Isoforms by Deep Proteome Sequencing." *Nature Biotechnology* 41 (12): 1776–86.
Steijger, Tamara, Josep F. Abril, Pär G. Engström, Felix Kokocinski, RGASP Consortium, Tim J. Hubbard, Roderic Guigó, Jennifer Harrow, and Paul Bertone. 2013. "Assessment of Transcript Reconstruction Methods for RNA-Seq." *Nature Methods* 10 (12): 1177–84.
Sterne-Weiler, Timothy, Robert J. Weatheritt, Andrew J. Best, Kevin C. H. Ha, and Benjamin J. Blencowe. 2018. "Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop." *Molecular Cell* 72 (1): 187–200.e6.
Talavera, David, Modesto Orozco, and Xavier de la Cruz. 2009. "Alternative Splicing of Transcription Factors' Genes: Beyond the Increase of Proteome Diversity." *Comparative and Functional Genomics* 2009 (July):905894.
Tang, Alison D., Cameron M. Soulette, Marijke J. van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J. Wu, and Angela N. Brooks. 2020. "Full-Length Transcript Characterization of SF3B1 Mutation in Chronic Lymphocytic Leukemia Reveals Downregulation of Retained Introns." *Nature Communications* 11 (1): 1–12.
Tapial, Javier, Kevin C. H. Ha, Timothy Sterne-Weiler, André Gohr, Ulrich Braunschweig, Antonio Hermoso-Pulido, Mathieu Quesnel-Vallières, et al. 2017. "An Atlas of Alternative

Splicing Profiles and Functional Associations Reveals New Regulatory Programs and Genes That Simultaneously Express Multiple Major Isoforms." *Genome Research* 27 (10): 1759–68.

Tardaguila, Manuel, Lorena de la Fuente, Cristina Marti, Cécile Pereira, Francisco Jose Pardo-Palacios, Hector Del Risco, Marc Ferrell, et al. 2018. "SQANTI: Extensive Characterization of Long-Read Transcript Sequences for Quality Control in Full-Length Transcriptome Identification and Quantification." *Genome Research* 28 (3): 396–411.

Trapnell, Cole, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. 2010. "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation." *Nature Biotechnology* 28 (5): 511–15.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. "Proteomics. Tissue-Based Map of the Human Proteome." *Science (New York, N.Y.)* 347 (6220): 1260419.

Van Nostrand, Eric L., Peter Freese, Gabriel A. Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M. Blue, et al. 2020. "A Large-Scale Binding and Functional Map of Human RNA-Binding Proteins." *Nature* 583 (7818): 711–19.

Varshavsky, Alexander. 2011. "The N-End Rule Pathway and Regulation by Proteolysis: The N-End Rule Pathway." *Protein Science: A Publication of the Protein Society* 20 (8): 1298–1345.

Wang, Eric T., Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. 2008. "Alternative Isoform Regulation in Human Tissue Transcriptomes." *Nature* 456 (7221): 470–76.

Wang, Liguo, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean-Pierre Kocher, and Wei Li. 2013. "CPAT: Coding-Potential Assessment Tool Using an Alignment-Free Logistic Regression Model." *Nucleic Acids Research* 41 (6): e74.

Wan, Yihan, and Daniel R. Larson. 2018. "Splicing Heterogeneity: Separating Signal from Noise." *Genome Biology* 19 (1): 86.

Weatheritt, Robert J., Timothy Sterne-Weiler, and Benjamin J. Blencowe. 2016. "The Ribosome-Engaged Landscape of Alternative Splicing." *Nature Structural & Molecular Biology* 23 (12): 1117–23.

Workman, Rachael E., Alison D. Tang, Paul S. Tang, Miten Jain, John R. Tyson, Roham Razaghi, Philip C. Zuzarte, et al. 2019. "Nanopore Native RNA Sequencing of a Human poly(A) Transcriptome." *Nature Methods* 16 (12): 1297–1305.

Zhang, Chi, Baohong Zhang, Lih-Ling Lin, and Shanrong Zhao. 2017. "Evaluation and Comparison of Computational Tools for RNA-Seq Isoform Quantification." *BMC Genomics* 18 (1): 583.