UNIVERSITY OF CALIFORNIA
SANTA CRUZ

CONNECTING ALTERNATIVE RNA PROCESSING TO POST-TRANSCRIPTIONAL REGULATORY OUTCOMES

A thesis submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY
in
BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Alexander Julian Ritter

June 2024

The Dissertation of A.J. Ritter
is approved:

_____
Professor Rebecca DuBois, Chair

_____
Professor Joshua Stuart

_____
Professor Jeremy Sanford

_____
Peter Biehl
Vice Provost and Dean of Graduate Studies

CONTENTS

Abstract

Alexander J. Ritter

Connecting Alternative RNA Processing to Post-transcriptional Regulatory Outcomes

High throughput RNA sequencing (RNA-seq), and more recently, long read (LR) RNA-seq have revolutionized the study of gene expression. We're able to sequence massive libraries of transcriptomic data upon which we can apply manifold analytical approaches to extract meaningful and actionable biological findings. Short read RNA-seq remains the prevailing method for transcriptomic characterization, owed to its capacity to accurately quantify gene expression at the RNA-level and to capture a wealth of information for the study of alternative RNA processing. However, an intrinsic shortcoming of short read RNA-seq is its reliance on small fragments of messenger RNAs (mRNA) to infer complete transcript structures and to resolve isoform-level expression. Additionally, when used on its own, it lacks the multidimensionality necessary to comprehensively distinguish the modes of regulation (transcriptional vs. post-transcriptional) that underlie changes in RNA abundance between conditions or to accurately infer the translational output of mRNAs.

Here, I present my work to integrate small RNA (sRNA) and mRNA sequencing approaches to explore SARS-CoV-2 (SC2) infection-mediated perturbations to the host mRNA and sRNA landscape (Chapter 2). I show that dozens of human microRNAs (miR) and novel SC2-derived small viral RNAs (svRNA) are dynamically expressed during SC2 infection, and I propose the intriguing hypothesis that several of the svRNAs may function like miRs to confer pleiotropic regulatory impacts to the host transcriptome. I further present my work on a bioinformatic tool called *junctionCounts*, which seeks to comprehensively characterize alternative splicing (AS) events in RNA-seq data (Chapter 3). In concert with its partner utilities *cdsInsertion* and *findSwitchEvents*, *junctionCounts* stands apart from other

AS analysis tools both by profiling non-canonical event types and by predicting functional outcomes of AS events including nonsense-mediated decay (NMD) and coding-to-noncoding switches induced by the inclusion or exclusion of alternative exons, introns or splice sites.

Finally, in Chapter 4, I present my work on the development of a translatomic method called *long read subcellular fractionation and sequencing* (LR Frac-seq). I propose a framework for integrating both LR and short read Frac-seq data to faithfully capture the complete structures of ribosome-associated transcripts from long reads, and to accurately quantify them utilizing the superior throughput of short reads. I show that isoform-specific ribosome association is pervasive and consistent across embryonic stem cells and neuronal progenitor cells, and I propose this approach as a novel way to study AS coupled with translational control (ASTC).

Acknowledgements

I'm grateful to my advisor, Jeremy Sanford, for entrusting me with several, diverse projects that allowed me to gain an appreciation for a breadth of topics in post-transcriptional regulation and to gain the kind of experience and scientific intuition that can only be earned through thousands of hours of research work. Besides providing the raw material for the main work, he facilitated two of the most pivotal experiences thus far in my development as a bioinformatician and as a scientist. The first was a Summer internship in 2022 as a Computational Biologist at Genentech, Inc. under the supervision of Dr. Timothy Sterne-Weiler. It was there that I learned how much room I had still to grow as a bioinformatician, and where I began to think more creatively about data analysis. The second was when I gave a talk about my work on LR Frac-seq at the 2023 Cold Spring Harbor Eukaryotic mRNA Processing meeting. It was that experience that re-ignited my aspiration to pursue a career in academic research and taught me how to think about research projects in new and improved ways.

Chapter 1: Introduction

1.1 Pre-mRNA processing and the establishment of the *cis*-regulatory landscape

The central dogma of molecular biology posits a paradigm in which genetic information flows unidirectionally as such: DNA is transcribed into mRNA, and mRNA is subsequently translated into protein. Within this model are myriad nuanced complexities that can affect the fate of mRNAs before translation, and contradictory to it are classes of RNAs that play functional roles without ever being translated, for example. While the factors that control the process of DNA being transcribed into RNA, termed "transcriptional regulation", are beyond the scope of the work herein, I'll briefly mention that *trans*-acting factors (molecular machines including transcription factors, RNA polymerases, etc.) associate with DNA-encoded *cis*-elements (i.e. distinct sequence motifs and/or secondary structures) to initiate, terminate and otherwise regulate transcription.

Analogous to this framework is the relationship between RNAs and *trans*-acting factors that recognize RNA-encoded *cis*-elements to enact the regulatory functions comprising "post-transcriptional regulation". These *trans*-acting factors include RNA-binding proteins (RBP) and ribonucleoproteins (RNP) that orchestrate and modulate: pre-mRNA processing, mRNA export, localization, stability and translation. One critical step in eukaryotic pre-mRNA processing, which is an emphasis of my work, is alternative splicing (AS). AS is a process that occurs in the vast majority of human protein coding genes by which splicing factors excise introns from multi-exon genes to produce distinct isoforms from a single genomic locus. While AS most obviously diversifies the proteome, it importantly also shapes the *cis*-regulatory landscape of individual isoforms by the inclusion and exclusion of particular sequences. Thus, AS calibrates the repertoire of *trans*-acting factors that can associate with alternative isoforms to affect their cytosolic fate.

1.2 Modes of post-transcriptional regulation that impact mRNA stability and translation

Because the protein products of mRNAs are customarily considered the pinnacle of gene expression, two factors in particular come into frame as important determinants of a given transcript's translational output: mRNA stability and translatability. Nucleotide sequence and concomitant GC-content can in themselves affect the disposition of transcript local secondary structures in ways that promote or diminish stability. Additionally, the *cis*-regulatory code embedded in the 5' untranslated region (UTR), the 3' UTR *and* the coding sequence (CDS) has been shown to affect mRNA stability by facilitating interactions with *trans*-acting factors that can either directly or indirectly impose stabilizing or degradative effects. One such factor is an RNP called the RNA-induced silencing complex (RISC) which is composed of an Argonaute protein associated with a microRNA (miR) that can guide it to complementary mRNA sequences for subsequent cleavage or translational repression. I will expand upon the topic of RISC and small noncoding RNAs in Chapter 2.

Arriving at the point of translation – upon which one or more ribosomes traverse the open reading frame(s) (ORF) of a transcript to generate its encoded protein product – intrinsic features (i.e. optimality of codons, length and nucleotide composition of UTRs, etc.) and *trans*-acting factors can similarly confer translational control of transcripts at each stage: from translation initiation to elongation to termination. Evocative of Ouroboros, the serpent eating its own tail, in more ways than one, the nature of the ribosome-mRNA interaction itself can influence transcript stability. One extremely well-documented example of this phenomenon across virtually all eukaryotes is a translation-dependent mRNA surveillance mechanism called nonsense-mediated decay (NMD). This process involves ribosome encounter with a premature termination codon (PTC) during translation, which recruits NMD factors to accelerate the degradation of the associated transcript to prevent the generation of aberrant,

truncated proteins or to otherwise modulate gene expression. I will elaborate on the topics of translational control and NMD in Chapters 3 and 4.

Chapter 2: Small Noncoding RNAs as Regulators of Gene Expression

2.1 Chapter Introduction

The primary subject of this chapter's project is RISC. Initially, we set out to examine the effects of SARS-CoV-2 infection on the host transcriptome, with special interest to host small noncoding RNAs. We wanted to test the hypothesis that part of the widespread transcriptomic dysregulation observed during SARS-CoV-2 infection is due to infection-induced changes to miR expression. Because individual miRs can have hundreds of potential targets, perturbations to their steady-state levels can dramatically alter the transcriptome. To that end, our collaborators in the Mishra Lab at Columbia University conducted time course infection experiments in African green monkey kidney cells and human lung epithelial cells from which we isolated sRNA and mRNA for sequencing.

In support of our hypothesis, we identified 28 human miRs that were dynamically regulated during SARS-CoV-2 infection. More surprisingly, however, we identified a subset of sRNAs that mapped to the SARS-CoV-2 genome at distinct loci across both human and African green monkey cell models and were plausibly not degradation products. Dozens of these small viral RNAs (svRNA) resembled miRs by their lengths, were predicted to hybridize stably by their seed sequences with complementary sequences in human miRs and mRNAs, and some were attributable to precursor RNAs that were predicted to form hairpins *in silico*. This finding led to the captivating and somewhat controversial hypothesis that SARS-CoV-2-derived svRNAs could directly antagonize host miRs and mRNAs, with or without the aid of Argonaute 2.

My contributions as a co-first author on this project include: all data analysis, some data visualization, writing of the methods and editing of the manuscript. This work was published in *Scientific Reports* on December 15, 2022. All supplementary materials are available at the online publication.

2.2 A potential role for SARS-CoV-2 small viral RNAs in targeting host microRNAs and modulating gene expression

Zachary T. Neeb[1+], Alexander J. Ritter[2+], Lokendra V. Chauhan[3], Sol Katzman[4], W. Ian Lipkin[3], Nischay Mishra[3*], and Jeremy R. Sanford[1*]

[+]These authors contributed equally to this work.
[1]Department of Molecular, Cell and Developmental Biology and Center for Molecular Biology of RNA, University of California Santa Cruz, Santa Cruz, CA, USA.
[2]Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, CA, USA.
[3]Center for Infection and Immunity, Mailman School of Public Health, Columbia University, New York, NY, USA.
[4]Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA.

*Correspondence: nm2641@cumc.columbia.edu (N.M.), jsanfor2@ucsc.edu (J.R.S.)

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causes coronavirus disease (COVID-19) in humans, which may be fatal. We used a comparative transcriptomics approach to investigate the effects of SARS-CoV-2 infection on the host mRNA and sRNA expression machinery in a human lung epithelial cell line (Calu-3) and an African green monkey kidney cell line. Upon infection, we observed global changes in host gene expression and differential expression of dozens of host miRNAs, many with known links to viral infection and immune response. Importantly, we also discovered an expanded landscape of more than a hundred SARS-CoV-2-derived small viral RNAs (svRNAs), predicted to interact with differentially expressed host mRNAs and miRNAs. svRNAs are derived from distinct regions of the viral genome and sequence signatures suggest they are produced by a non-canonical biogenesis pathway. 52 of the 67 svRNAs identified in Calu-3 cells are predicted to interact with differentially expressed miRNAs, with many svRNAs having multiple targets. Accordingly, we speculate that these svRNAs may play a role in SARS-CoV-2 propagation by modulating post-transcriptional gene regulation, and that methods for antagonizing them may have therapeutic value.

Introduction

Small non-coding RNAs (sRNAs) play diverse roles in gene regulation and genome integrity. Ranging from ~20-30 nt, this functionally diverse class of RNAs plays important roles in the regulation of many biological processes [1,2]. Short interfering RNAs and microRNAs (siRNA and miRNA, respectively) function in post-transcriptional control of gene expression by regulating messenger RNA (mRNA) translation and stability [3,4]. By contrast, Piwi-associated RNAs (piRNAs) control transcriptional silencing of transposable elements and the elimination of entire regions of ciliate genomes [5,6].

sRNAs are derived from larger precursor transcripts. Although biogenesis pathways for the different types of sRNAs vary widely depending on type and species, processing of precursors typically involves cleavage by the endonucleases Drosha and Dicer, before they are loaded onto an Argonaute Family protein (either an AGO or a PIWI) to interact with downstream targets. However, there are a multitude of alternative mechanisms for sRNA biogenesis, some of which include Drosha/Dicer-independent mechanisms or even multiple Argonautes and Dicer-like proteins [7–10]. The level of complementarity of an sRNA to its target varies from as little as 8 nt in the 5' seed region (miRNAs) up to 100% complementarity of the full sRNA sequence (siRNAs and piRNAs), and often is a determining factor in how an sRNA interacts with its targets. It has also been suggested for human miRNAs that their abundance directly affects whether they repress translation by active mRNA degradation, or by interaction with the 3' UTR when miRNA levels are insufficient for widespread 3' UTR binding [11]. Taken as a whole, small RNA biogenesis and function is complex and exceptions to rules regarding roles and mechanisms are quite common.

Viruses such as SARS-CoV-2 remodel host gene expression programs. This occurs through a variety of mechanisms, including deregulation of post-transcriptional gene regulation. More recently, viral small non-coding RNAs have also been identified that may play important roles in adaptation and viral propagation [12–15] in infections with SARS-CoV-1 [12], Influenza [13], Hepatitis A [14] and EV71 [15]. During the COVID-19 pandemic, multiple groups also reported the existence of miRNA-like sRNAs derived from SARS-CoV-2, although the details of their biogenesis and specific function have yet to be elucidated [16–18]. Small viral RNAs (svRNAs) vary greatly in size and cellular abundance and have been implicated in a variety of different biological pathways related to infection and host immune evasion [19].

In this study, we characterize global changes in both host cell mRNA and miRNA expression using multiple sequencing-based methods. We also report the discovery of a diverse landscape of svRNAs produced by SARS-CoV-2. Sequence alignments suggest the intriguing hypothesis that SARS-CoV-2 small RNAs may directly regulate host transcripts, including microRNAs.

Results

SARS-CoV-2 infection induces global changes in host gene expression

To investigate how SARS-CoV-2 infection influences host gene expression, we analyzed the transcriptomes of both a human lung epithelial cell line (Calu-3) and African green monkey kidney cell line (Vero-E6) during the course of SARS-CoV-2 infection. Both cell lines were infected with SARS-CoV-2 at a multiplicity of infection (MOI) of 0.01. RNA extracts were made from harvested cell pellets of infected Calu-3 cells at 0 h, 12 h, 24 h, 48 h and 72 hours post-infection (hpi), and infected Vero-E6 cells at 4 h, 24 h, 48 h, 72 h, 120 h, 165 h and 216 hpi. RNA extracts of cell pellets from uninfected cells were also used as controls and

processed from all time points. Similar to previous studies [20], the majority of reads from infected cells mapped to the host genomes (60.3 - 100.0% for Calu-3 cells, 88.2 - 95.6% for Vero-E6 cells), with the remainder mapping to the SARS-CoV-2 genome (Figure 1A, Supplemental Table 1).

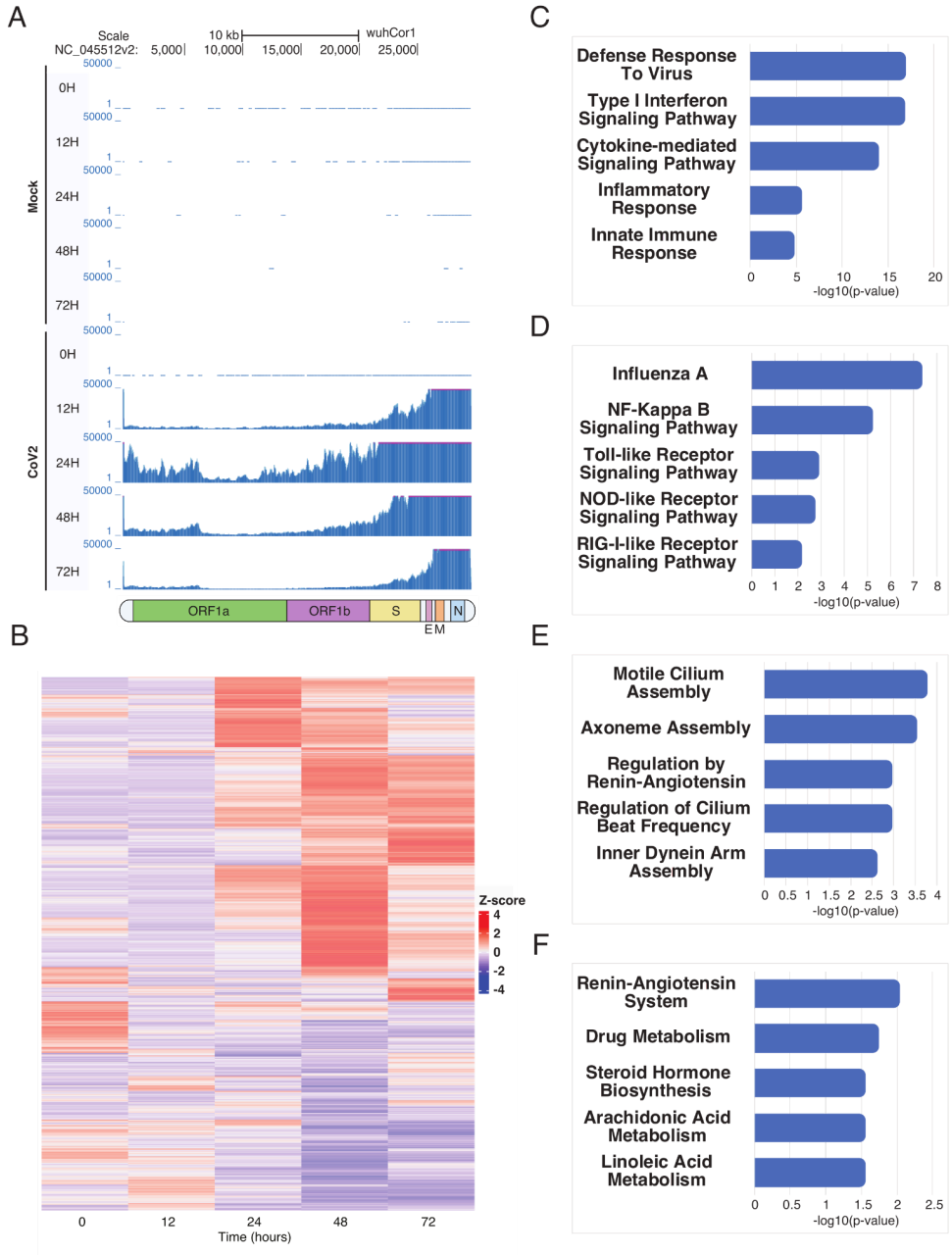


Figure 1. SARS-CoV-2 infection induces global changes in gene expression. A. UCSC Genome Browser screenshots showing Calu-3 sample-derived mRNA sequencing reads mapped to the SARS-CoV-2 genome for infected and control cells. Underneath the screenshots is a schematic representation of the SARS-CoV-2 genome depicting the regions of ORF1a/1b, the spike protein (S), the envelope protein (E), the membrane protein (M) and

the nucleocapsid protein (N). B. Heatmap representation of differentially expressed Calu-3 genes over the time course of infection. 4,593 differentially expressed genes are shown. Heatmaps were generated using hierarchical clustering. C. Enriched Gene Ontology Biological Process (GOBP) terms are shown for upregulated Calu-3 genes. Only a subset of significantly enriched GOBP terms are shown. D. Same as C, but with enriched KEGG Human Pathways. E. Enriched GOBP terms for downregulated Calu-3 genes. Only a subset of significantly enriched GOBP terms are shown. F. Same as E, but with enriched KEGG Human Pathways.

SARS-CoV-2 infection induces global changes in gene expression in both Calu-3 cells and Vero-E6 cells. DESeq2 [21] analysis revealed 4,593 and 1,040 significantly differentially expressed genes, respectively (Figure 1B, Supplemental Figure 1). Differentially expressed Calu-3 genes and Vero-E6 genes were used for downstream Gene Ontology Biological Process (GOBP) and KEGG Human Pathway enrichment analysis. Upregulated genes fell into enriched gene families related to innate immunity and the inflammatory response pathways, while downregulated genes tend to be involved in metabolic and other biosynthetic pathways (Figure 1C-F, Supplemental Table 2). For both Calu-3 and Vero-E6 cells, we observed significant overlap between GOBP terms for upregulated and downregulated genes throughout the time course. By contrast, KEGG pathway overlap was only observed for the upregulated gene sets (Supplemental Figure 2). Interestingly, for both cell types, we observed gene expression changes to be most significant from 24 hpi - 48 hpi, most likely reflective of the time required to elicit cellular response to viral infection. Overall, we found SARS-CoV-2 infection to cause global gene expression changes in both Calu-3 cells and Vero-E6 cells.

SARS-CoV-2 expresses a diverse landscape of small RNAs

We analyzed the small RNA transcriptome from mock and infected Calu-3 and Vero-E6 cells to uncover potential SARS-CoV-2-dependent gene regulatory mechanisms. Following infection, libraries from both cell lines contained reads mapping to the SARS-CoV-2 genome (Figure 2A, Supplemental Figure 3, Supplemental Table 1). We found that the distribution of svRNAs mapping to the SARS-CoV-2 genome was not uniform and that there were

“pile-ups" in particular regions indicating the RNAs identified are not likely to be degradation products of the full-length viral genomic or subgenomic RNA, but may be derived from specific loci (Figure 2A, Supplemental Figure 3). Using Piranha [22], we identified 67 svRNA loci from Calu-3 cells with a mean length of 22 nt and 97 svRNAs from Vero-E6 cells with a mean length of 25 nt (Figure 2B, Supplemental Figure 4, Supplemental Figure 5, Supplemental Table 3). 28 svRNA loci were shared between the two cell lines (Figure 2B). svRNA expression is dynamic throughout the time course with peak expression of the svRNAs between 24 – 48 hpi for both species (Figure 2C, Supplemental Figure 6). Calu-3 svRNAs tend to cluster into three distinct types of expression profiles. Two svRNA groups showed transient peak expression at 24 and 48 hpi respectively and then returned to base-level, while another group showed sustained expression from 24-48 hpi (Figure 2C). The dynamics of svRNA expression follows similar kinetics as changes in host mRNA and miRNA expression (Figure 1B, Figure 3A, Supplemental Figure 1).

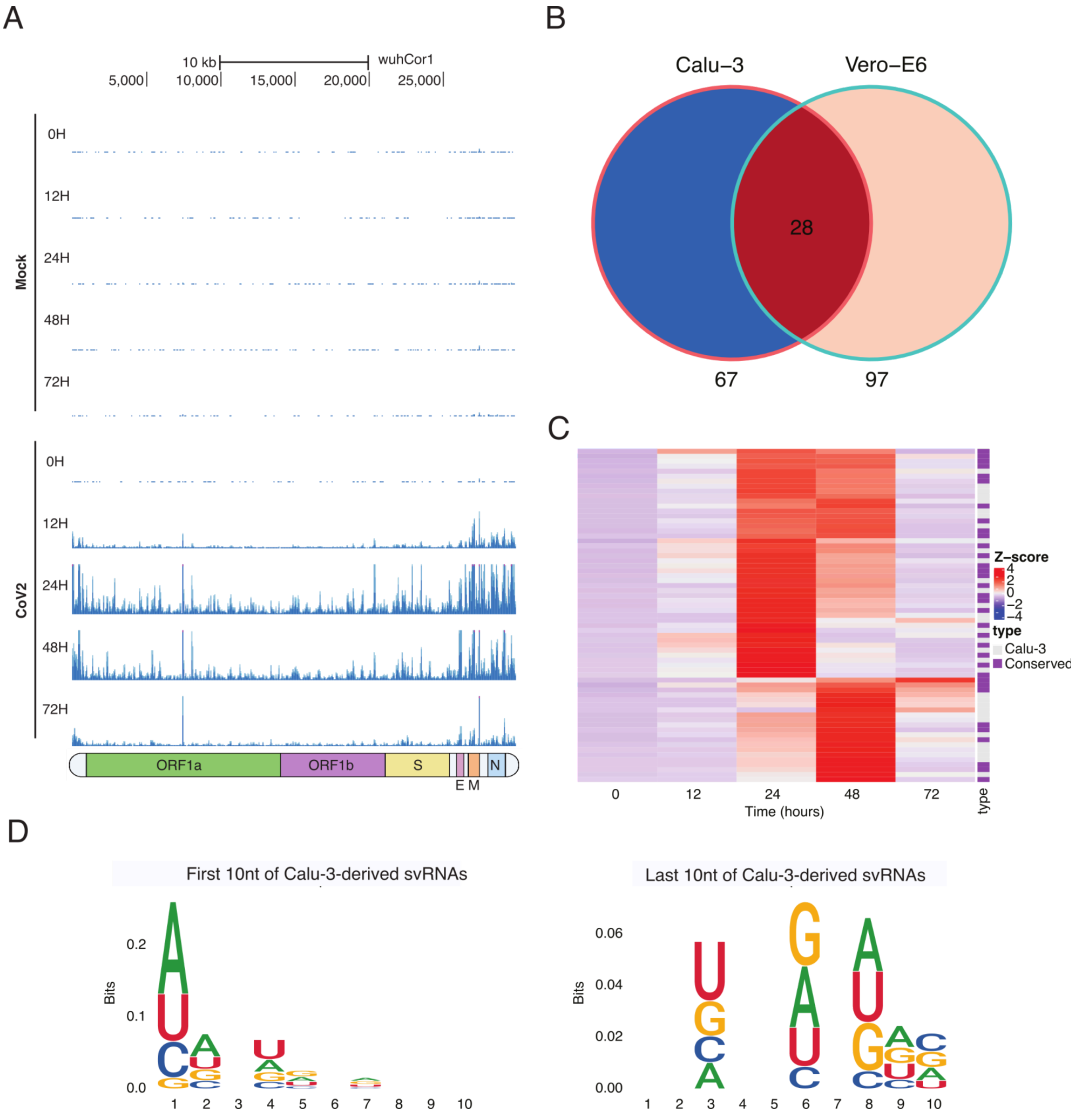


Figure 2. SARS-CoV-2 expresses a diverse landscape of small RNAs. A. UCSC Genome Browser screenshots showing Calu-3 sample-derived sRNA sequencing reads mapped to the SARS-CoV-2 genome for both infected and control cells. Reads from infected cells do not map uniformly to the genome but instead form "pile ups" in particular regions. Underneath the screenshots is a schematic representation of the SARS-CoV-2 genome depicting the regions of ORF1a/1b, the spike protein (S), the envelope protein (E), the membrane protein (M) and the nucleocapsid protein (N). B. Venn Diagram showing the number of svRNA loci identified in Calu-3 and Vero-E6 cells, with conserved loci indicated. C. Heatmap representation of Calu-3 cell-derived svRNA expression. Heatmaps were generated using hierarchical clustering. To the right of the heatmap is an annotation column indicating if the svRNA is species-specific or conserved between Calu-3 and Vero-E6 cells. D. Sequence

logos for the first ten and last ten nucleotides of the svRNAs identified in Calu-3 cells. svRNAs lack a 5'-U signature typical of many canonical miRNAs.

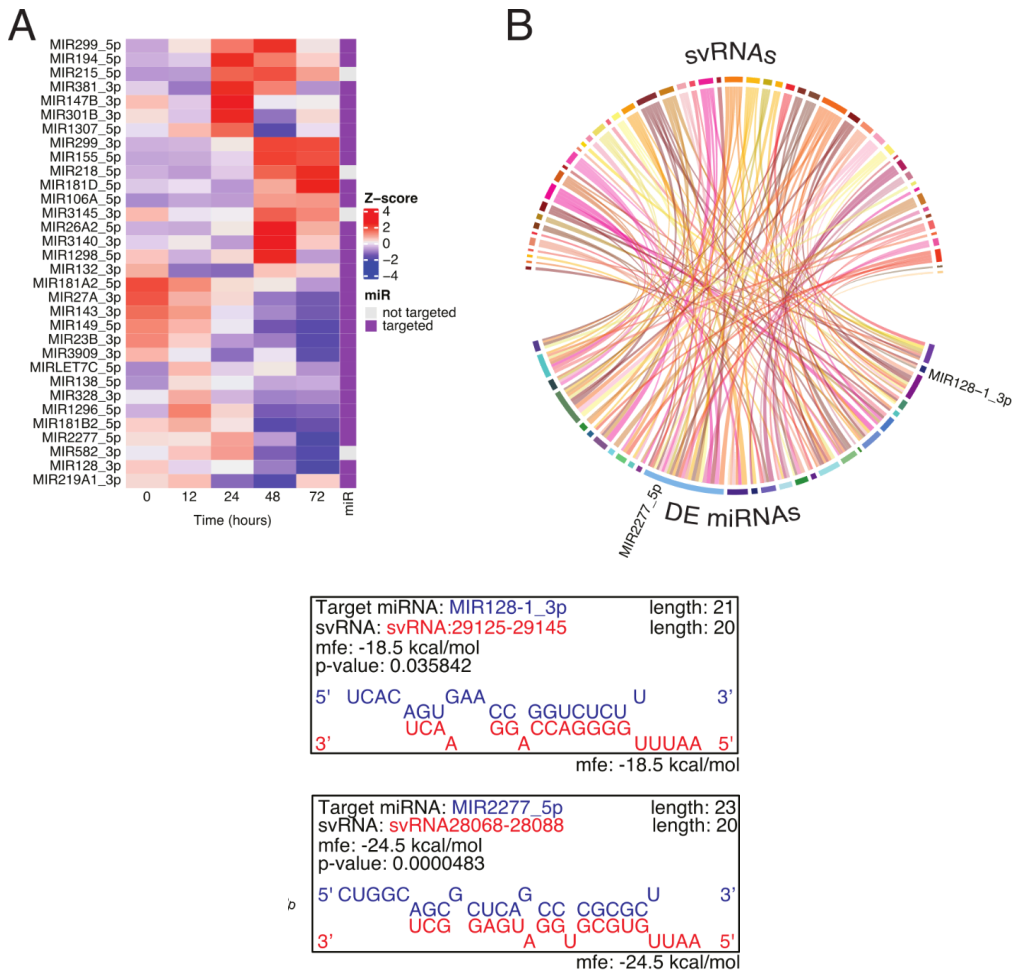


Figure 3. SARS-CoV-2 svRNAs may target host miRNAs. A. Heatmap representation of the 28 differentially expressed Calu-3 microRNAs over the time course of infection. Heatmaps were generated using hierarchical clustering. To the right of the heatmap is an annotation column indicating if the miRNA is a predicted target of an svRNA. B. Circos plot representation of predicted interactions between svRNAs and DE (human) miRNAs using RNAhybrid without seed-forcing. The ribbons are colored according to the svRNA that is predicted to interact with the miRNA to which it connects. There are many examples of svRNAs targeting more than one miRNA and also examples of miRNAs being targeted by more than one svRNA. DE miRNAs mentioned in the Results section are labeled. Shown below are two examples of RNAhybrid predictions for svRNAs and DE miRNAs pictured above.

To illuminate potential biogenesis pathways for SARS-CoV-2 svRNAs, we aligned the 5' and 3' ends of svRNAs searching for nucleotide biases. By contrast to miRNAs, the svRNAs did not possess sequence signatures with the typical 5'-U bias, suggesting a Dicer-independent processing pathway (Figure 2D, Supplemental Figure 7). Additionally, we found that only 10 of the 67 Calu-3 cell-derived svRNAs and 12 of the 97 Vero-E6 cell-derived svRNAs were predicted to have an upstream or downstream complementary sequence within 44 nt that may form a hairpin [23], often required for canonical miRNA processing. (Supplemental Table 4).

SARS-CoV-2 alters host microRNA expression

We hypothesized that global changes in gene expression could be mediated through changes in host sRNA expression, specifically miRNAs. We analyzed the host small RNA transcriptomes of infected and control Calu-3 cells (Figure 3A, Supplemental Figure 8-10, Supplemental Table 1) and discovered that 28 miRNAs are differentially expressed throughout the infection time course (Figure 3A, Supplemental Table 5). Consistent with our mRNA sequencing data, we observed differential expression of miRNAs at 24 hpi with the most significant changes occurring at 48 hpi (Figure 1B).

SARS-CoV-2 svRNAs may hybridize with differentially expressed mRNAs and miRNAs

To investigate the potential function of svRNAs we used RNAHybrid [24,25] to identify targets within differentially expressed Calu-3 mRNAs and miRNAs. Of the svRNAs identified in Calu-3 cells, 59 of 67 have predicted mRNA targets within the 3'UTRs of differentially expressed (DE) genes (Tables 1 and 2, Supplemental Table 3). RNAhybrid predictions identified svRNAs that can also form stable duplexes with miRNAs, 52 of which target differentially expressed miRNAs in our data set (Figure 3B, Table 3, Supplemental Figure 11, Supplemental Table 3). A larger portion of svRNA-target interactions were

predicted to occur with downregulated miRNAs than with downregulated mRNA targets throughout the time course (Supplemental Figure 10). Interestingly, we found numerous examples of host miRNAs that can pair with multiple svRNAs (Figure 3B, Figure 4, Supplemental Figure 11). In addition, we also observed examples of svRNAs that can potentially hybridize to a broad array of miRNAs (Figure 3B, Figure 4, Table 3, Supplemental Figure 11). Because the Vero-E6 genome has not been extensively annotated to include miRNA sequences, we were unable to perform RNAhybrid predictions against Vero-E6 miRNAs, but found that of the 61 Calu-3-derived svRNAs capable of targeting host miRNAs, 27 had conserved loci with Vero-E6-derived svRNAs.

svRNA	Counts	# mRNA targets	Top Enriched GO Term for Targets
svRNA2875-2897	165	39	Negative Regulation of CD4-positive, Alpha-beta T Cell Proliferation (GO:2000562)
svRNA26927-26947	134	18	Negative Regulation of CD4-positive, Alpha-beta T Cell Proliferation (GO:2000562)
svRNA28260-28289	89	1	Gamma-aminobutyric Acid Metabolic Process (GO:0009448)
svRNA27059-27089	88	16	Regulation of Protein Import into Nucleus (GO:0042306)
svRNA27107-27128	80	11	Heart Field Specification (GO:0003128)
svRNA26820-26840	77	2	Mesoderm Morphogenesis (GO:0048332)
svRNA5572-5597	64	4	Regulation of Oxidoreductase Activity (GO:0051341)
svRNA37-63	57	2	Ribosomal Large Subunit Assembly (GO:0000027)
svRNA18943-18965	47	1	N/A
svRNA29039-29064	45	5	Extracellular Structure Organization (GO:0043062)

Table 1. List of the ten most abundant Calu-3-derived svRNAs with DE mRNA targets. Table listing the ten most abundant svRNAs with DE mRNA targets, normalized counts, number of differentially expressed mRNA targets and the top enriched GO term for targets. Counts were normalized to the mean read depth of all sample libraries.

Top Enriched GO Terms for Downregulated mRNA Targets of svRNAs	Top Enriched GO Terms for Upregulated mRNA Targets of svRNAs
Quinone Catabolic Process (GO:1901662)	Cellular Response to Cytokine Stimulus (GO:0071345)
Leukotriene B4 Metabolic Process (GO:0036102)	Cytokine-mediated Signaling Pathway (GO:0019221)
Menaquinone Metabolic Process (GO:0009233)	Regulation of Interferon-gamma Production (GO:0032649)
Vitamin K Metabolic Process (GO:0042373)	Cellular Response to Interferon-gamma (GO:0071346)
Fat-soluble Vitamin Metabolic Process (GO:0006775)	Negative Regulation of Natural Killer Cell Mediated Cytotoxicity (GO:0045953)
Organic Hydroxy Compound Catabolic Process (GO:1901616)	Regulation of Apoptotic Cell Clearance (GO:2000425)
Fat-soluble Vitamin Catabolic Process (GO:0042363)	Positive Regulation of Apoptotic Cell Clearance (GO:2000427)
Axoneme Assembly (GO:0035082)	interferon-gamma-mediated Signaling Pathway (GO:0060333)
Diterpenoid Metabolic Process (GO:0016101)	Regulation of Immune Response (GO:0050776)
Phosphate Ion Transport (GO:0006817)	Negative Regulation of Natural Killer Cell Mediated Immunity (GO:0002716)

Table 2. List of the top enriched GO terms for DE mRNA targets of Calu-3-derived svRNAs. Table listing the top ten significantly enriched GO terms for all upregulated and all downregulated mRNA targets of Calu-3-derived svRNAs in our data set.

svRNA	Counts	# miRNA targets	miRNA
svRNA26741-26761	165	4	**MIR128_3p**, MIR1298_5p, **MIR1307_5p**, **MIR2277_5p**
svRNA28573-28595	134	1	**MIR2277_5p**
svRNA2868-2897	88	3	**MIR106A_5p**, **MIR2277_5p**, MIR381_3p
svRNA28704-28725	80	2	**MIR138_5p**, **MIR2277_5p**
svRNA28553-28573	77	4	**MIR132_3p**, **MIR2277_5p**, **MIR23B_3p**, MIR299_5p
svRNA18943-18965	57	1	**MIR149_5p**
svRNA27578-27602	47	2	MIR26A2_5p, MIR381_3p
svRNA524-546	45	1	**MIR2277_5p**
svRNA5423-5445	45	3	MIR1298_5p, **MIR181B2_5p, MIR3909_3p**
svRNA28068-28088	41	1	**MIR132_3p**

Table 3. List of the ten most abundant Calu-3-derived svRNAs with DE miRNA targets. Table listing the ten most abundant svRNAs with DE miRNA targets, normalized counts, number of differentially expressed miRNA targets and miRNA target names. miRNAs in bold are downregulated in our data set. Counts were normalized to the mean read depth of all sample libraries.

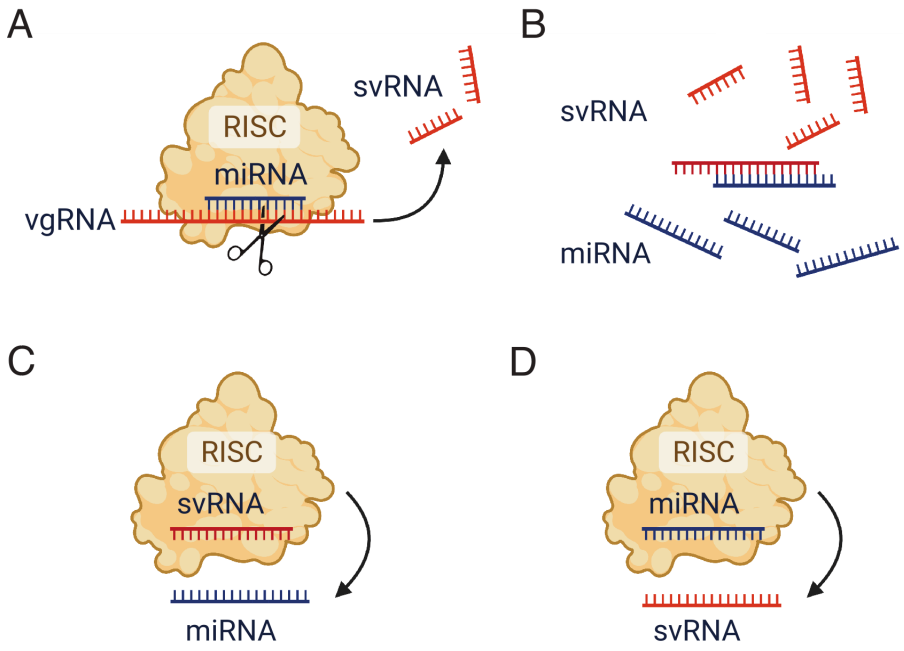


Figure 4. Proposed svRNA-miRNA interactions upon SARS-CoV-2 infection. The illustrated model is briefly described in the Discussion section. svRNAs may interact with miRNAs in multiple ways, none of which are mutually exclusive. A. A RISC-miRNA complex targets complementary sequences in the viral genomic RNA. svRNAs are produced by cleavage and go on to perform downstream functions, such as interacting with host mRNAs and miRNAs. B. There is direct hybridization of svRNAs and miRNAs without association with RISC, leading to sequestration of miRNAs, preventing them from interacting with endogenous targets. C. A RISC-svRNA complex targets miRNAs and prevents them from interacting with endogenous targets. D. A RISC-miRNA complex targets svRNAs, preventing the complex from interacting with endogenous targets. Created with BioRender.com.

Discussion

We analyzed the transcriptomes of two primate cell lines during SARS-CoV-2 infection and observed global changes in gene expression. Consistent with previous work, we found enrichment for genes involved in the RIG-I pathway, Toll-like Receptor pathway (TLR) signaling pathways, NF-kappa Beta pathway, and type 1 interferon (IFN) signaling [16,26–30] (Supplemental Figure 12). We also identified more than a hundred svRNAs capable of interacting with host transcripts, including miRNAs. svRNA are expressed from specific loci, within ORF7a, and the spike (S) and nucleocapsid (N) genes. In addition to

having predicted target sites in differentially expressed host transcript 3'-UTRs like canonical miRNAs, 52 of the 67 Calu-3 sample-derived svRNAs are also predicted to form stable duplexes with differentially expressed miRNAs, many of which have been implicated in other types of viral respiratory infections (VRIs) [31]. For example, miR-128, miR-2277 and miR-155 are known to be associated with rhinovirus infection, Middle Eastern Respiratory Syndrome (MERS) and SARS-CoV-2, respectively. We see many examples of svRNAs with predicted target sites in more than one differentially expressed miRNA, with 38 of the 67 having two or more predicted targets. svRNAs are most highly expressed at 24 h and 48 hpi, which corresponds to the time points at which we see the most significant changes in host gene and miRNA expression. This expression dynamic suggests an interplay between svRNAs and host miRNAs, in which svRNAs have the capability to interact with host miRNAs, therefore affecting all downstream target mRNA expression.

Here we describe an expanded landscape of SARS-CoV-2 derived small RNAs including all four svRNAs previously reported by the Cheng, Meng and Pawlica groups [16–18] (Supplemental Figure 13). In agreement with what the Meng and Cheng groups observed, we identified dozens of svRNAs derived from the nucleocapsid (N) gene, many that overlap the sequences reported in those studies, albeit with varying starting coordinates and lengths [16,17] (Supplemental Figure 13). In addition to recapitulating the existence of svRNAs derived from the N gene, we identified an svRNA (svRNA27406-27428) derived from ORF7a that almost perfectly overlaps with the sequence reported by Pawlica and colleagues [18] (Supplemental Figure 12). We reason that our experiment has increased sensitivity because we use a lower MOI (0.01) than previous studies (0.05-5), and performed an extended time course experiment [16,18].

miRNAs are typically processed via a Dicer-mediated biogenesis pathway. Our data suggest that many of the svRNAs are generated instead via a Dicer-independent processing mechanism. Lack of a 5'-U bias in the sequence analyses for svRNAs indicates that biogenesis may occur in a non-canonical Dicer-independent manner. Other groups have also reported svRNAs derived from SARS viruses lacking a 5'-U sequence signature and it has been suggested that these small RNAs may utilize the slicer activity of Ago2 for biogenesis rather than the traditional miRNA/siRNA processing machinery, similar to what has been observed for the well-studied human miR-451[10,12,17]. Morales and colleagues observed svRNA expression following knockdown of the RNase III nucleases Drosha and Dicer [12], suggesting these canonical enzymes are not necessary for svRNA biogenesis. We also identified svRNAs derived from ORF7a and the N gene that overlap with svRNAs reported by the Meng, Cheng and Pawlica groups, but with varied starting coordinates and/or sizes. This may suggest that isomiRs exist for svRNAs, potentially caused by alternative processing events such as 5'/3' trimming variants due to imprecise cleavage of precursors [32]. Further studies on svRNAs derived from SARS-CoV-2 are necessary to elucidate key players in their biogenesis and to validate their potential to interact with host transcripts to affect host gene expression.

Based on our observations, we propose a functional role for svRNAs in which deregulation of the host mRNA expression program may be a direct effect of svRNAs sequestering or "sponging" host miRNAs (Figure 4). There are many examples of known miRNA sponges in the literature, such as *HSUR 1* and *HSUR 2* from *Herpesvirus saimiri [33]*, and it is entirely possible that the svRNAs presented here may function similarly [34–36]. An efficient way to post-transcriptionally alter a gene network or biological system would be to sequester/act on regulators of the network itself, rather than individual genes separately. As canonical human miRNAs often have hundreds of predicted target genes each, production of svRNAs that

interfere with this regulation may confer an evolutionary advantage for SARS-CoV-2 and other viruses by modulating host gene expression on a global scale, increasing the likelihood of viral propagation. It has been previously suggested that viral RNA derived from SARS-CoV-2 may act as a host miRNA sponge to aid in avoiding the host immune response and our findings add more credence to this hypothesis [37]. Based on RNAhybrid seed-target predictions, 22 of 33 differentially expressed host miRNAs are capable of targeting svRNAs, 11 of which are upregulated during the time course. It is possible that in addition to svRNAs sponging host miRNAs, they also directly target miRNAs with or without RISC, preventing the host miRNAs from interacting with endogenous targets (Figure 4). These potential svRNA-miRNA interactions are not mutually exclusive and may occur simultaneously, ultimately leading to the same outcome: global changes in host gene expression. Direct interactions between svRNAs and their predicted targets must still be validated experimentally to assess the full extent to which these interactions occur during viral infection and also to tease apart the actual targeting mechanism itself. In the future, these interactions could be easily antagonized or blocked using locked nucleic acids (LNAs) or other methods and potentially used as targets for the therapeutic treatment of patients infected with the virus.

Data and Code Availability

All RNA-sequencing data can be accessed at the National Center for Biotechnology Institute (NCBI) Gene Expression Omnibus (GEO) database: GSE197521. No new code was developed for this study. ***Supplementary materials are

Acknowledgments

This work was supported by the National Institutes of Health (NIH) (grant no. R35 GM130361), the Tong Tsung and Wei Fong Chao Foundation (grant no. GT007457) and the Chau Hoi Sheun Foundation (grant no. GT007457).

Author Contributions

Conceptualization, Z.T.N., N.M., and J.R.S.; Methodology, Z.T.N., N.M., J.R.S.; Software, A.J.R. and S.K.; Formal Analysis, A.J.R. and S.K.; Investigation, Z.T.N. and L.V.C.; Resources, N.M. and J.R.S.; Data Curation, A.J.R. and S.K.; Writing – Original Draft, Z.T.N., A.J.R.; Writing – Review & Editing, Z.T.N., A.J.R., L.V.C., S.K., W.I.L., N.M., and J.R.S.; Visualization, Z.T.N. and A.J.R.; Supervision, W.I.L., N.M. and J.R.S.; Funding Acquisition, N.M. and J.R.S.

Declaration of Interest

The authors declare no competing interests.

STAR Methods

Cell Culture and Viruses

African green monkey kidney (Vero-E6) cells and human lung epithelial (Calu-3) cells were obtained from the ATCC (Manassas, VA). The cells were cultured in Dulbecco's modified Eagle's medium (ThermoFisher Scientific, Waltham, MA, USA) containing 1% heat inactivated fetal bovine serum (ThermoFisher Scientific), 100 U/mL penicillin, and 100 µg/mL streptomycin. The cells were incubated in 95% air and 5% CO_2 at 37°C.

Virus Infection and total RNA extraction

Vero-E6 and Calu-3 cells (2×10^5 cells/well in 6-well plates) were cultured in DMEM containing 1% FBS at 37°C in a CO_2 incubator overnight. The cells were washed once with phosphate-buffered saline (PBS), prototypic SARS-CoV-2 Washington strain (USA/WA1/2020) in growth media at MOI 0.01 was added into each desired well, and the plates were incubated for 1.5 h at 37° C in a CO_2 incubator. After incubation, 3 mL of DMEM containing 1% FBS was added to each well and the plates were incubated at 37° C in a CO_2 incubator. Infected Calu-3 cells were harvested at 0 h, 12 h, 24 h, 48 h and 72 hpi (in triplicate) and Vero-E6 cells were harvested at 4 h, 24 h, 48 h, 72h, 120 h, 168 h and 216 hpi (in duplicate).

For harvesting, cell culture supernatants were collected and mixed with Trizol (Fisher Scientific) at a ratio of 1:1. Cells were detached using Trypsin-EDTA 0.25% (ThermoFIsher Scientific), followed by centrifugation at 1,000 rpm for 5 min at 4°C to form a cell pellet which in turn was dissolved in 250 ul of Trizol. Both cell supernatant in Trizol and cell pellet in Trizol were stored at -80'C until further processing. SARS-CoV-2 amplification and cell culture procedures were performed according to biosafety level 3 (BSL-3) conditions.

Total RNA was extracted using the TRI Reagent protocol for isolation of RNA with the following modification: one additional RNA ethanol wash step was included. After the total RNA was solubilized in ddH_20, one overnight ethanol precipitation step was included for further purification of the total RNA.

Illumina sequencing of sRNA libraries

Total RNA was isolated from cell culture pellets as described above. Total RNA from infected Calu-3 cells, infected Vero-E6 cells and uninfected controls were used for sRNA library preparation. 1 ug of total RNA for each sample was used for sRNA library preparation

using the NEXTFLEX small RNA-Seq Kit v3 following the manufacturer's protocol (Perkin Elmer Applied Genomics). Adapter dilution was not performed and 17 cycles of PCR amplification were used before using the gel-free size selection and cleanup protocol. Pooled sRNA sequencing libraries were sequenced on an Illumina HiSeq 4000 at the UC Davis Sequencing Core Facility, generating 100 bp single-end reads.

Illumina sequencing of mRNA libraries

Total RNA was isolated for cell culture pellets as described above. Total RNA from infected Calu-3 cells, infected Vero-E6 cells and uninfected controls were used for mRNA library preparation. 2 ug of total RNA for each sample was used for mRNA library preparation using the NEXTFLEX Rapid Directional RNA-Seq Kit 2.0 following the manufacturer's protocol (Perkin Elmer Applied Genomics). Before library preparation, total RNA samples were subjected to Poly(A) selection and purification using the NEXTFLEX Poly(A) Beads Kit 2.0 following the manufacturer's protocol (Perkin Elmer Applied Genomics). Based on sample RNA integrity numbers (RINS) and total RNA input, 9 minutes fragmentation time was used, adapter concentration was diluted by half and 9 cycles of PCR amplification were used before bead clean up and elution. For Calu-3 cell samples, pooled mRNA sequencing libraries were sequenced on an Illumina NovaSeq S4 at the UC Davis Sequencing Core Facility, generating 150 bp paired-end reads. For Vero-E6 samples, pooled mRNA sequencing libraries were sequenced on an Illumina HiSeq 4000, generating 100 bp single-end reads.

Analyses of Calu-3 mRNA sequencing data

The full-length paired-end reads were mapped with Bowtie 2 [38] against the set of human repeat-masker elements in the hg38 assembly, as well as a constant poly-A sequence. Reads that were mapped to these elements were removed from further processing. The

repeat-filtered full-length reads were mapped with STAR 2.5.3a [39], using the "--alignEndsType EndToEnd" option, to a single joint target consisting of the hg38 genome assembly plus the wuhCor1 genome. In cases of multiply-mapped reads, only the best mapping was retained. The STAR mapped bam files were divided into separate hg38 and wuhCor1 files for further analysis.

Gene-by-gene coverage was extracted from the hg38 mappings that overlapped any potential exon for each gene in a gene model of hg38. The total coverage for each gene was divided by the total paired-end read length of each mapping to extract read counts as input to DESeq2 [21]. DESeq2 was run to compare all replicates of conditions as follows:

1. At each time point, the control vs the infected samples
2. For the controls, comparisons between the time points
3. For the infected samples, comparisons between the time points.

Analyses of Vero-E6 mRNA sequencing data

The single-end reads were trimmed at the 3' end to leave 50bp for mapping. The trimmed reads were mapped with Bowtie 2 [38] against a set of repeat elements in the Vero-E6 genome. The trimmed, repeat-filtered and wuhCor1-filtered, single-end reads were mapped with STAR 2.5.3a, using the "--alignEndsType EndToEnd" option, to chlSab2 genome assembly. In cases of multiply-mapped reads, only the best mapping was retained. Potential PCR duplicates (single-end reads mapped to the identical coordinates) were removed from the chlSab2 mappings.

Gene-by-gene coverage was extracted from the chlSab2 mappings that overlapped any potential exon for each gene in a gene model of chlSab2. The total coverage for each gene

was divided by the single-end read length of each mapping to extract read counts as input to DESeq2.

DESeq2 was run to compare all replicates of conditions:

1. At each time point, the control vs the infected samples
2. For the controls, comparisons between the time points
3. For the infected samples, comparisons between the time points.

Gene ontology (GO) and KEGG human pathway analyses

GO process annotations and KEGG human pathway annotations were retrieved using the web tool Enrichr [40–43]. The p-values of enrichment of GO process and KEGG pathways were determined using Enrichr on differentially expressed upregulated and downregulated genes from both Calu-3 and Vero-E6 data sets.

svRNA discovery from sRNA-seq data

Potential svRNAs were identified from sRNA-seq reads using a stepwise bioinformatic approach which incorporates established tools and bespoke analysis methods. Prerequisite packages include: STAR v2.7.8a [39], BEDOPS v2.4.40 [44], bedtools v2.28.0 [45], Piranha v1.2.1 [22], Bowtie 2 v2.4.4 [38] and RNAhybrid v2.1.2 [25]. The svRNA-discovery pipeline entails the following steps:

1. sRNA-seq libraries were first mapped against their respective host genomes (GRCh38 for human, chlSab2 for African Green Monkey) and against the SARS-CoV-2 genome (wuhCor1) using STAR [46].
2. Mapping files from the infected samples from Calu-3 and Vero-E6 were subsetted for sequences between 20-30 nt in length that mapped to the SARS-CoV-2 genome.

Piranha peaks were called for pooled reads from infected samples for each time point in each of the two cell types using bins of size 50 and a background threshold of 0.95.

3. Piranha peaks were pooled for each cell type, and unique reads mapping to them were aligned to their respective host genomes using Bowtie 2 to remove any potential multi-mapping or host-mapped reads from downstream analysis.
4. Of the remaining sequences, those that met our count-based cutoffs were filtered for sequences whose counts across infected replicates and time points were the highest among overlapping sequences. Some overlapping sequences whose counts were especially relatively high were retained. The count-based cutoffs for a given sequence were as follows:
    a. It must be present in at least 5 infected samples for Calu-3 and at least 4 for Vero-E6 (due to the difference in replicates).
    b. Its total raw counts from infected samples must be at least 20.
    c. If present in any control samples, it must not be present in all control samples.
    d. If present in any control samples, its total raw counts from control samples must be less than or equal to 10% of the total raw counts from infected samples.
5. Duplexes between host transcripts (all expressed mRNAs and miRNAs) and SARS-CoV-2-derived sequences from Step 4 were then predicted using RNAhybrid [24,25], which is a tool that is typically used to predict the hybridization, with the smallest minimum free energy, of a short RNA molecule with a longer one. SARS-CoV-2-derived sequences with at least 1 statistically significant (p-value < 0.05) predicted duplex with a host transcript were then considered potential SARS-CoV-2-derived svRNAs. The resultant target predictions were then subsetted

for those involving host targets that were significantly differentially expressed (p-value < 0.05, $\log_2$ fold change ≥ 1.0).

6. Potential svRNAs from both cell types with overlaps to each other (across cell types) greater than or equal to 60% were categorized as svRNAs derived from "conserved loci". We define conserved loci as loci in the SARS-CoV-2 genome from which potential svRNAs were detected in both the Calu-3 and Vero-E6 experiments. Additionally, overlapping svRNAs within cell type (i.e. shifted 1-2nt down the SARS-CoV-2 genome sequence) that were retained due to their high relative counts were denoted as potential viral isomiRs – in this case being svRNAs from the same approximate locus but with slightly different ends owed to imprecise precursor cleavage.
7. Lastly, to investigate the potential of identified svRNAs to be products of hairpin precursors, we used RNAfold [23] to predict the structure of each svRNA with the addition of 30-45 nt upstream and downstream separately. In order to call potential hairpin precursors for svRNAs, the structure predictions needed to have: only a single loop – with no shared nucleotides with the mature svRNA sequence, at least 14 base pairings in the stem, and a maximum minimum free energy (MFE) of -4.3kcal/mol [47]. Hairpins that passed these filters were visualized using Forna [48].

Chapter 3: Methodological Advancements in the Study of Alternative Splicing

3.1 Chapter Introduction

This chapter describes a project in which we sought to design an alternative splicing (AS) analysis tool, *junctionCounts*, that resolves a number of shortcomings of existing approaches. The two primary gaps we aimed to address were the limited event type definitions individual tools are capable of characterizing and the lack of utilities for predicting transcript-level functional consequences of splicing events (i.e. induction of NMD). Importantly, this tool was initially designed by Dr. Andrew Wallace and was a substantial portion of his doctoral dissertation. Nonetheless, *junctionCounts* was left unfinished despite several of our collaborators having since published work that made use of it.

Most crucially, *junctionCounts* did not include utilities for handling replicates or for statistical testing of AS events across conditions. Additionally, it was not comprehensively evaluated against existing tools. To finish this work, I benchmarked *junctionCounts* against two well-established AS analysis tools on RT-PCR and simulated datasets. *junctionCounts* performed competitively, especially in terms of quantification accuracy, and it notably captured complex event types beyond the scope of most existing tools. I further designed a bridge between *junctionCounts* and DEXSeq [49] to add statistical testing with flexible replicate-handling to the event-level analysis pipeline. To test the accuracy of NMD predictions generated by partner utilities, *cdsInsertion* and *findSwitchEvents*, I applied them to a published dataset profiling NMD in emetine-treated HEK 293 T cells [50] and found that nearly 75% of predicted NMD substrates followed the expected expression changes.

Beyond evaluating *junctionCounts* and its partner utilities, I next implemented them on a primate neuronal differentiation RNA-seq dataset [51] to profile conserved and

species-specific temporal splicing dynamics during neuronal differentiation across human, chimpanzee, orangutan and rhesus macaque. The findings described in detail in the following manuscript demonstrate the power of *junctionCounts* to generate high-quality results and substantiate its status as a fully-realized and user-friendly tool for researchers studying AS. I executed all bioinformatic analyses, data visualization and writing and editing of the manuscript. This manuscript was submitted to *Nucleic Acids Research Genomics and Bioinformatics* on February 23rd, 2024.

3.2 junctionCounts: comprehensive alternative splicing analysis and prediction of isoform-level impacts to the coding sequence

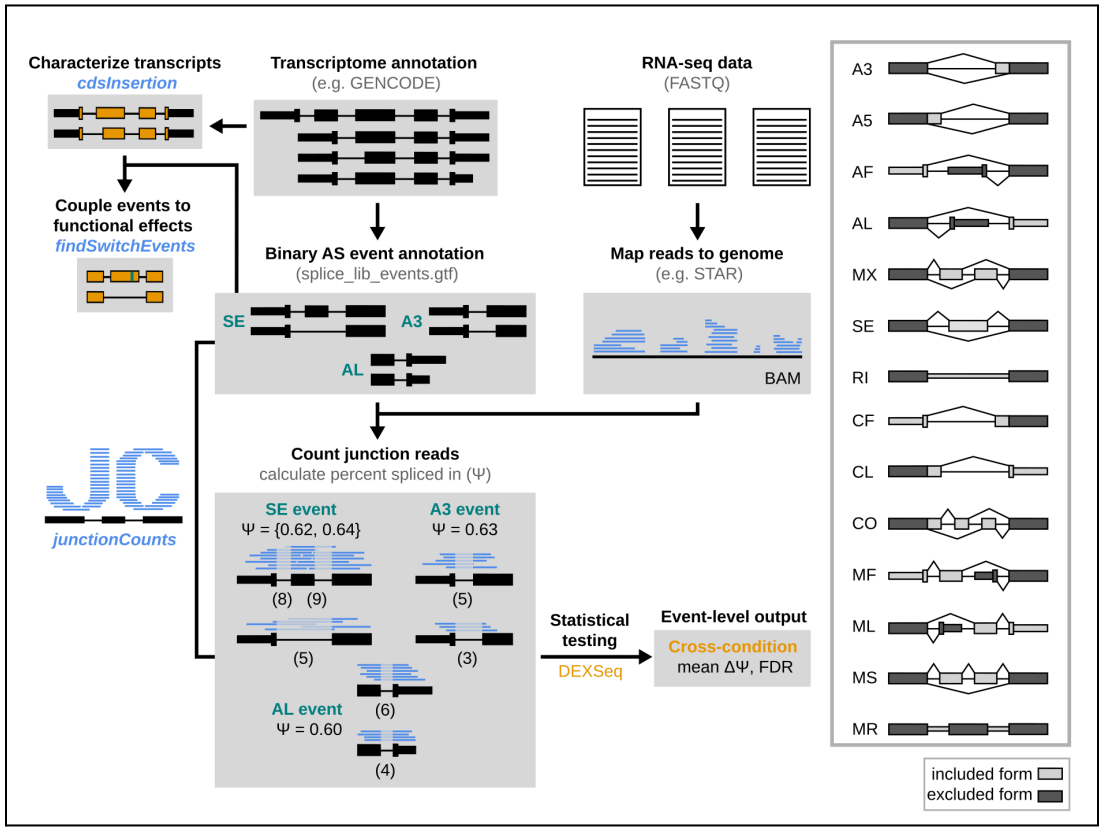
Alex Ritter[1†], Andrew Wallace[2†], Neda Ronaghi[2] and Jeremy R Sanford[2]

[†] These authors contributed equally to this work
[1] Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, 95064, USA
[2] Department of Molecular, Cell and Developmental Biology, University of California Santa Cruz, Santa Cruz, CA, 95064, USA

* To whom correspondence should be addressed. Tel: 831-459-1822; Fax: 831-459-3139
Email: jsanfor2@ucsc.edu

Present Address: Jeremy Sanford, Department of Molecular, Cell and Developmental Biology, University of California Santa Cruz, Santa Cruz, CA, 95064, USA

GRAPHICAL ABSTRACT



junctionCounts is an alternative splicing analysis tool that identifies both simple and complex splicing events from a gene annotation and then measures their percent spliced-in from mapped RNA-seq junction reads.

ABSTRACT

Alternative splicing (AS) is emerging as an important regulatory process for complex biological processes. Transcriptomic studies therefore commonly involve the identification and quantification of alternative processing events, but the need for predicting the functional consequences of changes to the relative inclusion of alternative events remains largely unaddressed. Many tools exist for the former task, albeit each constrained to its own event type definitions. Few tools exist for the latter task; each with significant limitations. To address these issues we developed junctionCounts, which captures both simple and complex pairwise AS events and quantifies them with straightforward exon-exon and exon-intron junction reads in RNA-seq data, performing competitively among similar tools in terms of sensitivity, false discovery rate and quantification accuracy. Its partner utility, cdsInsertion, identifies transcript coding sequence (CDS) information via *in silico* translation from annotated start codons, including the presence of premature termination codons. Finally, findSwitchEvents connects AS events with CDS information to predict the impact of individual events to the isoform-level CDS. We used junctionCounts to characterize splicing dynamics and NMD regulation during neuronal differentiation across four primates, demonstrating junctionCounts' capacity to robustly characterize AS in a variety of organisms and to predict its effect on mRNA isoform fate.

INTRODUCTION

Alternative splicing (AS) generates a diverse array of mRNA isoforms from a single locus. The consequences of this process can have manifold effects on gene expression by altering mRNA half-life [52], intracellular localization [53], translation efficiency [54], and most obviously, by producing different protein isoforms [55]. AS plays critical roles in a variety of biological processes including disease pathology [56], cancer [57], and cellular differentiation [58]. In cellular homeostasis, AS is tightly regulated to control the precise expression of diverse mRNA isoforms, allowing cells to adapt to changing conditions and to achieve different states of activation in immune cells, for example [59]. Regulatory elements, including *cis*-acting splicing enhancers and silencers, as well as *trans*-acting splicing factors, coordinate the inclusion or exclusion of alternative exons or splice sites during transcription [60].

Dysregulation of AS can contribute to the production of aberrant protein isoforms, impacting critical cellular functions. In cancer, this can result in the generation of oncogenic isoforms, altered signaling pathways, and evasion of regulatory mechanisms [61]. Often, mutations in splicing factors and in *cis*-elements underlie oncogenic AS dysregulation [62]. During cellular differentiation from pluripotent stem cells, AS orchestrates the precise control of gene expression, directing the development of specialized cell types with distinct functions [63]. This process is intricately involved in shaping the cellular landscape, driving fate decisions, and maintaining tissue homeostasis [64]. The important impact of AS in disease and cellular differentiation underscores its significance as a regulatory force in biological processes and highlights its potential as a therapeutic target in pathological conditions.

It is thus important, in any eukaryotic cellular context, to understand gene expression at the isoform level. The complex nature of AS, however, creates numerous obstacles to its accurate

study. mRNA isoforms can, and have canonically been, characterized in terms of binary events that either include or exclude an alternative feature (exon, intron or splice site). This mode of characterization faces the challenge of differentiating between complex and overlapping event features, and also relies on comprehensively annotated gene structures. The latter problem is highlighted by the lack of records for transcripts clearly supported by mapped reads in available references. To address this, reference-guided or *de novo* transcriptome assembly has become a widely used step in RNA-seq analysis. In this process, transcript structure is predicted from the data with or without the use of a reference gene annotation as a template [65]. This process allows analysts to consider potential novel transcripts and novel alternative event features that may be important to the biological phenomena under study.

Upon establishing comprehensive gene models, it subsequently becomes important to understand how AS configures the coding and noncoding regions of resultant mRNA isoforms. Unfortunately, common tools for transcriptome assembly [66,67] are unable to provide information on the presence and nature of open reading frames (ORFs) that may be contained within predicted novel transcripts. One example of an available tool, Transdecoder [68], somewhat addresses this limitation – however, it was developed for use with Trinity [68] which is intended for completely *de novo* transcriptome assembly in the absence of a reference genome assembly or any annotations. Consequently, Transdecoder performs *de novo* ORF prediction with intent towards identifying all ORFs that could convincingly give rise to proteins. In most well-annotated genomes, however, existing ORF annotations are available for the majority of genes in which new transcripts might be identified. In these cases a potentially more reliable approach is to examine novel coding sequences that begin with high-confidence annotated start codons. This prediction approach is useful not only because it informs on potential novel peptides, but also because it has the ability to identify the presence

of premature termination codons (PTCs) or high-confidence start codons that lack an in-frame downstream stop codon. In these latter cases, translation is expected to result in surveillance of host transcripts via nonsense-mediated decay (NMD) and non-stop decay (NSD) respectively.

NMD exemplifies an important potential outcome of AS. This translation-dependent surveillance mechanism identifies transcripts containing PTCs ≥ 55 nt upstream of a splice junction and triggers its degradation and translational suppression [69]. PTCs can be introduced through AS by inclusion of PTC-containing exons (poison exons), through splicing events that shift the reading frame of the message and by splicing events occurring within the 3' untranslated region (3'UTR) [70]. Other outcomes that can dramatically affect the fate of mRNAs may involve coding-to-noncoding switches, long-to-short UTR switches, or the inclusion of rare codons. It is thus important to profile mRNA coding features, and to connect binary AS events to their potential impacts to mRNA fate and function. The functional impacts of AS, however, remain difficult to predict.

To address this problem we developed junctionCounts comprising: junctionCounts event identification and quantification modules, cdsInsertion and findSwitchEvents. junctionCounts identifies and quantifies a diverse array of AS events. cdsInsertion translates provided transcripts *in silico* from user-provided overlapping start codons and determines resulting transcript characteristics such as UTR lengths, putative primary structure, the presence of PTCs, PTC distances from downstream splice sites, and more. Its partner utility, findSwitchEvents, bridges the gap between isoform- and event-level analysis, allowing one to take advantage of the potential superior quantification accuracy and regulatory interpretation of event-level analysis while still leveraging information that can only be derived from full-length transcripts [71]. In this study, we present junctionCounts as a powerful and

flexible tool for studying AS in a variety of cellular contexts, and we demonstrate its utility in not only identifying significantly regulated splicing events, but also inferring their functional outcomes.

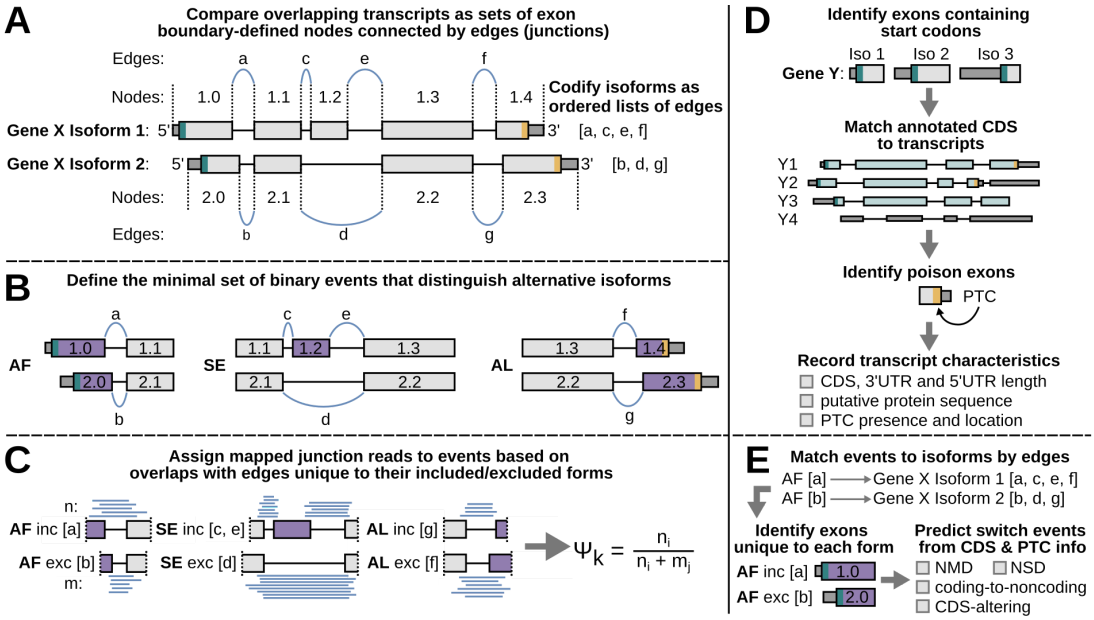


Figure 1. Overview of junctionCounts and partner utilities: cdsInsertion and findSwitchEvents. (A) junctionCounts takes a transcriptome annotation and makes pairwise comparisons of overlapping transcripts within genes, handling them as nodes (exons) connected by edges (junctions). (B) This iterative process identifies and classifies the minimal set of binary events that differentiates alternative isoforms. (C) junctionCounts then assigns mapped junction reads to the unique junctions that define the included and excluded forms of events and quantifies percent spliced in (PSI) values based on the ratio of included form and total junction reads mapping to the event. (D) cdsInsertion identifies exons overlapping annotated start codons, matches annotated coding sequences (CDS) to transcripts, identifies exons containing PTCs (poison exons), and finally records isoform-level information about the CDS. (E) Lastly, findSwitchEvents matches the included and excluded forms of events to isoforms, distinguishes the unique exons of each form, and predicts the impact of an event's inclusion/exclusion on the CDS. These predictions specify events whose inclusion/exclusion may confer: NMD, NSD, a coding-to-noncoding switch, or simply an alteration of the CDS relative to its opposite form.

MATERIALS AND METHODS

Alternative event definition and classification in junctionCounts

Alternative events are defined as instances in which pairs of: a) identical upstream 5'- and identical downstream 3'-exon boundaries, b) non-identical upstream 5'-transcript termini and identical downstream 3'-exon boundaries or c) identical 5'-exon boundaries and non-identical 3'-transcript termini are separated by any two combinations of distinct exon coordinates. Cases in which the aforementioned pairs are separated by more than two sets of distinct exon coordinates result in distinct alternative events for all pairwise combinations of those sets. Cases in which two or more events share the same splicing structure result in a single representative event in which the most proximal outer exon boundaries are used. This approach to event identification encompasses standard alternative event types and further identifies non-standard events of complex exon structure.

junctionCounts begins by identifying alternative events solely from a user-provided transcriptome annotation, in contrast to many other tools that also require BAM files, via its event identification module: infer_pairwise_events.py. This approach allows users to generate a single event catalog for the latest GENCODE transcriptome annotation [72], for example, and to use it across multiple datasets involving samples from a common species. If detection of novel splice sites and consideration of unannotated exons is desired, users can assemble a transcriptome *de novo* from the RNA-seq data under study with a tool like StringTie [67] and provide it as input to the event identification module. junctionCounts begins inferring pairwise events by generating a dictionary of annotated transcripts containing chromosome and strand information, a sorted list of exon coordinates and a sorted list of accompanying junction coordinates which are simply the inner coordinates of any pair of exons. It then filters the transcript dictionary based on user-defined cutoffs with the following defaults: exon length within 3 nt - 35 kb and intron length within 20 nt - 1 Mb. Then, it codifies

transcripts as nodes based on exon boundaries, connected by edges based on junction coordinates (Figure 1A). After establishing these simplified transcript structures, it makes pairwise comparisons of overlapping transcripts and iteratively defines the unique nodes that differentiate them; ultimately arriving at the minimal set of binary events that distinguish alternative isoforms in a gene (Figure 1B). During this process, events are also classified based on the nature of component features. The majority of event types in junctionCounts correspond to types explicitly defined elsewhere [73], but it also adopts previous usage of the term ”complex” [74,75] to refer to non-standard event types, distinguishing between internal (CO), 5’-terminal (CF), and 3’-terminal (CL) contexts. Below are descriptions of the criteria that define each event type:

A3 - *alternative 3’-splice site*: an event in which an upstream exon and the 3’-boundary of the downstream exon are common to both isoforms, but the 3’-splice site is distinct. The form containing the most upstream of the alternative 3’-splice sites is the included form.

A5 - *alternative 5’-splice site*: an event in which a downstream exon and the 5’-boundary of the upstream exon are common to both isoforms, but the 5’-splice site is distinct. The form containing the most downstream of the alternative 5’-splice sites is the included form.

AF - *alternative first exon*: an event in which each isoform has its own distinct 5’-terminal exon; each with a distinct 5’-terminus and 3’-splice site. The exon immediately downstream of the terminal exon is common to both isoforms. The form with the most upstream 5’-terminus is the included form.

AL - *alternative last exon*: an event in which each isoform has its own distinct 3'-terminal exon (i.e. each with a distinct 3'-terminus and 5'-splice site). The exon immediately upstream of the terminal exons is common to both isoforms. The form with the most downstream 3'-terminus is the included form.

MX - *mutually exclusive exons*: an event in which a pair of 3'- and 5'-splice sites are separated by a distinct exon in each isoform. The only requirement for the two exons being distinct is that they do not share either splice site. It is possible for the alternative exons in an MX event to partially and completely overlap, provided their boundaries do not coincide. The form in which the alternative exon's 3'-splice site is the most upstream of the mutually exclusive exons is the included form.

RI - *retained intron*: an event in which a pair of adjacent exons are spliced together in one isoform, but joined together in another by retention of the intron separating them. The form with the retained intron is the included form.

SE - *skipped exon*: an event in which a pair of 3'- and 5'-splice sites are separated by a single exon (the skipped exon) in one isoform and spliced directly together in another. The form containing the intermediate exon is the included form.

MF - *multiple alternative first exons*: an event in which each isoform has its own distinct 5'-terminal set of one or more exons upstream of a single shared exon. This event type is distinguished from AF in that either isoform must contain more than one unique exon. The form with the most upstream 5'-terminus is the included form.

ML - *multiple alternative last exons*: an event in which each isoform has its own distinct 3'-terminal set of one or more exons downstream of a single shared exon. This event type is distinguished from AL in that either isoform must contain more than one unique exon. The form with the most downstream 3'-terminus is the included form.

MR - *multiple retained intron*: an event in which a set of three or more exons are spliced together in one isoform, but connected in the other by two or more consecutive retained introns. The form with the retained introns is the included form.

MS - *multiple skipped exons*: an event in which a pair of 3'- and 5'-splice sites are separated by multiple exons in one isoform, and spliced directly together in another. The form containing the intermediate exons is the included form.

CO* - *complex internal*: this is a general category for events that do not meet any of the above criteria and do not involve transcript termini. The isoform with the longest spliced length is the included form.

CF* - *complex 5'-terminal*: this is a general category for events that do not meet any of the above criteria and involve alternative 5' transcript termini. The form with the most upstream 5'-terminus is the included form.

CL* - *complex 3'-terminal*: this is a general category for events that do not meet any of the above criteria and involve alternative 3' transcript termini. The form with the most downstream 3'-terminus is the included form.

*Complex event types capture less straightforward cases of binary splicing events that can involve combinations of multiple alternative features (terminal and internal exons and/or splice sites).

After generating the initial event dictionary, junctionCounts writes BED files for introns and exons separately and uses bedtools intersect [45] with the arguments -wa -wb -s -f 1.00 to identify introns that completely overlap an exon. These introns are then considered putative RI events. Finally, junctionCounts organizes events by edges (junctions) that are unique to their included and excluded forms, collapses any redundant events that have identical included and excluded form edges, and filters them for events that can be quantified by exon-exon junction reads alone (or by exon-intron junction reads in the case of RI/MR events).

Alternative event quantification in junctionCounts

junctionCounts employs a junction read-centric approach to alternative event quantification. First, the event quantification module, junctionCounts.py, generates a nested containment list [76] of all event junction coordinates. Then, for each read or read pair, junctionCounts considers matches between splice junctions identified in the alignment and event splice junctions, as well as overlaps between contiguous mapped read sequence and informative exon-intron junctions (Figure 1C). Informative exon-intron junctions are those that are overlapped by an exon in the alternative isoform. Reads overlapping such an exon-intron junction are considered consistent with the alternative isoform. Key examples of this occur in the excluded isoform of RI events, which are overlapped by the exon of the included form.

After establishing the event isoforms with which a read is consistent, junctionCounts attempts to disambiguate the read assignment using exon-exon and exon-intron junctions that are

unique to specific isoforms, when possible. With this approach, junctionCounts goes beyond simple junction-by-junction read counting. Both the event and the informative exon-intron junction definition prohibit scenarios in which reads are assigned to both isoforms of the same event. With read-to-event isoform consistencies established, read counts are tallied for each exon-exon and informative exon-intron junction for each isoform of each event. A percent spliced in (PSI or Ψ) value is calculated for all pairwise combinations of included and excluded junction counts, yielding the ratio between an included form's junction counts and the sum of the included *and* excluded form's junction counts:

$$\Psi_k = \frac{n_i}{n_i + m_j}$$

(1.0)

Where n_i is the number of reads assigned to the included form junction i and m_j is the number of reads assigned to the excluded form junction j. A set of PSI values is established for each sample. Additionally, junctionCounts calculates the minimum (Ψ_{min}), maximum (Ψ_{max}) and span PSI (Ψ_{span}):

$$\Psi_{min} = min(\Psi_k) \quad (1.1) \qquad \Psi_{max} = max(\Psi_k) \quad (1.2) \qquad \Psi_{span} = \Psi_{max} - \Psi_{min} \quad (1.3)$$

The Ψ_{span} serves as a rough measure of within-sample uncertainty. junctionCounts reports these values as well as the included and excluded junction read counts (n and m in Equation 1.0) for each event. As an optional method of assessing within-sample uncertainty, junctionCounts offers bootstrap quantification, which repeats a user-specified number of rounds of bootstrap read selection and re-quantification. For each bootstrap round, junctionCounts reports all measurements, in addition to the initial non-resampled quantification.

Statistical testing of events between conditions in experiments with at least two replicates per condition is done with the condition comparison module, DEXSeq_comparison.R, which employs DEXSeq [49]. First, it produces a DEXSeq-compatible GFF file that defines the included and excluded form of each event as pairs of "exonic parts". It subsequently writes included and excluded form junction counts per sample as separate count files. DEXSeq then normalizes included and excluded junction counts as per its documentation [49], and estimates included and excluded form junction count dispersions using a Cox-Reid adjusted profile likelihood estimation followed by fitting of a dispersion-mean relation to the individual dispersion values and shrinkage of per-form estimates toward fitted values. Next, DEXSeq compares the deviances of the included and excluded form junction counts in each event across conditions using a χ^2-distribution to produce p-values for each form. Event-level Q-values are then calculated from the p-values with the Benjamini-Hochberg procedure. Finally, events are filtered based on user-defined cutoffs with the following defaults: ≥ 0.1 mean PSI in at least 1 condition, ≥ 15 total mean junction counts across all replicates, and ≤ 0.03 span PSI for RI/MR events. The final results provide event coordinates, quantification results, and classification of events as significant based on user-defined cutoffs with the following defaults: $|dPSI| \geq 0.1$ and Q-value ≤ 0.05.

Description and validation of partner utilities: cdsInsertion and findSwitchEvents, which couple alternative splicing events to isoform-level impacts to the CDS

cdsInsertion translates transcripts *in silico* from user-provided overlapping start codons and determines resulting transcript characteristics including: UTR lengths, putative protein sequences, the presence of PTCs, PTC distances from downstream splice sites and more (Figure 1D). For a given codon and transcript, cdsInsertion first checks whether the start codon's first position overlaps the genomic coordinates of a transcript's exons. If it does, the

start codon sequence is checked in the spliced transcript's sequence. Currently, only AUG initiation is supported. If the start codon sequence is AUG, cdsInsertion translates the spliced transcript *in silico* by looking for an in-frame downstream stop codon, which can be: UAA, UGA or UAG. If a downstream stop codon is found, the resulting CDS is recorded along with associated information such as CDS length, coding sequence, PTC presence and PTC distance. If more than one CDS is found within the transcript, additional CDS features are associated to the given transcript as distinct CDS features. If no in-frame downstream stop codon is found, the transcript is recorded as a possible NSD substrate. cdsInsertion outputs a table with summary information about each transcript and a separate GTF file for potential non-PTC, PTC, and non-stop transcript-CDS combinations. The GTF file contains separate transcript records for every CDS-transcript combination. Optionally, cdsInsertion can additionally output bigGenePred files which enable codon visualization on the UCSC Genome Browser. cdsInsertion further outputs a pickled Python dictionary containing all of the aforementioned information associated with each transcript.

Its partner utility, findSwitchEvents, takes an IOE file; a format originally introduced by SUPPA [77], which is generated by junctionCounts to associate events with transcripts, and the pickled Python dictionary containing transcript CDS information and associated details from cdsInsertion. With this information, it evaluates whether isoforms with a specific property (NMD, NSD or unique CDS) are exclusive to one form of an alternative event (Figure 1E). Switch events are AS events that meet this condition, meaning that one (either the included or excluded) form is coupled with a switch to PTC-containing, non-stop or noncoding isoforms within a gene. We evaluated these tools on two published datasets. First, we procured RT-PCR data for well-characterized NMD targets and accompanying RNA-seq data from HEK-293 cells upon UPF1 knockdown versus non-targeting siRNA [78]. We then ran junctionCounts and its partner utilities with default settings on the RNA-seq data to

predict NMD switch events within the NMD targets verified with RT-PCR in the original study. Out of the 13 predicted NMD switch events associated with the NMD targets, 11 had the expected dPSI directionality (Figure 2L). Additionally, we used the same approach on a dataset in which treatment with emetine, a translation elongation inhibitor, was reported to increase the abundance of NMD substrates in HEK-293 T cells [50]. Indeed, cdsInsertion and findSwitchEvents identified 636 potential NMD switch events with significant changes in splicing ($|dPSI| \geq 0.1$ and Q-value ≤ 0.05) upon emetine treatment, out of which 472 (74.2%) exhibited the expected dPSI directionality (Figure 2M, N).

Benchmarking performance across five AS analysis tools

junctionCounts was evaluated on its performance relative to four established splicing analysis tools: MAJIQ v2.5.6.dev1+g8423f68b [79], rMATS-turbo v4.3.0 [80], splAdder v3.0.4 [81] and Whippet v1.6.2 [71]. We generated four paired-end simulated datasets from real RNA-seq data using polyester v1.36.0 [82] with the arguments: read.length = 100, fragment.length.min = 100, fragment.length.max = 500, fragment.length.mean = 180, fragment.length.sd = 40 and simulated.sequencing.error = TRUE. Three of the simulated datasets: 25M, 50M and 75M, were used to evaluate performance at different library depths (25, 50 and 75 million reads per library respectively). These datasets were modeled on mouse cerebellum and liver RNA-seq data in triplicate for each cell type from Vaquero-Garcia et al. (2016) [79], who replicated the experiments in Zhang et al. (2014) [83]. The fourth simulated dataset was modeled on human RNA-seq data from HeLa cells treated with either spliceostatin A or DMSO [78] at 50 million reads per library with triplicates for each condition. This dataset introduced a larger pool of potential AS events to test compared to the murine simulated datasets, and importantly increased the total number of MR events tested.

To generate these simulated datasets, we first downloaded the FASTQ files for the aforementioned mouse and human RNA-seq data [78,79] from their respective data repositories and verified their quality with FASTQC v.0.12.1. Next, we mapped them to their respective genomes and transcriptomes (GRCm38.p6 with the GENCODE vM20 primary assembly annotation for mouse data and GRCh38 with the GENCODE v41 primary assembly annotation for human data) [72] with STAR v2.7.8a [39]. Then, we quantified transcript-level expression with Salmon v.1.10.2 [84]. The transcript-level quantification results were used as input for polyester to simulate RNA-seq libraries reflecting the exact transcripts per million (TPM) specified for each sample in each simulated dataset – thus producing datasets with known ground truth transcript expression values. Ground truth PSI and dPSI values for junctionCounts-defined AS events were calculated from the TPM values with a custom Python script that leverages the event-transcript associations in the junctionCounts IOE file using the following equation:

$$\Psi = \frac{\sum_{i=1}^{n} TPM_i}{\sum_{i=1}^{n} TPM_i + \sum_{j=1}^{m} TPM_j} \quad (2.0)$$

where transcripts *i* in the numerator are those consistent with the included form of the event, and transcripts *j* in the denominator are those consistent with the excluded form of the event. We then mapped the simulated data to the appropriate genomes with STAR [39] and evaluated each tool's performance on the resulting BAMs (and FASTQ files in the case of Whippet). All tools were run on a System76 Lemur Pro laptop with an Intel® Core™ Ultra 5 125U @4.3GHz processor, 14 total cores and 40GB RAM. The total time elapsed for all steps of each tool's pipeline and the memory peak among all steps of each tool's pipeline were measured with the default linux package "time" (http://man7.org/linux/man-pages/man1/time.1.html) with the "-v" flag. Below are the run

parameters for each tool with specific arguments and flags noted, excluding arguments related to user-specific input/output files and directories. The following variables refer to either the mouse or human versions of genome sequences and annotations: “$GTF” (GENCODE vM20/GENCODE v41 primary assembly annotation GTF) “$GFF3” (GENCODE vM20/GENCODE v41 primary assembly annotation GFF3) and “$FASTA” (GRCm38.p6/GRCh38 genome sequence FASTA).

junctionCounts was run with default settings:

1. Event identification step: infer_pairwise_events.py --transcript_gtf $GTF
2. Event quantification step: junctionCounts.py
3. Condition comparison step: DEXSeq_comparison.R

MAJIQ:

1. Event identification step: majiq build $GFF3 --minreads 5
2. Event quantification step: majiq psi
3. Condition comparison step: majiq deltapsi
4. Produce output TSV file: voila tsv --changing-between-group-dpsi 0.1 --threshold 0.1
5. Modulize Voila files: voila modulize --show-all --changing-between-group-dpsi 0.1

rMATS-turbo:

1. All steps: rmats.py --gtf $GTF -t paired --libType fr-secondstrand --readLength 100 --variable-read-length --nthread 4

splAdder:

1. Event identification and quantification step: spladder build -a $GTF --remove-se --validate-sg-count 3 --confidence 2
2. Condition comparison step: spladder test --confidence 2

Whippet:

1. Preparation step (not included as part of Whippet's time/memory evaluation because it's not a required step): generate a merged, deduplicated and indexed BAM file from all the sample BAMs as input with samtools merge, sort, rmdup and index commands.
2. Event identification step: julia whippet-index.jl --fasta $FASTA --bam merged.sort.rmdup.bam --gtf $GTF --suppress-low-tsl --bam-min-reads 5
3. Event quantification step: julia whippet-quant.jl --biascorrect
4. Condition comparison step: julia whippet-delta.jl

The next step to evaluate performance on the simulated datasets was to identify events defined by each tool that approximately reproduce the junctionCounts-defined events for which we have ground truth PSI values. Each tool defines events with their own unique and valid approach, leading to many events with slightly different exon/splice site/intron node edges across tools. Therefore, each tool was evaluated on its individual subset of events that matched junctionCounts-defined events in the simulated datasets with ≥ 95% overlap at each participating alternative feature. Performance at the event detection and quantification level (PSI) and at the event change level (dPSI) were then assessed in terms of the following metrics with specific adjustments to maximize fairness:

$$TPR = \frac{TP}{TP + FN} \quad (2.1) \qquad FDR = \frac{FP}{FP + TP} \quad (2.2) \qquad MAE = \frac{\sum|GT - OB|}{n} \quad (2.3)$$

$$PPV = \frac{TP}{TP + FP} \quad (2.4) \qquad NPV = \frac{TN}{TN + FN} \quad (2.5)$$

For event detection and quantification (PSI) metrics, we measured: sensitivity (TPR), false discovery rate (FDR) and mean absolute error (MAE). True positives (*TP*) were defined as events for which the ground truth (*GT*) and a given tool's measured/observed PSI values (*OB*) were both > 0. False negatives (*FN*) were defined as events with *GT* PSI > 0 and *OB* PSI = 0. False positives (*FP*) were defined as events with *GT* PSI < 0.05 and *OB* PSI ≥ 0.05. We chose

this definition of *FP* because each tool commonly reported miniscule *OB* PSI values < 0.05 for events with *GT* PSI = 0, which would in most normal use cases be excluded or filtered in subsequent analyses unless they had more substantial (typically ≥ 0.1) PSI values in another condition. TPR and FDR were calculated with the described definitions of *TP*, *FP* and *FN*. MAE was calculated by summing the absolute differences between *GT* and *OB* PSI values for *TP* events and dividing it by the number of observations, *n*.

For event change (dPSI) metrics, we measured: positive predictive value (PPV), negative predictive value (NPV), FDR and MAE. Each metric was measured at cumulative |dPSI| thresholds starting at 0.1 and increasing stepwise by 0.05 to 1.0, such that each measurement considers the subset of events with |*GT* dPSI| both at and below the given threshold. Each tool was evaluated based on its accuracy of significant/insignificant calls and quantification of event changes. Ground truth events with |*GT* dPSI| ≥ 0.1 were considered significant. Therefore, each tool's condition comparison step was given the appropriate argument specifying a dPSI threshold of 0.1 to be considered statistically significant. Significant events were defined for each tool as those with |*OB* dPSI| ≥ 0.1 and the accompanying tool-specific statistical cutoff: Q-value ≤ 0.05 (junctionCounts), probability_changing ≥ 0.95 (MAJIQ), p-value ≤ 0.05 (rMATS-turbo), adjusted p-value ≤ 0.05 (splAdder) or probability ≥ 0.95 (Whippet). We did not measure TPR for the dPSI evaluation because while a given event may meet the |*GT* dPSI| ≥ 0.1 threshold across conditions, each tool's statistical evaluation of *OB* dPSI measurements may justifiably call the event statistically insignificant based on numerous factors including junction read support and dispersion characteristics. To mitigate this possibility, we filtered the ground truth events for those with minimum total junction read support ≥ 15 in both conditions. For this test, *TP* were defined as events with either *GT* dPSI ≥ 0.1 and *OB* dPSI ≥ 0.1 or *GT* dPSI ≤ -0.1 and *OB* dPSI ≤ -0.1 (with the tool-specific statistical cutoff described earlier). *FP* were defined as events with |*GT* dPSI| < 0.02 that were

called significant by a given tool. We chose this definition of *FP* because events with |*GT* dPSI| < 0.02 were a higher-confidence "not changing" set of events which each tool should correctly identify as insignificant. *TN* were defined as events with |*GT* dPSI| < 0.1 that were called insignificant by a given tool. *FN* were defined as events with |*GT* dPSI| ≥ 0.1 that were called insignificant by a given tool. PPV, NPV and FDR were calculated with the described definitions of *TP*, *FP*, *TN* and *FN*. MAE was calculated as described earlier for the PSI metrics, but at cumulative |GT dPSI| thresholds.

Analysis of interspecies and temporal alternative splicing dynamics during neuronal differentiation in primate PSCs

We analyzed RNA-seq data from human and rhesus macaque embryonic stem cells (ESCs) as well as chimpanzee and orangutan induced pluripotent stem cells (iPSCs) [51]. Field et al. induced differentiation of the stem cells to cortical neurospheres to model prenatal brain development. Duplicate RNA-seq libraries from each time point (0, 1, 2, 3, 4 and 5 weeks of neuronal differentiation) were downloaded as compressed FASTQ files from SRA, deduplicated and mapped with STAR to the appropriate genome: GRCh38 [72], panTro4 [85], ponAbe2 [86] and rheMac8 [87] for human, chimpanzee, orangutan, and rhesus macaque respectively. The GENCODE v27 basic gene annotation [72] was used as a basis for CAT [88] to generate gene annotations of similar complexity for all species. To reduce the complexity of the input transcriptomes, only basic transcripts were retained. Following this filtration, these annotations were used as input along with the mapped RNA-seq reads for StringTie v1.3.6 [67] to assemble unannotated transcripts. Using the StringTie merge command, comprehensive gene annotations were produced for each species.

To identify orthologous AS events, the whole genome sequences of human (GRCh38), chimpanzee (PanTro4), orangutan (ponAbe2), and rhesus macaque (rheMac8) were mapped

to one another using minimap2 v2.11-r797 [89] with parameters *--cs* and *-asm20*. The resulting mappings were used to lift the coordinates of alternative event exons to other species with a modified version (altered such that the input BED file coordinates are semicolon-delimited rather than underscore-delimited in the name field of the output BED file) of the minimap partner utility paftools. Events were reassembled from the lifted coordinates of component exons, and checked for exon count and event type-concordance with the original event. Lifted events passing these checks were proposed as putative orthologs, and then checked against events natively identified in the target species to identify orthologous relationships. Non-one-to-one relationships were removed from consideration.

Temporal (time points 1-5 weeks of neuronal differentiation versus t_0) and interspecies (paired time points compared across species) AS analyses were performed using junctionCounts. Events with $|dPSI| \geq 0.1$ and Q-value ≤ 0.05 across conditions were considered significantly different. Events exhibiting significant splicing differences in at least one temporal comparison for each species were clustered by their temporal PSI trajectories with CLARA (cluster v2.1.6) [90] using euclidean distance and with 500 iterations. Conservation of primate temporal splicing dynamics was assessed based on concordance with human temporal splicing dynamics with regard to PSI measurements. Genes with mean $|dPSI| \leq 0.1$ for events in chimpanzee, orangutan and rhesus macaque relative to human in pairwise comparisons at each time point were categorized as genes with conserved splicing, while those with mean $|dPSI| \geq 0.3$ were categorized as genes with non-conserved splicing. Gene ontology analyses for genes in the temporal event clusters and conserved and non-conserved splicing gene sets was done with Metascape [91].

Functional splicing analyses included assessment of switch events and exonic features. Switch events, which we define as events in which all transcripts consistent with one

alternative form contain a particular feature while all transcripts consistent with the other form do not, were identified using cdsInsertion and findSwitchEvents (https://github.com/ajw2329/cds_insertion). In NMD switch events one form, but not the other, introduces a premature termination codon (in-frame stop codon ≥ 55 nt upstream of the final exon-exon junction when translated *in silico* from any overlapping consensus coding sequence start codon). In NSD switch events one form, but not the other, results in transcripts lacking an in-frame stop codon. In coding (to noncoding) switch events one form, but not the other, results in transcripts lacking a coding sequence. Exon ontology analysis was performed with Exon Ontology [92] using the included form coordinates of alternative exons as the test list, and the excluded form coordinates as the background.

Data sources

RNA-seq data from mouse cerebellum and liver cells [83] and RNA-seq data from spliceostatin A and DMSO-treated HeLa cells [78] was used to generate simulated data for the benchmarking experiment. RNA-seq data and NMD target RT-PCR data from UPF1 siRNA and non-targeting siRNA-treated HEK-293 cells [93] and RNA-seq data from emetine and DMSO-treated HEK-293 T cells [50] was used to validate cdsInsertion and findSwitchEvents NMD predictions. RNA-seq data from human and rhesus macaque ESCs as well as chimpanzee and orangutan iPSCs [51] was used to demonstrate junctionCounts' utility in a variety of applications.

RESULTS

junctionCounts and its partner utilities facilitate user-friendly isoform-level analysis

junctionCounts can characterize alternative events in any user-provided transcriptome annotation. It defines the minimal set of binary splicing events that distinguish alternative

isoforms within gene models and classifies events into event types that range from simple, canonical events to complex events that capture coordinated splicing of multiple event features comprising multiple event types. To be clear, in the context of this work, we define complex events as those involving multiple alternative feature types (i.e. a CF event representing coordinated AF-SE-A3 splicing), which excludes MS events, for example. Other AS analysis tools, including MAJIQ [79] and Whippet [71], can also characterize complex events. However, junctionCounts uniquely summarizes complex event junction read support within an easily interpretable binary context; assigning a single PSI value to the included and excluded form of events rather than to individual splice sites or features. Beyond alternative event definition and classification, junctionCounts' partner utilities enable valuable prediction of functional outcomes, including NMD, by connecting individual splicing events to their effects on the CDS at the isoform level. cdsInsertion derives CDS information from a transcriptome annotation and findSwitchEvents uses junction coordinate keys to associate the included and excluded form of events to their respective alternative isoforms. Altogether, junctionCounts presents an easy to install, easy to use set of tools with relatively few dependencies and the novel capability of event-to-isoform CDS characterization within the AS analysis milieu.

junctionCounts accurately quantifies alternative splicing events

We evaluated junctionCounts' performance on simulated data, with known ground truth PSI values for junctionCounts-defined events, modeled on real RNA-seq data with that of four established AS analysis tools: MAJIQ [79], rMATS-turbo [80], splAdder [81] and Whippet [71]. We generated four simulated datasets in total: three datasets at 25, 50 and 75 million reads per library were modeled on mouse cerebellum and liver RNA-seq data [79] with triplicates per cell type to evaluate performance at different library sizes. The fourth dataset

was modeled on human RNA-seq data at 50 million reads per library with triplicates for two conditions: spliceostatin A (SSA) and DMSO treatment. This dataset provided a larger pool of events to test relative to the murine datasets, including over 2000 MR events of which there were less than 100 in the murine datasets. Each tool was run on a laptop, as described in the Materials and Methods, and the time from start to finish of all analysis steps and the peak memory cost at any point during that time were recorded (Figure 2A, B). junctionCounts exhibited the median time and memory cost among the other tools, which scaled with library size and transcriptome complexity for all tools.

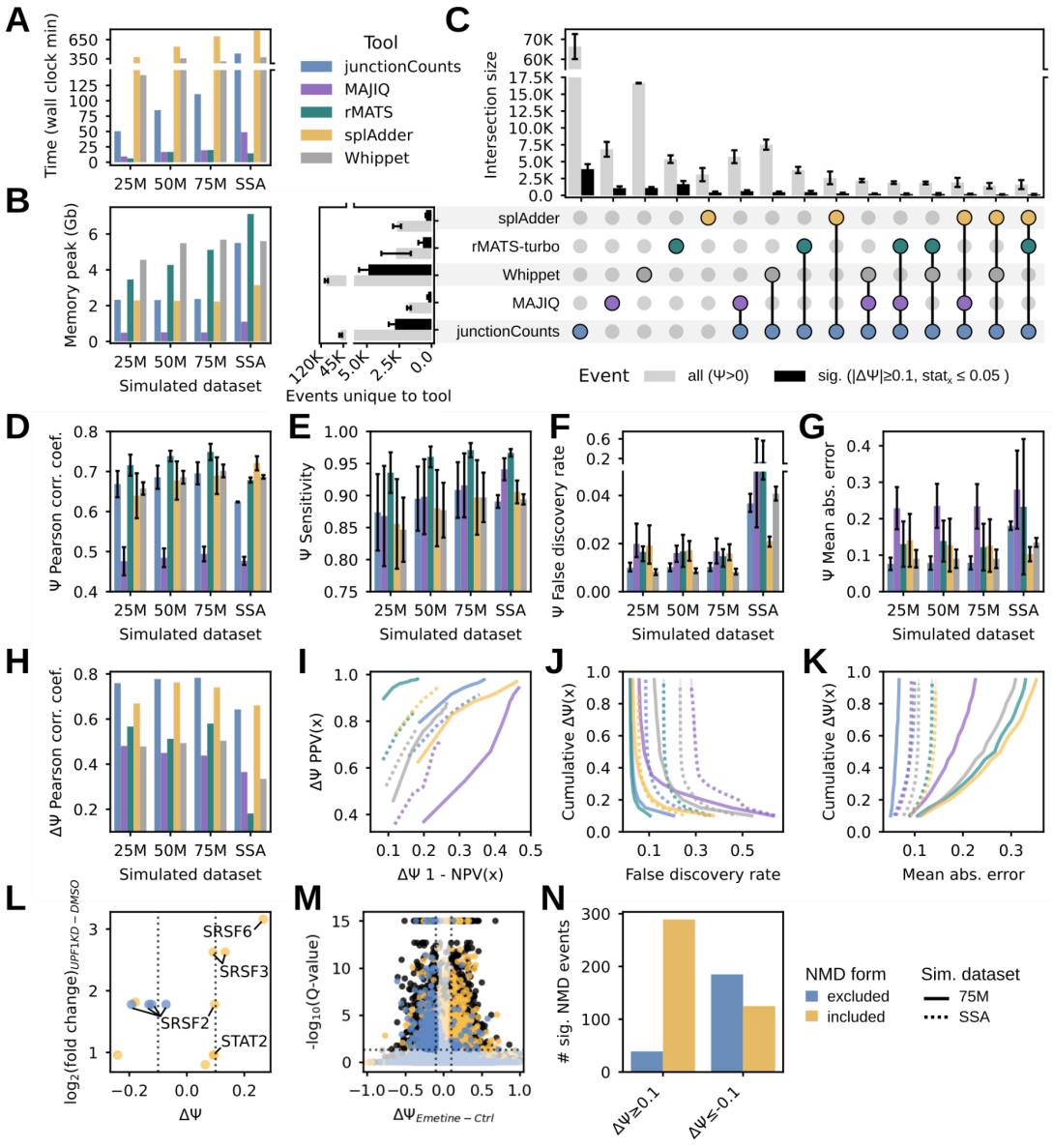


Figure 2. Benchmarking experiment to evaluate junctionCounts performance relative to similar tools. Four simulated datasets were generated: 25M, 50M and 75M refer to the library depths of samples simulated from mouse cerebellum and liver RNA-seq in triplicate for each cell type. SSA refers to samples simulated from human cells treated with spliceostatin A or DMSO in triplicate for each condition, at 50M reads per library. dPSI measurements were made across cell types in the mouse data and across conditions in the human data respectively. (A) Measurements of elapsed wall clock time upon running full tool pipelines to completion. (B) Peak memory consumption at any point of each tool's full pipeline. (C) Upset plot showing the intersection of each tool's alternative events with junctionCounts-defined ground truth events in the 25M, 50M and 75M datasets. Gray bars represent all events detected by a given tool or set of tools which had PSI > 0 in at least 1 condition. Black bars represent events determined to be differentially spliced across

conditions by a given tool or set of tools (|dPSI| ≥ 0.1 and the tool's associated statistic meeting a probability of 0.95 or a FDR/p-value of 0.05 for the event). The horizontal barplot on the left shows the number of events that didn't overlap between each tool and junctionCounts – or in the case of junctionCounts itself, the number of events that were not reproduced by any other tool. (D) Pearson correlation coefficient of measured PSI values for each sample compared with its cognate ground truth (6 total replicates and comparisons per dataset). (E) Sensitivity and (F) false discovery rate of event detection. (G) Mean absolute error of PSI measurements. Error bars for (D, E and G) depict standard deviation, while those for (F) show the full range of observations. (H) Pearson correlation coefficient of measured dPSI values derived from each tool's respective condition comparison steps with accompanying statistical filtering (|dPSI| ≥ 0.1 and Q-value/p-value ≤ 0.05 or probability ≥ 0.95). (I) Predictive receiver operating characteristic (PROC) showing positive predictive value (PPV) and negative predictive value (NPV) of significant dPSI calls at cumulative ground truth dPSI thresholds. (J) False discovery rate of dPSI calls at cumulative ground truth dPSI thresholds. (K) Mean absolute error of dPSI measurements at cumulative ground truth dPSI thresholds. (I-K) Only the 75M and SSA datasets (solid and dotted lines respectively) are shown because the 25M and 50M datasets had nearly identical curves to 75M. (L) RT-PCR $\log_2$(fold change) of NMD targets in UPF1-KD HEK-293 cells vs. DMSO compared with dPSI measurements of junctionCounts-predicted NMD events within those targets. (M) Volcano plot of junctionCounts-predicted NMD events in emetine-treated HEK-293 T cells vs. DMSO. (N) The number of significant predicted NMD events stratified by dPSI directionality upon emetine treatment in HEK-293 T cells. Events in (L-N) are categorized as included or excluded NMD form, meaning that the included form or the excluded form, respectively, is predicted to confer NMD to the resulting transcript. Performance metrics in (E-G and I-K) are described in detail in the Materials and Methods.

Each publicly available AS analysis tool identifies and quantifies AS events within its own event type repertoires and definitions, thus complicating their comparison. rMATS-turbo and splAdder are limited to non-terminal event types, while junctionCounts, MAJIQ and Whippet are each capable of characterizing terminal events and additionally non-canonical, complex event types involving coordinated splicing of multiple alternative feature types. Each tool's set of events identified from the simulated data were matched to junctionCounts-defined events to compare ground truth to observed PSI values. In order for an event, identified by a given tool, to match with a junctionCounts event, the coordinates of its participating features each had to overlap by ≥ 95%, which provides latitude for minute variation of event exon/intron nodes across tools. Despite this flexibility, no complex events across junctionCounts, MAJIQ or Whippet met the requirements for comparison, which reflects

each tool's unique approach to event definition, even with canonical event types. The mean overlap of junctionCounts events with other tools across the four simulated datasets was: 78% for MAJIQ, 75% for rMATS-turbo, 51% for splAdder and 13% for Whippet – amounting to thousands of events per tool. Each tool was tested individually on the subset of junctionCounts-defined events they approximately reproduced (Figure 2C, Supplemental Figure 1A-E).

We first measured performance at the PSI level (in-depth descriptions of testing procedures in Materials and Methods). At the PSI level, we evaluated each tool's event detection and quantification capabilities. Here, we defined sensitivity (true positive rate; TPR) as the proportion of events detected that had both a measured and ground truth PSI > 0. In this context, the sensitivity test measured a tool's ability to correctly assign read support to an event rather than its ability to reproduce ground truth events. Next, we tested false discovery rate (FDR) as the proportion of events that had a measured PSI ≥ 0.05 and a ground truth PSI < 0.05 relative to all events with a measured PSI > 0. This threshold was applied because each tool commonly misattributed miniscule PSI values to events with ground truth PSI 0, which in most normal use cases isn't a problem as final results are typically filtered for events meeting a minimal PSI value among conditions or replicates. Instead, this FDR test quantified the rate at which tools misattributed read support to a substantial degree (≥ 0.05 PSI) that would be problematic in typical AS analysis settings. Finally, we tested the accuracy of true positive event (measured and ground truth PSI > 0) quantification in terms of mean absolute error (MAE). junctionCounts generally represented the median of the five tools across these metrics (Figure 2D-G), with the caveat that it was tested on the largest number of events and event types (Supplemental Figure 1A-E). When stratified by event type, we found that junctionCounts consistently maintains a FDR < 3% and the largest area under curve of

cumulative mean absolute error distribution relative to the other tools for AF, AL, MX and MS event types among those directly compared (Supplemental Figure 1F-I).

Next, we assessed performance at the event change (dPSI) level, focusing on each tool's accuracy in calling and quantifying significant and insignificant event changes at cumulative dPSI thresholds. We measured positive predictive value (PPV), which is the proportion of correct significant event calls made by a given tool that have |ground truth dPSI| $\geq$ 0.1 relative to all significant event calls and negative predictive value (NPV), or the proportion of correct insignificant event calls with |ground truth dPSI| < 0.1 relative to all insignificant event calls. We also measured FDR, which we defined as the proportion of incorrect significant event calls with |ground truth dPSI| < 0.02 among all significant calls, and finally the MAE of measured vs. ground truth dPSI values. junctionCounts had the highest dPSI Pearson correlation coefficient, just under 0.8, among the other tools on the murine datasets, and notably never surpassed 0.1 dPSI MAE at any dPSI threshold, while Whippet, rMATS-turbo and splAdder all surpassed 0.3 dPSI MAE on the SSA dataset. Whippet and junctionCounts maintained the smallest distance between cumulative PPV/NPV and FDR curves across the 75M and SSA datasets, while junctionCounts achieved the minimal distance of cumulative MAE curves among the other tools across datasets (Figure 2H-K). Taken together, junctionCounts had the most consistent performance across dPSI metrics between murine and human datasets compared to the other tools.

Characterizing temporal and species-specific alternative splicing dynamics during primate neuronal differentiation

After establishing that junctionCounts competently characterizes AS in simulated data, we next wanted to examine its utility on real data. To that end, we analyzed a primate neuronal

differentiation RNA-seq dataset comprising human and rhesus macaque ESCs as well as chimpanzee and orangutan iPSCs [51] (Figure 3A). We hypothesized that because the four primates share 90-99% genome sequence conservation [94], junctionCounts should identify a substantial number of orthologous AS events across the four primates [95]. We further expected to observe substantial species-specific splicing dynamics during neuronal differentiation as previous interprimate studies have reported [95].

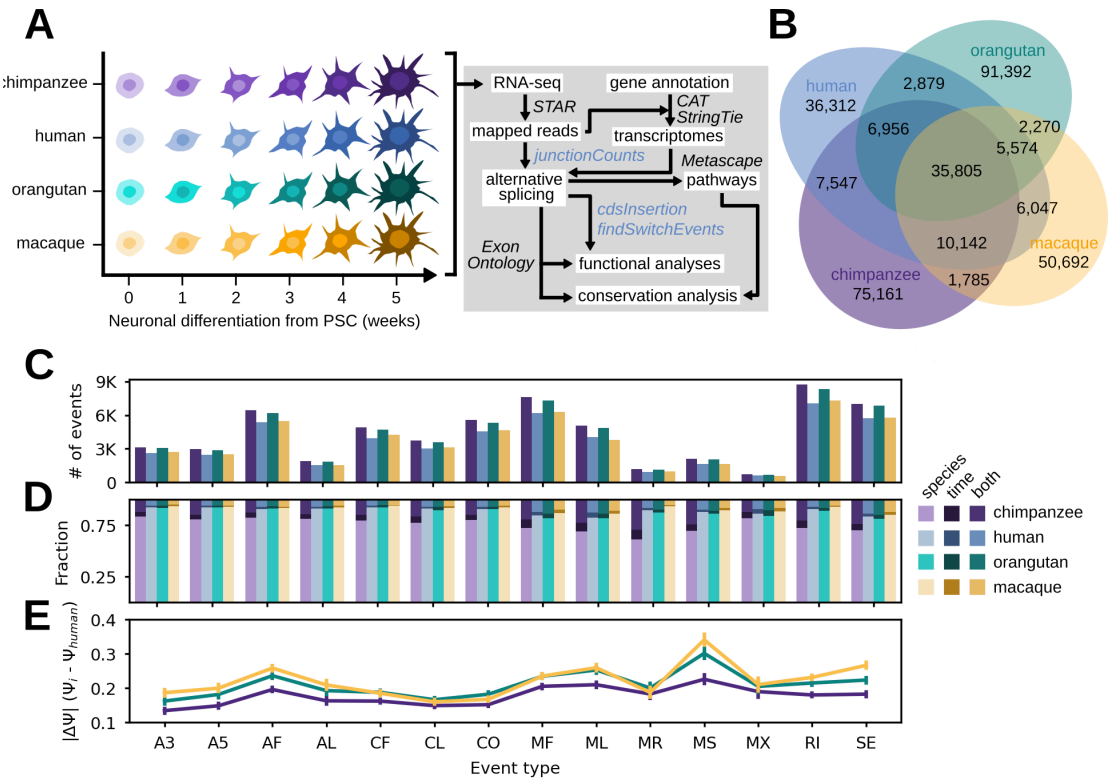


Figure 3. Application of junctionCounts to a primate neuronal differentiation time course experiment. (A) Schematic of five-week chimpanzee, human, orangutan and rhesus macaque neuronal differentiation from pluripotent stem cells and subsequent RNA-seq analysis workflow. (B) Venn diagram of the events identified by junctionCounts in each primate transcriptome. (C) The total number of significant events ($|dPSI| \geq 0.1$ and Q-value $\leq$ 0.05 in at least 1 temporal or interspecies comparison) by event type for each species. (D) The fraction of events that were significantly different across species, time or both factors. (E) Evaluation of conserved splicing by event type, measured by |dPSI| against human PSI values.

We used CAT [88] on the GENCODE v27 [72] basic gene annotation to generate gene annotations of similar complexity for all species. We then used StringTie v1.3.6 [67] on the resultant gene annotations along with the mapped RNA-seq reads to assemble unannotated transcripts. Thus we produced comprehensive gene annotations for each species. Using junctionCounts, we identified approximately 143K, 111K, 151K and 113K possible events in the chimpanzee, human, orangutan and rhesus macaque gene annotations respectively. And to identify orthologous AS events, we performed pairwise mapping of the whole genome sequences of human (GRCh38), chimpanzee (PanTro4), orangutan (ponAbe2), and rhesus macaque (rheMac8) using minimap2 [89]. Using the mappings, we lifted the coordinates of alternative event exons to other species using paftools. We then reassembled events from the lifted coordinates of component exons, assessed exon count and event type-concordance with the original events and checked these against events identified in the target species to establish orthologous relationships for which only one-to-one relationships were considered.

In all pairwise interspecies event set comparisons, at least 40% of events were not species-specific, with over 35K orthologous events common to all four primates (Figure 3B). We next quantified these events with junctionCounts which uses junction reads from the mapped RNA-seq data, after which we performed event-level – statistically tested with DEXSeq v3.19 [49] – pairwise temporal comparisons ($week_i$ versus $week_0$ of neuronal differentiation) and interspecies comparisons (between corresponding time points; $week_i$ versus $week_i$) with duplicates per condition. We identified 61K, 50K, 59K and 51K events that were significantly differentially spliced ($|dPSI| \geq 0.1$ and Q-value ≤ 0.05) in at least one temporal or interspecies comparison in chimpanzee, human, orangutan and rhesus macaque respectively. We observed that the majority of splicing changes were in interspecies comparisons (Figure 3D), with RI, MF, SE and AF constituting the most commonly differentially spliced event types (Figure 3C). Intriguingly, when we compared orthologous

event PSI values by event type between each primate and human across corresponding time points – as a proxy for conservation of splicing dynamics – we found that complex event types (CF, CL and CO) displayed the closest central tendency of PSI values to those of human cells (Figure 3E). This finding may lend credence to the value of characterizing complex event types and their involvement in primate neuronal differentiation.

junctionCounts uncovers novel splicing dynamics in genes relevant to neuronal differentiation and function

Among the 17K significant events ($|dPSI| \geq 0.1$ and Q-value ≤ 0.05 in at least one comparison) that were orthologous in all four primates (Figure 4A), we hypothesized that junctionCounts would both recapitulate previously reported splicing phenomena and identify novel events in genes involved in neuronal differentiation and function. Here, we highlight several such findings. Amphiphysin 1 (AMPH) and Amphiphysin 2 (BIN1) are both enriched in the mammalian brain and participate in synaptic vesicle endocytosis [96,97]. Splice variants of BIN1 have been reported in the brain as well as other tissue types (43), but the implications of AS in AMPH1 remain unexplored. We report a SE event involving exon 17 of AMPH1 (Figure 4B), which is the only scenario of AS that affects the CDS among AMPH1 isoforms annotated in GENCODE V44. According to Exon Ontology [92], AMPH1 exon 17 encodes an intrinsically unstructured polypeptide region which contains an O-phospho-L-serine modification site. This exon is increasingly spliced in over the time course with species-specific trajectories and magnitudes, possibly indicating a functional role for AMPH1 exon 17 inclusion in neurons.

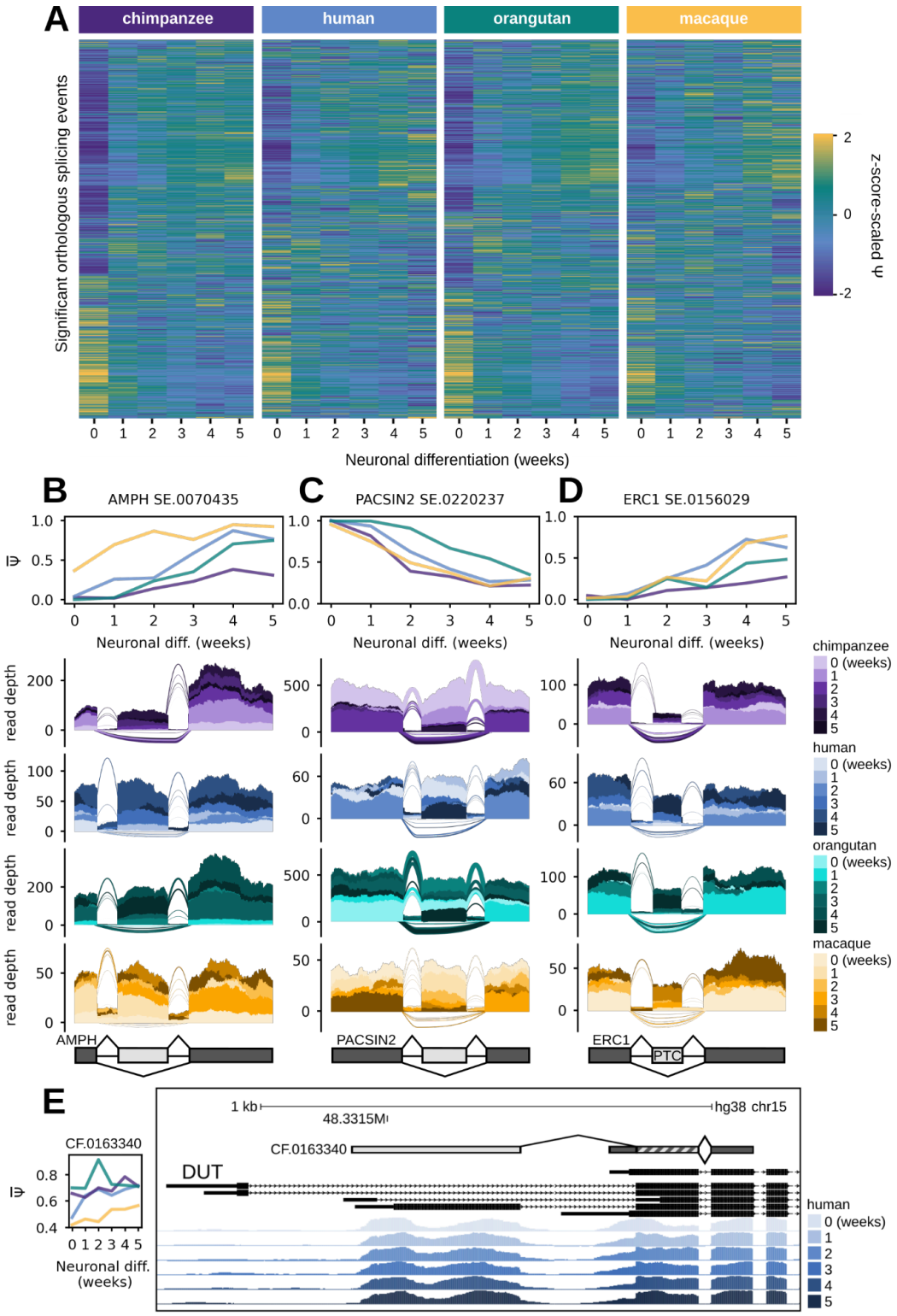


Figure 4. junctionCounts uncovers conserved and species-specific temporal splicing patterns among orthologous splicing events across the four primates. (A) Heatmap of Z-score-scaled PSI values for significant orthologous splicing events (|dPSI| ≥ 0.1 and

Q-value ≤ 0.05 in at least 1 temporal or interspecies comparison) with each row corresponding to the same event across all four primates. (B) Mean PSI trajectories and RNA-seq coverage at a skipped exon event in AMPH with species-specific temporal splicing patterns across chimpanzee, human and rhesus macaque. (C) Mean PSI trajectories and RNA-seq coverage at a skipped exon event in PACSIN2 with a conserved temporal splicing pattern. (D) Mean PSI trajectories and RNA-seq coverage at an PTC-containing skipped exon event in ERC1 with a conserved temporal splicing pattern. (E) UCSC Genome Browser snapshot of human read support at a complex first exon event in DUT, which measures the inclusion of one of several distal alternative first exons and its subsequent second exon versus the proximal first exon which overlaps with the alternative second exon. The subpanel to the left shows the mean PSI of the included form at each time point of neuronal differentiation in each primate. In panels (B), (C) and (D) the included form of the alternative event contains both the dark and light gray components, while the excluded form only contains the dark gray components. In panel (E), the same is true except the included form does not contain the dark gray fragment at the 5' end of the central exon. In panels (B), (C), (D) and (E), the upright and inverted arches represent junction read coverage for the included and excluded form respectively.

Protein kinase C and casein kinase II substrate in neurons 2 (PACSIN2) is the only known member of the PACSINs whose expression isn't cell type-specific, in humans. All three PACSINs have been reported to play a role in trafficking AMPA receptors in and out of synapses, which is a crucial factor in important neuronal processes including synaptic transmission and plasticity [98,99]. We identified a SE event involving exon 9 of PACSIN2 for which the included form uniformly decreases from the dominant to the minor form over the course of neuronal differentiation (Figure 4C). Similarly to the aforementioned SE event in AMPH1, this SE event is the only CDS-altering event among PACSIN2 isoforms annotated in GENCODE V44. According to the GTEx V8 RNA-Seq Read Coverage by Tissue track on the UCSC Genome Browser [100], PACSIN2 exon 9 inclusion is dominant in all non-neuronal tissue types, while the excluded form is dominant in 9 out of 14 neuronal tissue types. These observations make a compelling case for neuron-specific AS of PACSIN2, resulting in the preferential exclusion of exon 9 in several neuronal cell types.

We discovered a conserved ERC1 PTC-inducing SE event involving exon 18 in isoform ENST00000355446.9 (in GENCODE V44), that to our knowledge has not been previously

reported by other groups (Figure 4D). Inclusion of this exon may produce an NMD substrate, but could potentially yield a functional protein isoform at least 30 residues shorter at the C-terminus relative to isoforms consistent with the excluded form. ERC1 has been described to undergo neuron-specific AS and is implicated in important functions including neurotransmitter release and neuronal differentiation [101,102]. Over the five week course of neuronal differentiation, the PTC-inducing SE event follows a consistent pattern of becoming increasingly spliced in across the four primates. Interestingly, AS at the C-terminus of Erc1 in rats was shown to generate two isoforms: Erc1a and Erc1b. The latter of which is the brain-specific, shorter isoform that alone can bind to presynaptic active zone proteins, called RIMs, that regulate neurotransmitter release [103]. Taken together, these observations suggest a potential functional role for the inclusion of the ERC1 poison exon in differentiating neurons.

Deoxyuridine 5′-triphosphate nucleotidohydrolase (DUT) is an important enzyme involved in genome integrity maintenance that prevents uracil misincorporation into DNA. DUT expression has been shown, through knockout studies, to be essential to embryonic development and especially to later stages of differentiation in mice [104]. We identified a CF event in DUT (Figure 4E), in which the included form corresponds to the DUT-M isoform and the excluded form corresponds to the DUT-N isoform [105]. The DUT-M isoform localizes to mitochondria via a mitochondrial targeting presequence located in the first exon consistent with the included form of the CF event and is expressed constitutively. The DUT-N isoform localizes to the nucleus and its expression is induced during the G_0 to S phase transition. Exit from the cell cycle into G_0 phase triggers DUT-N protein degradation. Thus, DUT-N isoform expression is tightly linked to nuclear DNA replication [105]. We observed that the CF event is increasingly spliced in – meaning DUT-M isoform expression gradually eclipses DUT-N isoform expression over the time course – which is to be expected as the

primate pluripotent stem cells (PSCs) progressively commit to neuronal cell fates with decreasing cell cycle activity [106]. These findings demonstrate that junctionCounts can handily uncover novel splicing phenomena.

Temporal regulation of alternative splicing directs the transition from pluripotent to neuronal cell fate

High levels of AS and cell type-specific isoform expression are observed in neurons and during neuronal differentiation [58,107]. We postulated that genes exhibiting dynamic temporal splicing would be enriched for neuronal biological pathways. Taking the subset of significant AS events ($|dPSI| \geq 0.1$ and Q-value ≤ 0.05 in at least one temporal comparison), we generated the four most distinct clusters of events based on the Euclidean distance of their temporal PSI trajectories using CLARA [90] for each species (Figure 5A). Additionally, we identified subsets of genes with conserved (mean $|dPSI| \leq 0.1$ for all events per gene) and nonconserved (mean $|dPSI| \geq 0.3$ for all events per gene) splicing patterns in chimpanzee, orangutan and rhesus macaque relative to human in pairwise comparisons at each time point (Figure 5B). We then used Metascape [91] to identify enriched biological pathways in the sets of genes from each cluster of temporally regulated events (Figure 5C) and for the conserved and nonconserved splicing gene sets (Supplementary Figure 2).

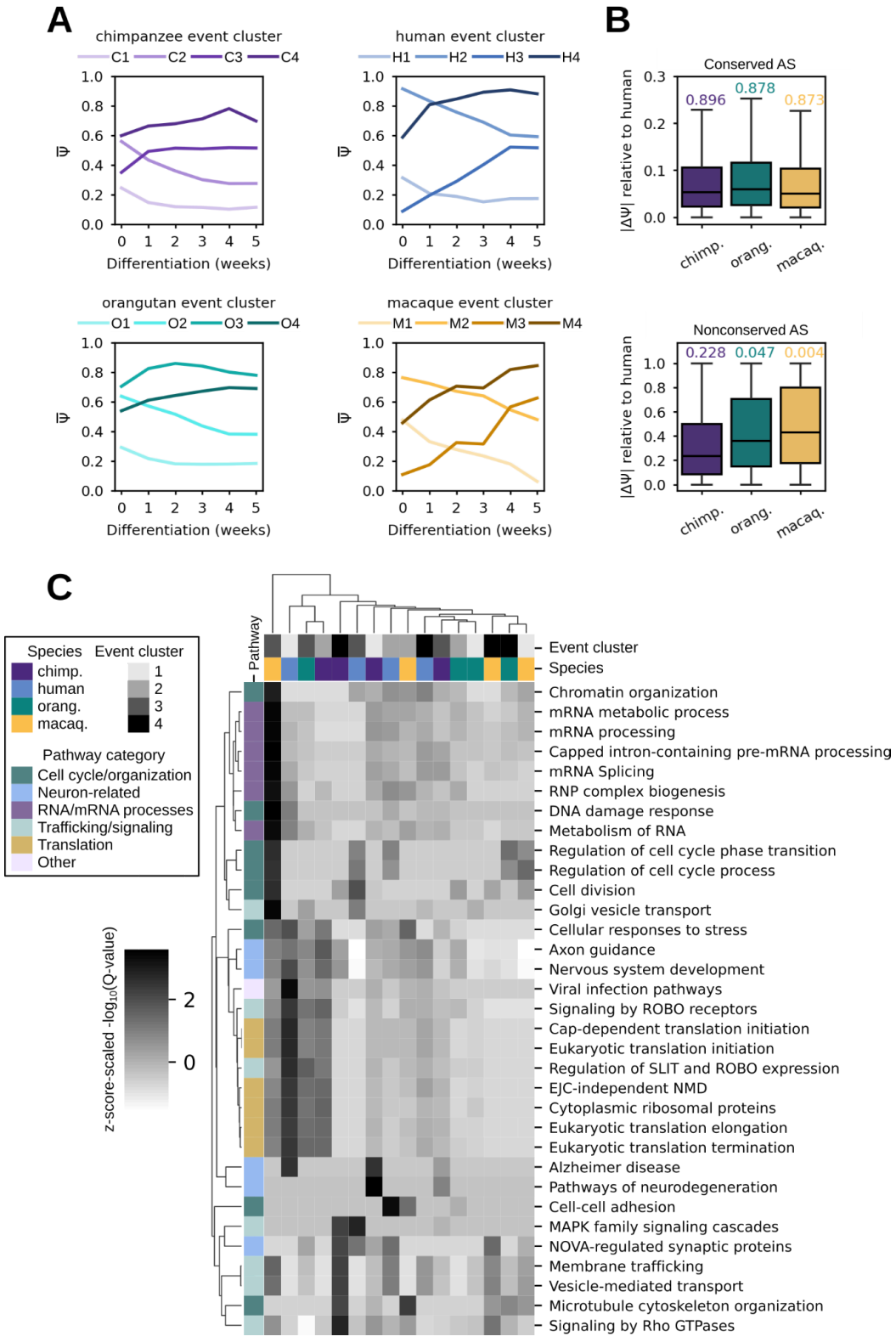


Figure 5. Gene ontology analyses for grouped event sets by temporal PSI trajectories.

(A) Four gene-level clusters derived with CLARA from temporal expression trajectories for each species. (B) Genes with mean |dPSI| ≤ 0.1 for events in chimpanzee, orangutan and rhesus macaque relative to human were categorized as genes with conserved splicing. Boxplots showing the distribution of each primate's per-gene mean |dPSI| relative to human (upper subpanel) with Pearson correlation coefficient of species-specific PSI values against human PSI values above them. The same for genes with nonconserved splicing based on mean |dPSI| ≥ 0.3 (lower subpanel). (C) Heatmap of Z-score-scaled-$\log_{10}$(Q-value) of Metascape (38) pathway enrichment in temporal event clusters.

Four similar but distinct event clusters were identified in each primate. Across the primates, cluster 1 (chimpanzee, human, orangutan and rhesus macaque corresponding to C1, H1, O1 and M1 respectively) generally represents alternative features (exons, introns, splice sites, etc.) whose inclusion is marginal in PSCs and declines over the course of differentiation. Cluster 2 (C2, H2, O2 and M2) represents alternative features whose inclusion is dominant in PSCs and declines during differentiation. Pathways similarly enriched in clusters H1, M1 and C2 indicate the preferred exclusion of particular alternative features in the mature splicing program of genes related to: translation, NMD, axon guidance and nervous system development. Cluster 3 (C3, H3, O3 and M3) generally contains alternative features whose inclusion is marginal in PSCs and increases during differentiation, while cluster 4 (C4, H4, O4 and M4) contains dominantly included alternative features that further increase until peaking at week 4 during differentiation. Clusters H3 and O3 indicate increasing inclusion of alternative features in the mature splicing program of genes related to: NOVA-regulated synaptic proteins, axon guidance and nervous system development. Cluster C3 indicates a slight increase in alternative feature inclusion in genes related to mRNA processing and translation. The conserved splicing event set was enriched for genes in critical pathways including cell cycle processes, signaling, AS, and interestingly, in neurodegeneration pathways. The subset of complex conserved splicing events (CF, CL and CO) was enriched for nearly all the same pathways, revealing the prevalence of complex events in important pathways (Supplementary Figure 2). For example, we identified a conserved CO event in

NCKAP1, which is involved in Rho GTPase signaling, and a conserved CL event in QKI, which is involved in pre-mRNA processing and AS (Supplementary Figure 3). Taken together, these results highlight the intricate temporal regulation of splicing as PSCs develop into neuronal cells and shed light on the biological relevance of species-specific and conserved splicing dynamics.

Emergent alternative features underlie many instances of species-specific alternative splicing

Because we observed that some events had miniscule or zero PSI values in particular species, we hypothesized that a subset of the aforementioned nonconserved event set represents events that sufficiently map (sequence divergence ≤ 20%) pairwise between all four primate genomes but contain alternative features that are only used (included) by specific primates despite the apparent presence of splice site and branch site sequences. We call features used by specific species emergent alternative features, potentially indicating exonization events. Indeed, we identified 3753 events, in 1922 genes, exhibiting significant temporal regulation ($|dPSI| \geq 0.1$ and Q-value ≤ 0.05 in at least one temporal comparison) while having a $\min(PSI) \geq 0.05$ in only a subset of the four primates (Figure 6A). Of these events, the most prominent event types were SE, MF, AF and ML (Figure 6B). Protection of Telomeres 1 (POT1) is an example of a rhesus macaque-specific SE event in the 5'UTR (Figure 6C). This SE event is likely an instance of species-specific differences in exon induction related to neuronal differentiation, as the alternative exon is annotated in GENCODE V44 and exhibits cell type-specific expression in a number of human neuronal tissues according to GTEx V8 RNA-Seq Read Coverage by Tissue despite its lack of inclusion in our human samples. Transmembrane Protein 165 (TMEM165) is an example of a human-specific PTC-containing SE event that is potentially the product of Alu exonization [108], as it overlaps an antisense

AluJb element (Figure 6D). Furthermore, one piece of evidence that suggests that it may be a bona fide emergent alternative cassette exon is that the rhesus macaque genome sequence has an A-to-G mutation 3 nt upstream of the 3'SS while the other three primates have a canonical 3'SS dinucleotide (Figure 6E). In short, identification of AS events in orthologous sequences between species may be an effective approach to uncover potential emergent alternative features.

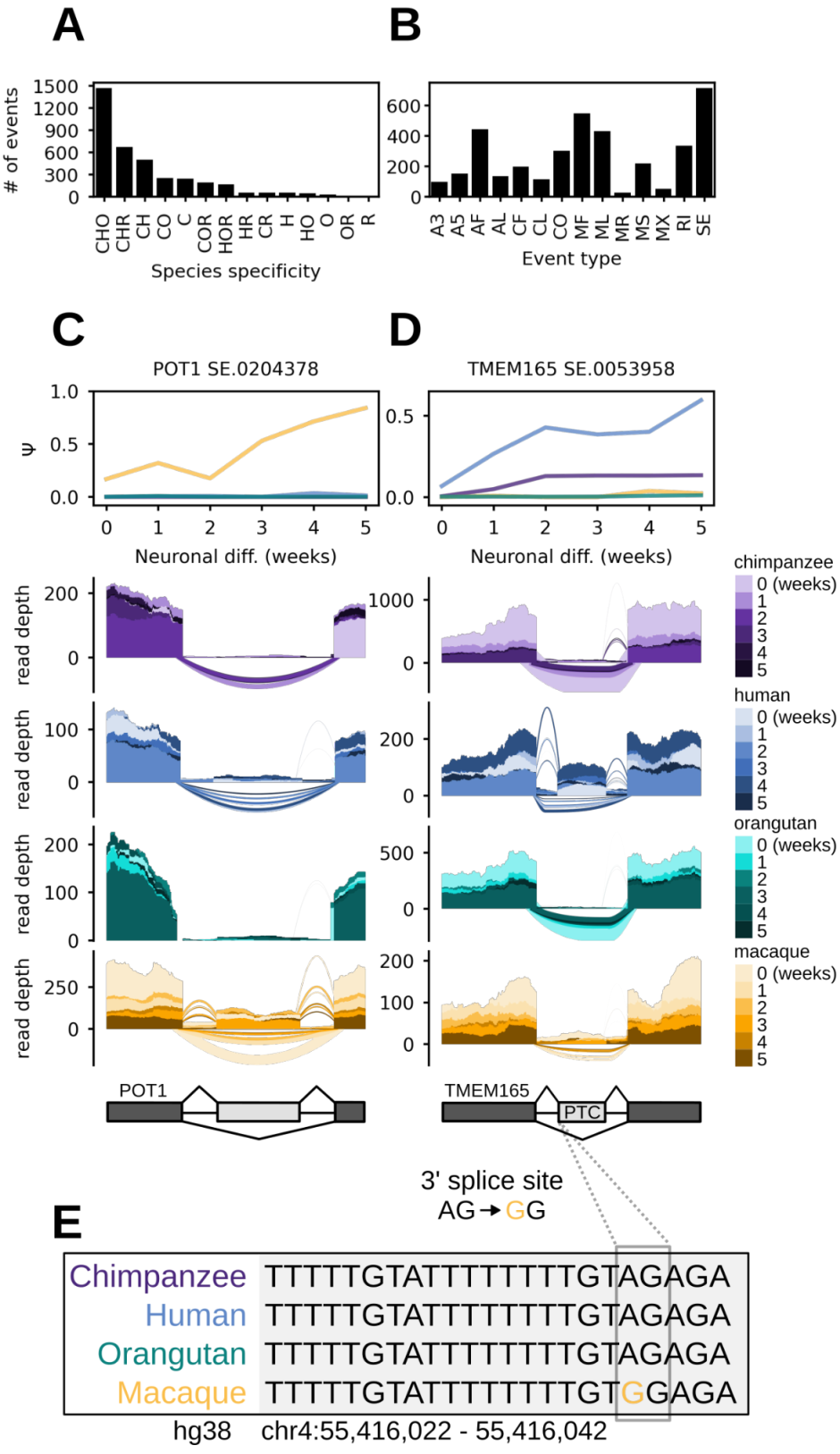

A
B
# of events
1500
1200
900
600
300
0
CHO
CHR
CH
CO
C
COR
HOR
HR
CR
H
HO
O
OR
R
Species specificity
600
400
200
0
A3
A5
AF
AL
CF
CL
CO
MF
ML
MR
MS
MX
RI
SE
Event type
C
D
POT1 SE.0204378
TMEM165 SE.0053958
ψ
1.0
0.5
0.0
0
1
2
3
4
5
Neuronal diff. (weeks)
read depth
chimpanzee
human
orangutan
macaque
0 (weeks)
POT1
TMEM165
PTC
3' splice site
AG→GG
E
Chimpanzee TTTTTGTATTTTTTTTGTAGAGA
Human TTTTTGTATTTTTTTTGTAGAGA
Orangutan TTTTTGTATTTTTTTTGTAGAGA
Macaque TTTTTGTATTTTTTTTGTGGAGA
hg38 chr4:55,416,022 - 55,416,042

Figure 6. Emergent/species-specific alternative feature usage. (A) Barplot showing the number of species-specific events (min(PSI) ≥ 0.05 in a given species) exhibiting significant temporal regulation (|dPSI| ≥ 0.1 and Q-value ≤ 0.05 in at least 1 temporal comparison). “C”, “H”, “O” and “R” are abbreviations for chimpanzee, human, orangutan and rhesus macaque respectively. Combinations of these abbreviations represent instances of events that meet the min(PSI) threshold in a set of species and are significantly temporally regulated in at least 1 species in the subset. (B) Barplot displaying the same set of species-specific events as in (A) but stratified by event type instead of species. (C) Mean PSI trajectories and RNA-seq coverage at a rhesus macaque-specific skipped exon event in POT1. (D) Mean PSI trajectories and RNA-seq coverage at a human-specific skipped exon event in TMEM165. (E) Macaque-specific point mutation just upstream of the 3’SS of the TMEM165 skipped exon event shown in (D).

cdsInsertion and findSwitchEvents connect alternative splicing events to potential functional impacts

An enduring problem in the study of AS is the challenging nature of connecting events to functional impacts, whether at the mRNA or protein level. We used cdsInsertion to annotate transcripts with information regarding the lengths of the UTRs and CDS, the presence of potential PTCs and other details gleaned from overlapping annotated start codons. Next, we employed findSwitchEvents to couple isoform-level CDS information to junctionCounts-defined events to identify “switch events”, which are instances in which a particular property is exclusive to transcripts consistent with the included or excluded form. We propose that this approach enables users to connect AS events to functional outcomes, comprehensively profile switch event regulation and to discover novel instances of NMD/NSD.

Among significant events (|dPSI| ≥ 0.1 and Q-value ≤ 0.05 in at least one comparison), we identified hundreds of events predicted to confer NMD, NSD and coding-to-noncoding switches as well as >1600 CDS-altering events in each species (Figure 7A). To investigate the potential structural and functional impacts of CDS-altering events, we mapped event coordinates to protein features with Exon Ontology [92]. The five protein feature categories

most frequently overlapping alternative exons were: post-translational modification (PTM), structure, binding, localization and catalytic activity (Figure 7B). Interestingly, intrinsically disordered regions (IDRs) were the most highly represented feature (Figure 7C). In agreement with these findings, IDRs have been described as preferred loci for both AS and PTMs [109,110].

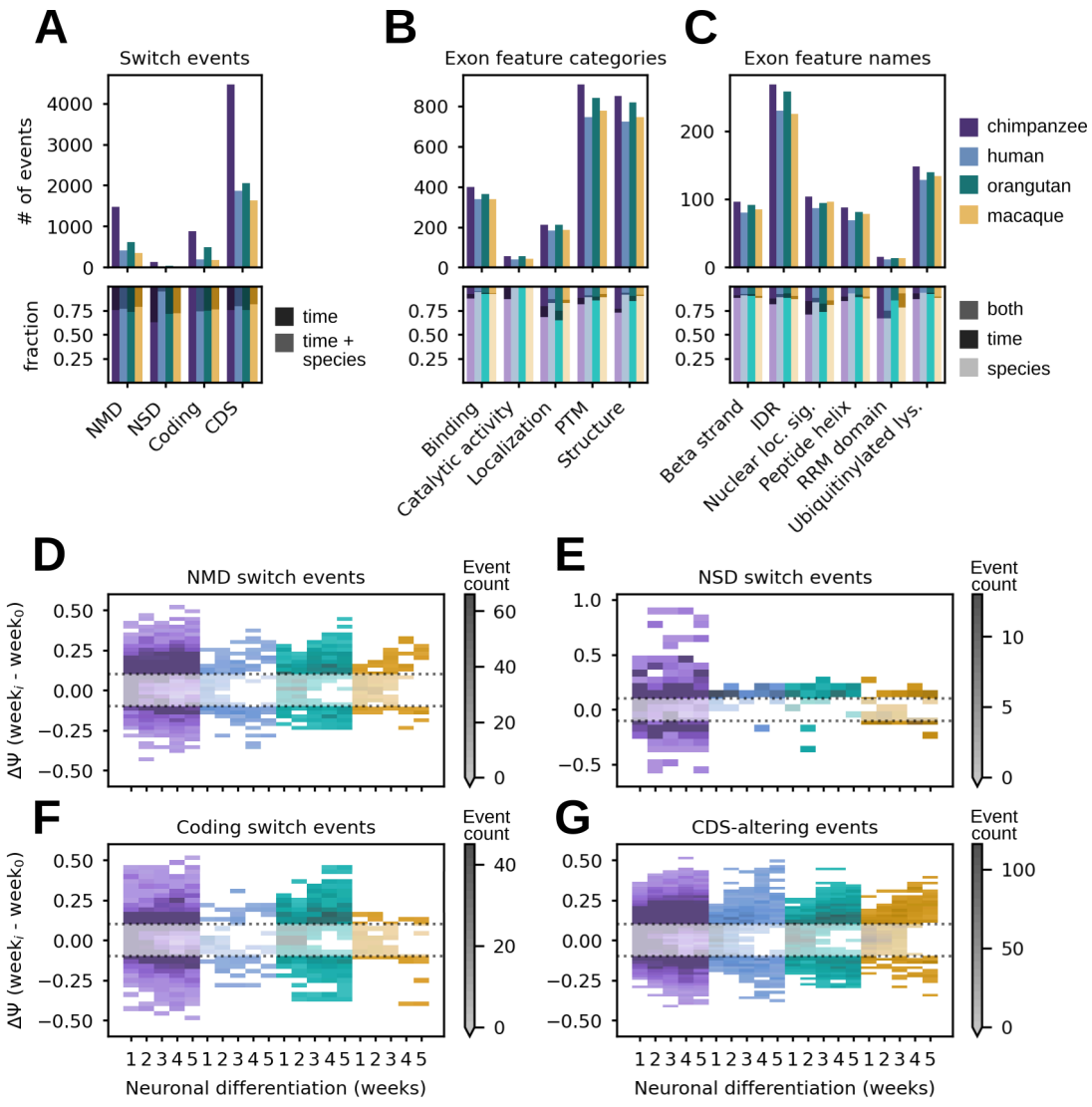


Figure 7. Analysis of splicing events with potential functional impacts to mRNA stability, coding capacity or protein function. (A) Total number of significant splicing events ($|dPSI| \geq 0.1$ and Q-value ≤ 0.05 in at least 1 temporal or interspecies comparison) that are predicted to: induce nonsense-mediated decay (NMD), induce non-stop decay (NSD), break the open reading frame (coding switch) or alter the coding sequence (CDS). (B) Exon

feature category support for significant event coordinates overlapping exon features in the Exon Ontology database. (C) Specific exon features affected by significant events, indicating potential impacts to structural or functional protein elements. The fraction of events in (A), (B) and (C) that were significantly different across species, time or both factors (lower subpanels). (D) Bivariate histogram showing the distribution of NMD switch event dPSIs over the time course of neuronal differentiation relative to pluripotent stem cells (week 0) in each primate. (E), (F) and (G) show the same for NSD switch, coding switch and CDS-altering events, respectively.

At the mRNA-level, NSD and NMD substrates are expected to be degraded via translation-dependent pathways to prevent the production of potentially harmful truncated proteins [111]. However, in certain contexts expression and translational activation of NMD substrates can play important roles in biological functions, including in neuronal differentiation [112]. Another possible outcome of AS is the generation of noncoding transcripts from protein coding genes [113]. We characterized the regulation of these three phenomena during neuronal differentiation and found that NMD and coding-to-noncoding switch events became progressively more differentially spliced over the course of neuronal differentiation relative to PSCs (Figure 7D-F). NSD switch events were relatively rare, but they were surprisingly overrepresented in chimpanzee relative to the other primates (Figure 7E). Overall, we did not observe a monotonic increase in NMD substrate abundance during neuronal differentiation and we found that NMD/NSD/coding-to-noncoding switch events generally followed species-specific temporal trajectories (Supplementary Figure 4). CDS-altering events were also increasingly differentially spliced over the course of differentiation (Figure 7G), likely owing to the gradual definition of a mature splicing program that requires cell-type specific expression of isoforms with distinct functions [114].

DISCUSSION

This paper describes our efforts to develop an accurate, rigorous, easy to use and interpret alternative splicing analysis tool capable of identifying and quantifying a comprehensive repertoire of splicing event types, with the addition of novel capabilities. junctionCounts accurately recapitulates ground truth PSI and dPSI values from simulated data and performed well in our benchmarking experiment against MAJIQ [79], rMATS-turbo [80], splAdder [81] and Whippet [71] (Figure 2, Supplemental Figure 1). It identifies and accurately quantifies a wide array of event types including complex event types that represent coordinated splicing of multiple alternative features. In contrast to MAJIQ, rMATS-turbo, splAdder and Whippet, junctionCounts identifies events from a gene annotation alone, eliminating the need to generate new splice graphs/event dictionaries for each individual dataset. Contrarily, the other tools use information from mapped RNA-seq data during splice graph generation, so event identification scales directly with library read depth and can include novel splice junctions. Novel splice junction detection can be achieved with junctionCounts by providing a transcriptome assembled from RNA-seq reads using a tool like StringTie [67]. Most uniquely, junctionCounts, in concert with its partner utilities: cdsInsertion and findSwitchEvents, couples events to their effects on the isoform-level CDS to predict functional consequences including NMD. We show through the analysis of published UPF1-knockdown [78] and Emetine [50] human RNA-seq data that junctionCounts accurately predicts NMD switch events (Figure 2L-N).

After rigorously testing junctionCounts and its partner utilities, we applied them to a primate neuronal differentiation RNA-seq dataset comprising human and rhesus macaque ESCs as well as chimpanzee and orangutan iPSCs [51]. We identified 50-61K significant splicing events ($|dPSI| \geq 0.1$ and Q-value ≤ 0.05 in at least one temporal or interspecies comparison) in each species, with 17K significant orthologous events across all four primates (Figure 4A).

Within these orthologous events, junctionCounts recapitulated previously reported splicing phenomena [105] and identified previously unreported events in several genes relevant to neuronal differentiation (Figure 4B-E). RT-PCR experiments were used to verify some of these events, including SE events in GABBR1 and MYCBP2 in human and macaque cells (Supplementary Figure 5). We additionally clustered events by their temporal splicing dynamics, uncovering distinct event trajectories that capture the tight regulation of splicing during development (Figure 5A). Highly relevant biological pathways were represented in these event clusters, including axon guidance, nervous system development and SLIT/ROBO signaling. Pathways in chromatin organization, mRNA processing and cell cycle processes were also shown to undergo and potentially underlie splicing regulation (Figure 5C). Within the set of events with nonconserved splicing patterns, we uncovered thousands of events containing alternative features that were used in some primates but not in others, suggesting potential emergent alternative features (Figure 6). Lastly, we used cdsInsertion and findSwitchEvents to connect events to predicted NMD/NSD/coding-to-noncoding switches based on isoform-level CDS properties exclusive to their included or excluded forms. This allowed us to profile temporal NMD/NSD regulation (Figure 7D-E) and to identify potential NMD substrates (Figure 4D and 6D). Altogether, we exhibited the functionality of junctionCounts in a variety of analysis contexts and presented its application to the characterization of splicing in evolution, neuronal differentiation and NMD.

Besides this work, we further note that several of our colleagues have already implemented and published results using a beta version of junctionCounts. These studies include a variety of model systems such as human and non-human primate cell lines, *C. elegans*, and yeast. Suzuki et al. (2022) looked at the effects of KIN17 and PRCC mutations on 5’ and 3’SS usage during development in *C. elegans*. They found both direct and potentially indirect changes in alternative 5’ and 3’SS usage, some of which were related to developmental and

population dynamics. They additionally RT-PCR-verified a number of these events to differentiate between embryonic-type splicing and somatic-type splicing [115]. Cartwright-Acar et al. (2022) characterized splicing changes in the presence of class II suppressors of uncoordination in an *unc-73(e936)* mutant forward genetic screen in *C. elegans*. They found that the majority of alternative 5'SS usage changes were in introns containing true alternative 5'SS and that suppressors rarely activated novel cryptic alternative 5'SS. They further RT-PCR verified several of the alternative 5'SS and 3'SS events, and finally asserted that the class II suppressors they studied may work at mutually exclusive stages of spliceosome assembly or use different mechanisms to maintain 5'SS identity based on their ability to differentiate between alternative 5' splicing events that are unique to particular suppressors [116]. Draper et al. (2023) quantified events across polyribosome fractions and between primates to assess the conservation of alternative splicing coupled to translational control (ASTC). They identified subsets of alternative events with either conserved or species-specific sedimentation profiles and discovered that alternative exons with conserved sedimentation had higher sequence conservation relative to those with species-specific sedimentation. They additionally tested three ASTC SE events using translational luciferase reporters [117]. Hunter et al. (2023) examined the effect of splicing inhibitors on intron splicing efficiency in *S. cerevisiae*. They found that individual introns had distinct sensitivities, including during co-transcriptional splicing, to different splicing inhibitors. Interestingly, they found that yeast sequences including the branch point consensus motif contribute to the differences in sensitivity [118]. Osterhoudt et al. (2024) explored changes in 3'SS usage upon SACY-1 perturbation in introns with pairs of 3' splice sites ≤ 18 nucleotides away from each other. They found that both SACY-1 depletion and a SACY-1 mutation lead to a clear unidirectional increase in proximal alternative 3'SS usage, which they RT-PCR-verified for several events [119]. Collectively, our collaborators found

junctionCounts easy to implement and show, through these works, its capacity to generate high quality results.

Beyond its flexibility and user-friendliness, junctionCounts stands out as a useful approach because it identifies both canonical and non-canonical alternative events. Many tools are limited to non-terminal and/or relatively rudimentary event types. The few that characterize complex or non-canonical event types are difficult to interpret. junctionCounts utilizes the concept of binary alternative events; identifying clear instances of the inclusion and exclusion of alternative features. This concept is well-established and pervades the splicing literature. It remains popular because binary alternative events can be accurately quantified relative to full-length transcripts, they likely accurately represent (a subset of) transcript structure as compared with full-length transcripts, and they exclude gene segments not relevant to the regulation of the event (i.e. introns and exons outside and distal to the event). The first two reasons will likely decrease in validity as improving long-read sequencing approaches provide more accurate representations of the ground-truth expressed transcriptome. The biggest problem may lie with the third reason, considering that the contribution of factors not necessarily local to an event itself can be important to its regulation [120,121].

At present, however, focusing on the local site of alternative events affords the opportunity to consider the behavior of hundreds or thousands of similar events and to look for trends in features that may explain their behavior. However, a key weakness of the traditional binary event is the existence of loci in which more than two alternative sub-transcripts overlap and are subject to simultaneous changes in relative abundance. While such non-binary events could be represented as the collection of binary events involving all possible pairs of sub-transcripts, this representation loses information as the regulatory decision is likely to be made in the context of all possibilities. A number of efforts such as MAJIQ [79] and Whippet

[71] have attempted to address this issue with several approaches. Nonetheless, junctionCounts presents a step in the right direction by characterizing events that don't fit into canonical binary event definitions.

Lastly, junctionCounts' main innovation lies in its ability to bridge the gap between event-level and isoform-level analysis with regard to the implications of AS events on transcript coding and translational capacity, via cdsInsertion. In an ideal case, studies that intend to consider translation and its implications on a transcriptome-wide scale would include an experimental technique to empirically define CDS regions or start codons. For example, ribosome profiling and approaches like TI-seq [122] can serve as a basis for empirically defining whole CDS or translation start sites respectively. However, as such data are typically unavailable due to the additional cost and complexity of these approaches, tools like cdsInsertion are useful. cdsInsertion fleshes out the putative characteristics of unannotated transcripts by performing *in silico* translation from known overlapping start codons, and thus permits the development of hypotheses to explain properties imparted by AS. Its partner tool, findSwitchEvents, infers alternative event characteristics from those of its constituent isoforms. Altogether, junctionCounts, cdsInsertion and findSwitchEvents comprise a method for the accurate characterization of AS and the novel capacity to couple events to potential functional outcomes.

DATA AVAILABILITY

RNA-seq data from mouse cerebellum and liver cells [83] used to generate simulated data for the benchmarking experiment: publicly available at the NCBI GEO (Accession no.: GSE54652).

RNA-seq data from spliceostatin A and DMSO-treated HeLa cells [78] used to generate simulated data for the benchmarking experiment: publicly available in the ArrayExpress database (Accession no.: E-MTAB-6060).

RNA-seq data and NMD target RT-PCR data from UPF1 siRNA and non-targeting siRNA-treated HEK-293 cells [93] used to validate cdsInsertion and findSwitchEvents NMD predictions: publicly available at the NCBI GEO (Accession no.: GSE176197).

RNA-seq data from emetine and DMSO-treated HEK-293 T cells [50] used to validate cdsInsertion and findSwitchEvents NMD predictions: publicly available at the NCBI GEO (Accession no.: GSE89774).

RNA-seq data from human and rhesus macaque ESCs as well as chimpanzee and orangutan iPSCs [51]: publicly available at the NCBI GEO (Accession no.: GSE106245).

The version of junctionCounts and partner utilities used in this study are published on Zenodo (https://doi.org/10.5281/zenodo.11186192). junctionCounts and its partner utilities are also available on GitHub: https://github.com/ajw2329/junctionCounts and https://github.com/ajw2329/cds_insertion.

SUPPLEMENTARY DATA

Supplementary data are available at NAR online.

ACKNOWLEDGEMENTS

We thank Sol Katzman, Julia Philipp and Jolene Draper for discussion on junctionCounts development and implementation. We thank Timothy Sterne-Weiler for discussion of benchmarking studies and feedback on the manuscript.

FUNDING

This work was supported by NIH grant GM130361 (JRS).

CONFLICT OF INTEREST

None declared.

Chapter 4: Alternative Splicing Coupled with Translational Control (ASTC)

4.1 Chapter Introduction

The project presented in this chapter reports our work to develop a new approach to study translational control utilizing long read sequencing to capture complete, ribosome-associated transcript structures. The impetus of this work is the phenomenon of transcripts exhibiting distinct ribosome association (sedimentation) profiles, most clearly exemplified by instances of isoform-specific sedimentation within genes. The motivation to incorporate LR RNA-seq with subcellular fractionation was to directly measure unabridged transcript sequences within the system to: 1) assign short reads to observed transcripts with high confidence, 2) discover novel transcripts, and 3) precisely characterize transcript features and termini. We not only found that isoform-specific sedimentation was widespread, but that it was largely consistent between stem cells and neuronal progenitor cells. Using machine learning approaches, we further discovered transcript features that were to a large extent predictive of sedimentation profiles.

The conceptualization, design and bench execution of this translatomic approach, *LR Frac-seq*, was carried out by Drs. Jeremy Sanford and Jolene Draper with the help of Dr. Chris Vollmers. I contributed all bioinformatic analyses involved in integrating LR and short read Frac-seq data, data visualization, and writing and editing of the manuscript. This manuscript was accepted by *Genome Research* on May 21st, 2024.

4.2 Long read subcellular fractionation and sequencing reveals the translational fate of full length mRNA isoforms during neuronal differentiation

Alexander J Ritter[1º], Jolene M Draper[2º] , Chris Vollmers[1] and Jeremy R Sanford[2]*

ºThese authors made equal contributions
[1] Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, USA
[2] Department of Molecular, Cell and Developmental Biology and Center for Molecular Biology of RNA, University of California Santa Cruz, Santa Cruz, CA, USA

* To whom correspondence should be addressed. Tel: 831-459-1822; Fax: 831-459-3139
Email: jsanfor2@ucsc.edu

ABSTRACT

Alternative splicing (AS) alters the cis-regulatory landscape of mRNA isoforms leading to transcripts with distinct localization, stability and translational efficiency. To rigorously investigate mRNA isoform-specific ribosome association, we generated subcellular fractionation and sequencing (Frac-seq) libraries using both conventional short reads and long reads from human embryonic stem cells (ESC) and neural progenitor cells (NPC) derived from the same ESC. We performed *de novo* transcriptome assembly from high-confidence long reads from cytosolic, monosomal, light and heavy polyribosomal fractions and quantified their abundance using short reads from their respective subcellular fractions. Thousands of transcripts in each cell type exhibited association with particular subcellular fractions relative to the cytosol. Of the multi-isoform genes, 27% and 19% exhibited significant differential isoform sedimentation in ESC and NPC respectively. Alternative promoter usage and internal exon skipping accounted for the majority of differences between isoforms from the same gene. Random forest classifiers implicated coding sequence (CDS) and UTR lengths as important determinants of isoform-specific sedimentation profiles, and motif analyses reveal potential cell type-specific and subcellular fraction-associated RNA-binding protein signatures. Taken together our data demonstrate that alternative mRNA

processing within the CDS and UTRs impacts the translational control of mRNA isoforms during stem cell differentiation, and highlights the utility of using a novel long read sequencing-based method to study translational control.

INTRODUCTION

Accurate eukaryotic gene expression requires messenger RNA (mRNA) assembly from precursor transcripts. Protein coding and regulatory sequences (exons) are distributed across expansive precursor messenger RNA transcripts. The spliceosome excises intervening non-coding sequences (introns) from pre-mRNA and ligates the exon sequences together to generate translation-competent mRNA [123]. Conserved sequences at exon-intron boundaries (splice sites) direct spliceosome assembly on each newly synthesized intron. Remarkably, the spliceosome can assemble different combinations of exon sequences to generate mRNA isoforms from a common pre-mRNA transcript [124,125]. Alternative splicing (AS) not only generates isoforms with distinct protein coding potential, but also with different post-transcriptional regulatory capacity. For example, AS decisions that introduce premature termination codons induce nonsense mediated decay while other splicing events generate transcripts with distinct subcellular localization or translational control. In addition to generating alternative isoforms with unique coding sequences (CDS), AS can produce isoforms that differ only in their untranslated regions (UTRs). Elements in the UTRs of mature mRNA play pivotal roles in post-transcriptional regulation. In the 5' UTR, regulatory sequences like upstream open reading frames (uORFs) and internal ribosome entry sites (IRES) influence translation initiation efficiency [126–128]. The 3' UTR contains various elements such as microRNA binding sites and RNA-binding protein (RBP) recognition sites that modulate mRNA stability, localization, and translation [129,130]. Regulatory elements in the CDS can also influence the fate of mRNAs. For example the RBP, HuR, stabilizes target mRNAs by binding to AU-rich elements (AREs) within the CDS, preventing their

degradation. Conversely, RBPs like TTP can promote mRNA degradation by binding to AREs in coding regions, leading to mRNA decay [131]. Proteins like IGF2BP1 can bind to coding region instability determinants in the CDS of target mRNAs to enhance their stability [132]. By and large, AS confers complex and multidimensional consequences to the fate of mRNAs through shaping the *cis*-regulatory landscape of alternative isoforms [133–135].

Importantly, there is poor correlation between steady-state mRNA and protein levels in eukaryotic systems [136–140]. And while factors like mRNA stability and translation initiation efficiency play a role in this disparity, the influence of AS on translational control is often overlooked. A number of methods exist to study translational control, which is the regulatory mechanism in eukaryotic cells that governs the efficiency and timing of protein synthesis from mRNA. A well-established method called Ribo-seq offers genome-wide insights into ribosome occupancy and translation dynamics by capturing single nucleotide-resolution ribosome footprints, but it can be vulnerable to artifacts and signal biases [141]. RNC-seq captures ribosome nascent-chain complex-bound mRNAs to characterize the translatome, but it doesn't provide ribosome footprints or ribosome density information [142]. TRAP-seq utilizes epitope-tagged ribosomes to enable cell type-specific translation profiling, which generates data similar to RNC-seq which can be modified to produce ribosome footprints, but it relies on transgenic models and may not fully replicate endogenous ribosome behavior [143,144]. Frac-seq, which our proposed method builds on, isolates actively translating ribosomes and assesses translation efficiency by stratifying transcripts by the number of ribosomes they are associated with [145]. However, it has the potential for selective bias toward highly abundant transcripts and it lacks single-nucleotide resolution of ribosome positions on mRNA. While each method has its strengths and weaknesses, one shared disadvantage is that they all involve the sequencing of short mRNA fragments from ribosome-protected or ribosome-associated mRNAs.

Short read RNA-sequencing methods struggle to accurately capture the complete structures of complex RNA isoforms [146]. In contrast, long read RNA-sequencing provides full-length reads that span entire transcripts, enabling precise characterization of intricate isoforms and annotation-agnostic detection of novel structures. The primary shortcoming of long read sequencing is its relatively lower throughput compared to short read sequencing platforms, limiting the depth of coverage for a given budget. To address this limitation and to maximize the benefits of both long read and short read methods, we employed a complementary approach. Here we introduce the development of long read Frac-seq to obtain full-length transcript isoforms with intact records of ribosome association, structural variation, and long-range interactions. We complement this data with short read Frac-seq to compensate for the loss of throughput and to provide a more complete and accurate representation of the translated transcriptome.

RESULTS

Characterization of a transcriptome supplemented with long read-derived novel transcripts

To investigate the relationship between alternative pre-mRNA splicing and isoform-specific mRNA translation we capitalized on the capability of long read sequencing to capture complete transcript structures of polyribosome-associated mRNA, without sacrificing throughput by generating both long read and short read Frac-seq libraries [145]. We used human embryonic stem cells and neural progenitor cells (ESC and NPC respectively) as a model system to characterize the translated transcriptome during early neuronal differentiation. The resulting samples were the cytosol, monosome, light polyribosome (2-4 ribosomes), and heavy polyribosome (≥5 ribosomes) fractions (Figure 1A). By utilizing the R2C2 method [147], our long read libraries, with mean read length 2 Kb and mean library

size 500 K, were reinforced with improved base calling accuracy (93%) and with high-confidence transcript starts and ends. Fractionation of the long reads was employed to enhance the likelihood of detecting transcripts which may be preferentially associated with distinct subcellular fractions. All long read libraries were pooled for *de novo* transcriptome assembly using Mandalorion [148], followed by rigorous quality control, filtering, and functional annotation using all three modules of the Functional IsoTranscriptomics analysis suite [149]. The resulting long read-derived transcriptome was then merged with GENCODE's GRCh38.p13 Release 41 primary assembly annotation[150] to account for transcripts that weren't captured by long read sequencing. The following analyses were done in the context of this "comprehensive transcriptome" containing both annotated and long read-derived novel transcripts.

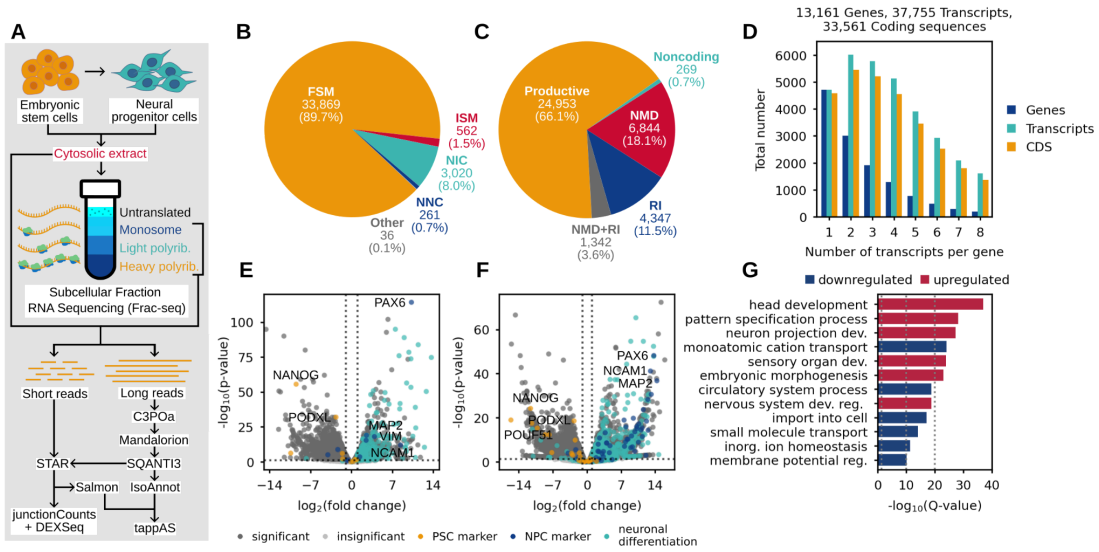


Figure 1. Experimental overview and characterization of the comprehensive transcriptome and the cytosol. (A) Schematic of the experiment and subsequent bioinformatic analysis workflow of the resulting cytosolic extract and fractionated, ribosome-associated short and long reads from ESC and NPC. (B) The transcriptome classified by SQANTI3-defined structural categories of spliced transcripts, including: Full Splice Match (FSM), Incomplete Splice Match (ISM), Novel in Catalog (NIC), Novel Not in Catalog (NNC) and intergenic or fusion transcripts (Other). FSM and ISM transcripts match

annotated splice sites and junctions (in GRCh38.p13 Release 41), while NIC transcripts comprise novel combinations of annotated splice sites and junctions and NNC transcripts contain at least one unannotated splice site. (C) The transcriptome classified by productivity based on the detection of complete or incomplete open reading frames (productive or noncoding respectively), premature stop codons (NMD) and retained introns (RI). (D) Stratification of the transcriptome by the number of isoforms and unique coding sequences per gene. (E-F) Gene-level (E) and transcript-level (F) differential expression between NPC and ESC cytosolic fractions. (G) Top 12 enriched Metascape pathways in differentially expressed genes between NPC and ESC cytosolic fractions.

The comprehensive transcriptome had short read coverage meeting a 1 count per million reads (CPM) cutoff for 37,755 transcripts with 33,561 unique coding sequences (CDS), arising from 13,161 genes (Figure 1D). Of these, 5,875 and 4,590 transcripts from 4,095 and 3,176 genes were uniquely expressed in ESC or NPC respectively. Transcripts were organized into SQANTI3-defined structural categories based on their fidelity to transcript structures in the GRCh38.p13 Release 41 primary assembly annotation [150]. 91.2% of transcripts matched the annotation, 8.7% were considered novel (containing either novel combinations of known splice sites and junctions or at least one novel splice site), and 0.1% were categorized as either genic or fusions (Figure 1B). Additionally, transcripts were categorized based on their productivity. We define productive transcripts as those encoding a full-length, canonical protein. Unproductive classes include: noncoding (lacking a complete open reading frame), nonsense-mediated decay (NMD) and retained intron (RI). 66.1% of transcripts were considered productive, 0.7% were predicted to be noncoding, 18.1% were classed as NMD based on the presence of a premature termination codon (PTC), 11.5% had a retained intron (RI) and the remaining 3.6% met both NMD and RI conditions (Figure 1C).

We used Salmon [84] to pseudoalign the fractionated short reads, with an average library size of 71.5 M reads, to the comprehensive transcriptome; producing transcript-level quantification across the gradient. Using the cytosolic fraction, which represents the raw output of the nucleus, we next tested the baseline transcriptomic differences in NPC relative to ESC at the gene-level (Figure 1E) and at the transcript-level (Figure 1F) to reveal

upregulation of NPC and neuronal differentiation markers and downregulation of pluripotency markers. Metascape [91] pathways further encapsulated these observations (Figure 1G). Taken together, these results present the framework for an approach to integrate fractionated long and short reads to study translational control at isoform-level resolution.

Thousands of transcripts exhibit distinct association with particular subcellular fractions

To discover if mRNA transcripts have distinct ribosome association profiles, we clustered transcript-level expression trajectories across the gradient using tappAS [149]; revealing subpopulations of transcripts with clear enrichment in one subcellular fraction over the others (Figure 2A). Interestingly, a subpopulation of transcripts enriched in both the monosome and light polyribosome fractions stood out as one of the most populous subsets: making up about 30% of transcripts with enrichment in subcellular fractions in both cell types. Thus, subsequent analyses class monosome-associated transcripts (Mono) and light polyribosome-associated (LPR) transcripts as those that are exclusively enriched in those fractions, leaving the set of monosome *and* light polyribosome-associated (M+L) transcripts as a standalone subpopulation. Thousands of transcripts were considered significantly enriched ($\log_2FC \geq 1.0$, p-value ≤ 0.05) in subcellular fractions relative to the cytosol (Figure 2B). Overall, 7.5% and 6.8% of transcripts were significantly associated with a subcellular fraction in ESC and NPC respectively. In the context of non-mutually exclusive enrichment in subcellular fractions, subpopulations of transcripts were generally dissimilar across fractions within cell type, with the exception of the Mono and LPR fractions with Jaccard similarity of 0.41 and 0.33 in ESC and NPC respectively (Supplemental Figure 1), due to the substantial M+L transcript subpopulations. When stratified by productivity, Mono- and LPR-associated transcript subpopulations exhibited pronounced incorporation of unproductive classes relative to the cytosol. Heavy polyribosome-associated (HPR) transcripts displayed a slight reduction

of unproductive classes relative to the cytosol in ESC, while an increase is observed in NPC (Figure 2C). These findings support the hypothesis that levels of ribosome association may correlate with levels of translatability.

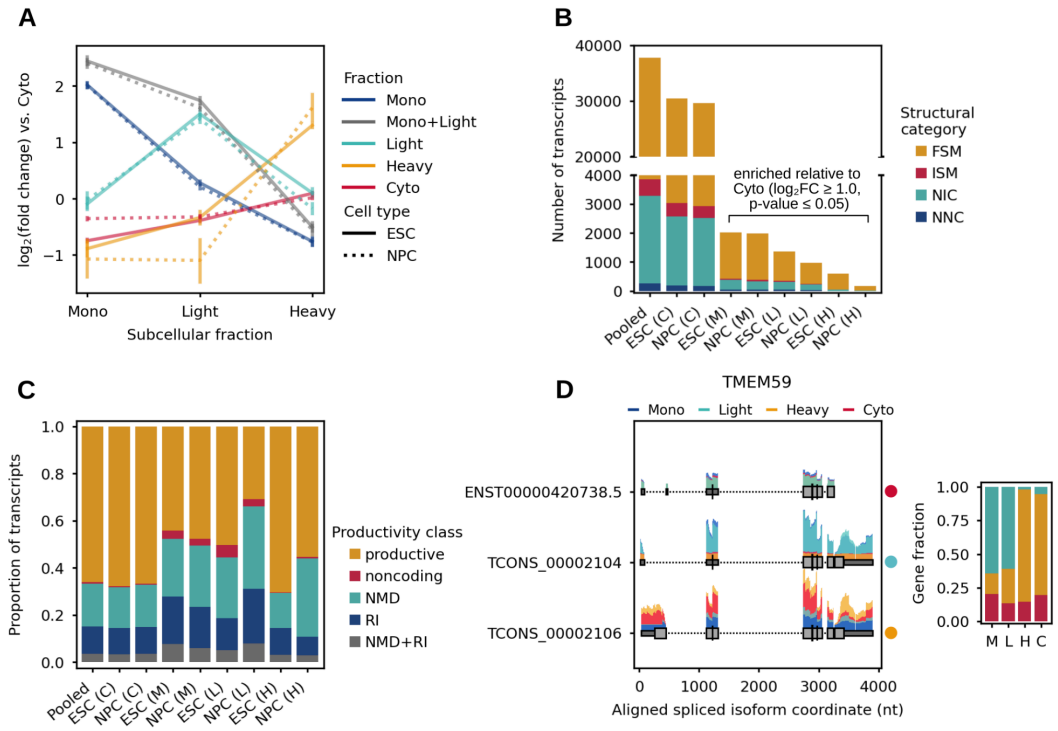


Figure 2. Establishing transcript ribosome association profiles. (**A**) Clustering of transcripts by their expression trajectories across the gradient relative to the cytosol. (**B**) Extraction of fraction-associated transcripts based on significant enrichment ($\log_2$FC ≥ 1.0, p-value ≤ 0.05) in the Mono, LPR or HPR fractions relative to their abundance in the cytosol. "C", "M", "L" and "H" represent the cytosol, Mono, LPR and HPR fractions respectively. (**C**) Categorization of fraction-associated transcripts by productivity. (**D**) Differential sedimentation of three isoforms in TMEM59. Above spliced isoform models, histograms of short read support at exons are colored by fraction. The stacked barplot summarizes the proportion of total gene expression each isoform contributes in each subcellular fraction.

Alternative splicing confers functional consequences to the stability and translation of mRNAs

Because we observed transcript subpopulations with distinct ribosome association profiles, we postulated that alternative mRNA isoforms may likewise sediment discretely. To test this hypothesis, we calculated the expression of individual isoforms relative to all isoforms from

the same gene and measured the difference between their gene fractions in subcellular fractions relative to those in the cytosol. TMEM59 is an example of a gene with three isoforms, two of which exhibit differential sedimentation in ESC (Figure 2D). TMEM59 expression in ESC is composed of a M+L-associated isoform, a HPR-associated isoform and a cytosol-associated (not differentially sedimenting) isoform. Interestingly, endogenous post-transcriptional silencing of TMEM59 by miR-351 in murine neural stem cells has been implicated to promote neuronal differentiation [151], although the two differentially sedimenting isoforms share all but the last base of their 3' UTRs. But along similar lines, it may be the case that *cis*-regulatory differences in their 5' UTRs and CDS influence the isoform-specific nature of their sedimentation.

We found 3,321 (26.5%) and 2,254 (19.2%) genes in ESC and NPC respectively exhibiting differential isoform sedimentation ($|\Delta \text{ Gene fraction}| \geq 0.1$ and Q-value ≤ 0.05) in a subcellular fraction relative to the cytosol (Figure 3A). Within those genes, 4,906 and 3,229 transcript isoforms preferentially sedimented (Δ Gene fraction ≥ 0.1) with a subcellular fraction in ESC and NPC respectively. These instances substantiate alternative splicing as an architect of isoform-specific translational control. We observed decreasing concordance of gene fraction changes across the gradient between cell types, with a Pearson correlation coefficient of 0.59, 0.54 and 0.09 in Mono, LPR and HPR respectively. In fact, 3085 transcripts in 1506 genes exhibit divergent patterns of isoform sedimentation (Supplemental Figure 2). Together, these findings suggest that isoform-specific sedimentation is likely cell type-specific, possibly owing to differences in the composition and environment of *trans*-acting factors.

Additionally, Figure 3A illustrates that isoforms associated with the Mono and LPR fractions have a greater magnitude of gene fraction differences than the HPR fraction relative to the

cytosol. This finding suggests that the Mono, LPR and HPR fractions gradate toward increasing similarity in transcript abundance and isoform ratios with the cytosol, which is consistent with findings from other groups [152]. Considering these results, we posit that the average number of ribosomes per mRNA in the cytosolic fractions of our ESC and NPC samples may be similar to that of the HPR fraction. Among genes displaying differential isoform sedimentation, pathways involved in chromatin organization, organophosphate biosynthesis and phosphorylation were enriched in ESC specifically, while DNA damage, stress response and cell cycle pathways were enriched in NPC (Figure 3B). Genes demonstrating divergent isoform sedimentation across cell types were enriched in similar pathways, with the addition of RNA metabolism. When comparing the gene ontology of cognate subpopulations of subcellular fraction-associated transcripts across cell types, we observed that cell cycle, translation and mRNA processing-related pathways were consistently represented, while cell type-specific pathway enrichment was much more apparent in the cytosolic fraction (Supplemental Figure 3).

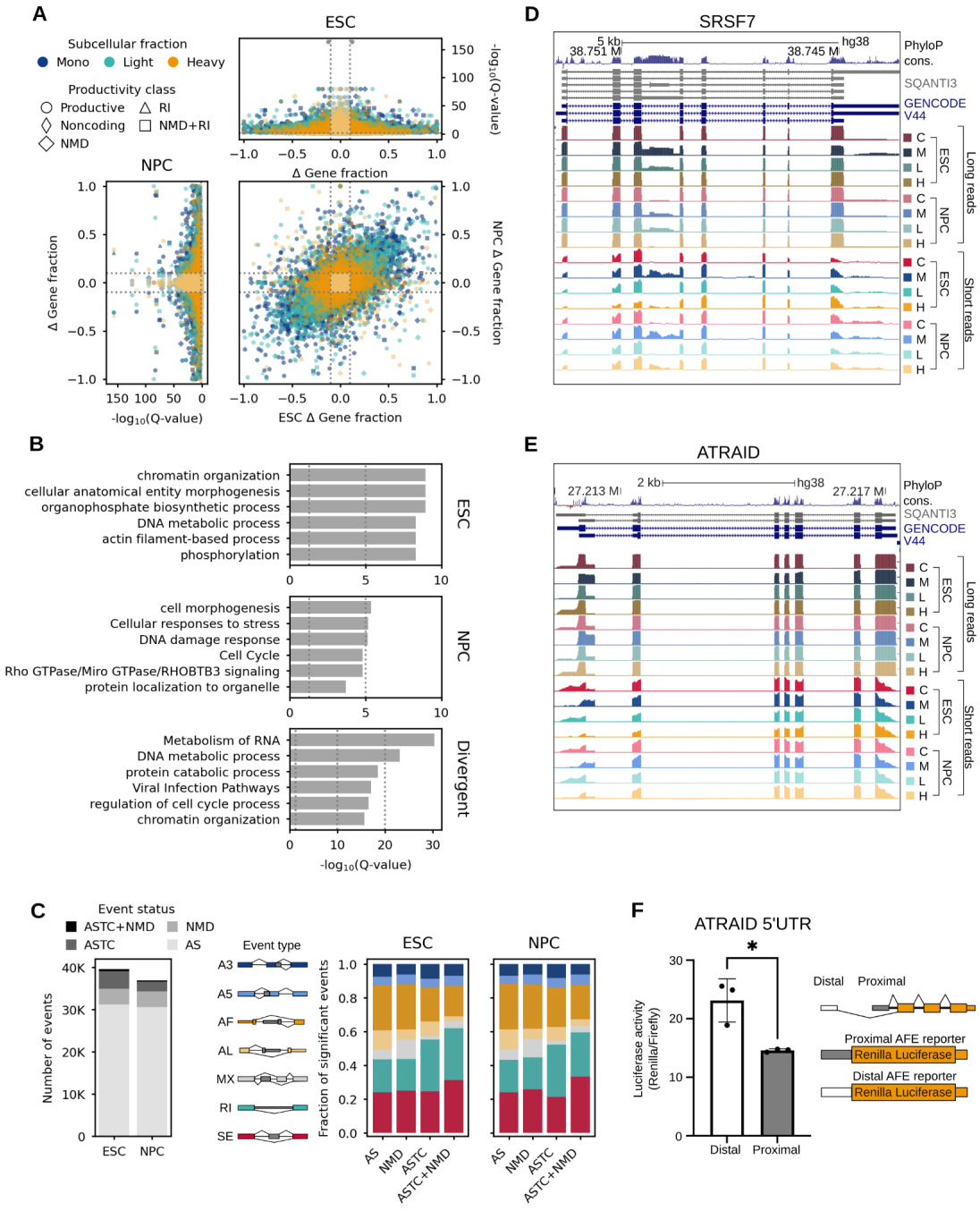


Figure 3. Differential isoform sedimentation across the gradient and functional outcomes of alternative splicing. (**A**) Volcano plots representing differential isoform sedimentation by changes in isoform gene fraction relative to the cytosol. 3,321 and 2,254 genes in ESC and NPC respectively exhibit significant differential isoform sedimentation (|Δ Gene fraction| ≥ 0.1, Q-value ≤ 0.05). The central plot shows changes in isoform gene fraction, with Q-value ≤ 0.05, of isoforms expressed in both ESC and NPC. (**B**) The first two bar plots show the top six enriched Metascape pathways in genes containing significant instances of differential isoform sedimentation exclusively in ESC or NPC. The third bar plot,

labeled “Divergent”, depicts enriched Metascape pathways in genes displaying contrasting patterns of isoform sedimentation between ESC and NPC. (**C**) The first stacked bar plot categorizes significant alternative splicing events ($|\Delta\Psi| \geq 0.1$, adjusted p-value or Q-value ≤ 0.05) as: alternative splicing (AS), alternative splicing coupled with translational control (ASTC), meaning splicing events that are differentially included across the gradient, NMD, and alternative splicing coupled with both translational control and nonsense-mediated decay (ASTC+NMD). The following two bar plots show the breakdown of event types comprising each category in ESC and NPC. (**D-E**) UCSC Genome Browser snapshot of long read and short read coverage at (**D**) SRSF7, exhibiting subcellular fraction-associated inclusion of a conserved retained intron, and at (**E**) ATRAID, exhibiting subcellular fraction-associated alternative first exon usage. “C”, “M”, “L” and “H” represent the cytosol, Mono, LPR and HPR fractions respectively. (**F**) Luciferase assay measuring the translational impact of using either the distal or the proximal ATRAID 5’ UTR in HEK-293 cells.

To examine the types of alternative splicing (AS) that give rise to the diversity of the transcriptome, we categorized AS events as: AS ($0.1 \leq \Psi \leq 0.9$, adjusted p-value ≤ 0.05 within condition), ASTC ($|\Delta\Psi| \geq 0.1$, Q-value ≤ 0.05 across subcellular fractions), NMD (events that introduce a PTC) and ASTC+NMD (NMD events that adhere to the mentioned cutoffs for significant ASTC events) (Figure 3C). Notably, 11.8% and 7.0% of significant AS events were classified as either ASTC or ASTC+NMD in ESC and NPC respectively. We found that alternative first exon, retained intron, and skipped exon events feature most prominently among ASTC events, while skipped exons and retained introns comprise the majority of ASTC+NMD events. 2,456 and 1,257 CDS-altering events (A3, A5, MS, MX and SE event types) and 526 and 359 terminal events (AF and AL event types) were linked to translational control (either ASTC or ASTC+NMD) in ESC and NPC respectively. Because of the mentioned similarity between the HPR fraction and the cytosol, the majority of ASTC/ASTC+NMD events were Mono (79.1% in ESC, 62.1% in NPC) and LPR-associated (42.5% in ESC, 54.8% in NPC).

In our dataset, SRSF7 presents one complete and one partial retained intron event associated with NMD via induction of a PTC in the highly conserved SRSF7 intron 3 locus, which has been previously described to contain a conserved poison exon [153,154] (Figure 3D). Preferential association of PTC-containing isoforms with the Mono, and modestly with the

LPR fraction, is consistent with our understanding of NMD's effect on translation [155–157]. ATRAID (also known as APR3), a relatively poorly understood gene implicated to play roles in all-*trans* retinoic acid-induced apoptosis, osteoblast differentiation and some cancer types [158,159], demonstrates marked patterns of alternative first exon usage across the gradient. The proximal first exon of ATRAID was preferentially spliced into the Mono-associated isoform, which may indicate its reduced translation. Indeed, a Renilla-firefly luciferase assay comparing Renilla incorporating either the proximal or the distal 5' UTR of ATRAID in HEK293 cells exhibited significant differences in fluorescence (Figure 3F). Interestingly, the two isoforms of ATRAID may be functionally different, as the distal first exon contains an upstream open reading frame which may encode a signal peptide with potential importance to its localization with lysosomes [159]. Collectively, these results demonstrate the widespread functional impacts of alternative splicing to the cytosolic fate of mRNAs.

Intrinsic features and *cis*-elements correlate with transcript polyribosome profiles

Given that AS defines the *cis*-regulatory landscape of mature mRNAs, we sought to identify intrinsic transcript features that may encode the underlying regulatory grammar of ASTC. We define intrinsic transcript features as measurements and functional elements that are native to the sequence of a spliced transcript. To extract the relative predictive weight of features on ASTC, we employed random forest classifiers (RFC) to perform feature selection. Across the transcriptome, we measured the length and GC-content of the transcript, CDS, and UTRs. Additionally, our feature set included 5' and 3' UTR motifs, uORFs, repeat sequences, coding capacity, codon frequency, the presence of PTCs and retained introns. Given these features, RFCs were assigned binary classification tasks to predict the correct subcellular fraction for transcripts between every combination of subcellular fraction-associated transcript subpopulations at an 80:20 train:test split using 300 estimators. From the results, we extracted

permutation feature importance, with 50 repeats, and found that the number of exons, length of the CDS and UTRs, and the GC-content of the UTRs were important features for our models to correctly classify transcript polyribosome profiles. We note, however, that while our RFC models outperformed unskilled models/chance levels, the combination of generally adequate receiver operating characteristic curves with suboptimal Precision-Recall performance highlights class imbalances, the need for more data points in each fraction, and further indicates the requirement of additional features beyond those included in this study (Supplemental Figure 4).

Our findings that CDS and 3' UTR length positively correlate with association with heavier polyribosome fractions is consistent with previous reports [152] (Figure 4A). This is not to be confused with ribosome density, which other groups have shown to be inversely correlated with CDS length [160,161]. Although a longer CDS can theoretically accommodate a greater number of ribosomes, the increased potential for incorporation of non-optimal or rare codons may trigger codon usage-dependent negative impacts to translation initiation and elongation [162]. Additionally, longer CDS and transcript lengths have been observed to be negatively correlated with translation initiation rates in the context of intrapolysomal ribosome reinitiation [163]. En masse, inference of ribosome association based on the CDS alone is likely too simplistic to make accurate predictions.

To more clearly understand changes in feature length that may impact ribosome association, we also measured the change in CDS, 5' UTR and 3' UTR length relative to the dominant cytosolic isoform among isoforms belonging to genes with differentially sedimenting isoforms (termed gene-linked isoforms). Distinctly, Mono- and LPR-associated isoforms displayed a clear signal of relatively shorter CDS and longer 3' UTR, while HPR-associated isoforms remained largely similar or equivalent to the dominant cytosolic isoform (Figure

4B). Relatively longer 3' UTRs in gene-linked isoforms are connected to strong effects on ribosome association, and isoforms with 5' UTRs ≥ 1000 nt in length have been observed to be relatively poorly ribosome-associated relative to their shorter 5' UTR-containing counterparts within the same gene [152]. These phenomena could be due, in part, to potential increased inclusion of *cis*-regulatory elements in UTRs including miRNA target sites, uORFs and iron-responsive elements which can negatively impact mRNA stability and translation. Our summary analyses of GC-content measurements yielded less clear patterns in relation to ribosome association (Figure 4B). Nonetheless, about 30% of isoforms preferentially sedimenting in lowly ribosome-associated fractions (Mono and LPR) exhibited decreasing 3' UTR GC-content relative to the dominant cytosolic isoform, which may support findings that relate lower 3' UTR GC-content to increased association with P-bodies and enhanced susceptibility to miRNA targeting [164]. The other 70% of lowly ribosome-associated isoforms showed the opposite characteristic regarding 3' UTR GC-content, which is concordant with reports that suggest an inverse relationship between 3' UTR GC-content and mRNA stability [165]. Overall, broad measurements like length and GC-content *were* predictive of ribosome association to some degree, but appear to lack the granularity required to definitively elucidate the mechanisms underlying instances of ASTC. Comparisons between cell types regarding feature measurements also reveals cell type-specific differences in trends that indicate further layers of complexity (Supplemental Figures 5-6).

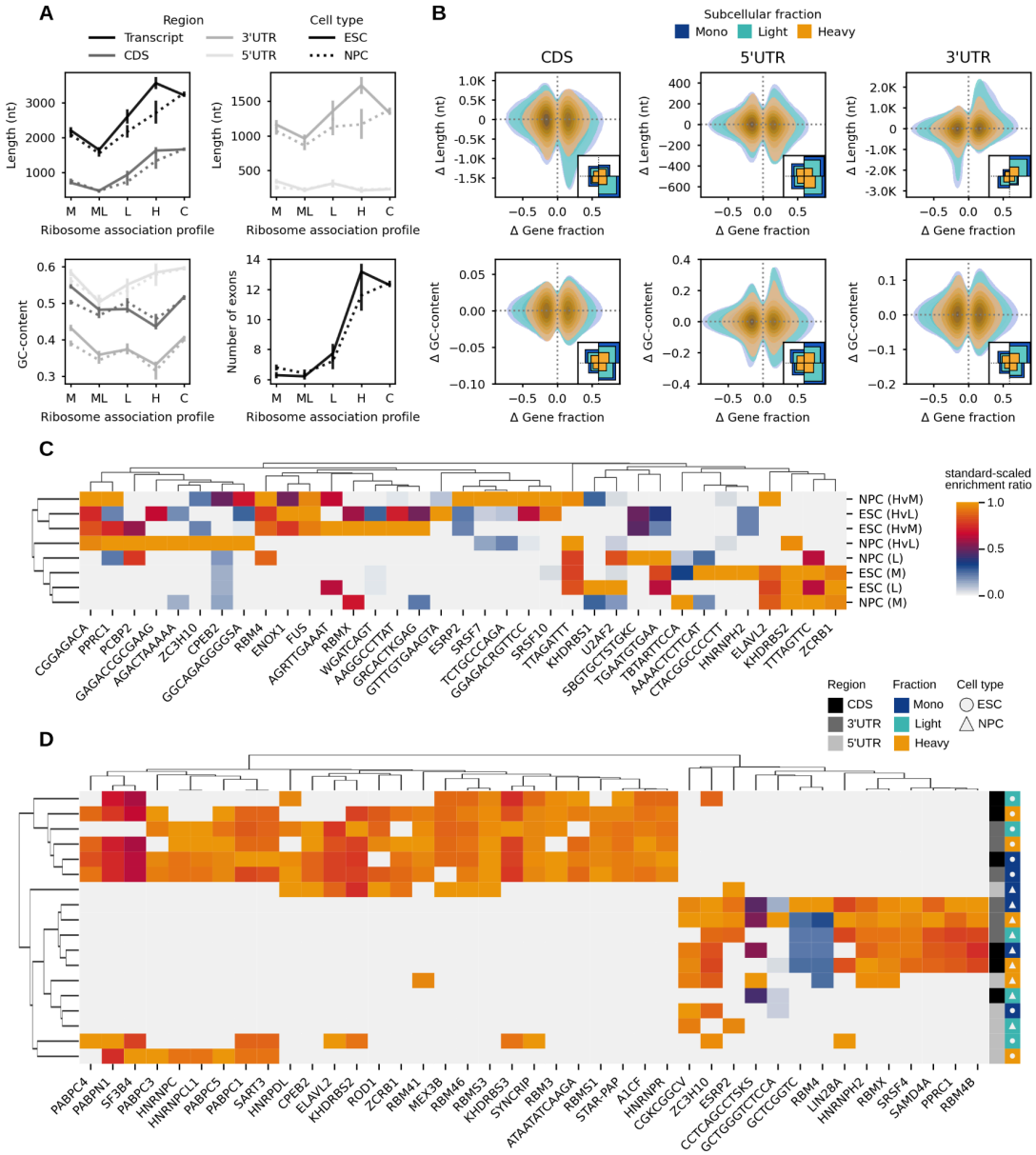


Figure 4. Analysis of features correlated with ribosome association profiles. (**A**) The number of exons and summaries of length and GC-content, with 90% confidence interval, of the CDS, 5' UTR and 3' UTR of transcripts associated with subcellular fractions. "M", "ML", "L", "H" and "C" represent the Mono, Mono+Light, LPR, HPR and cytosolic fractions respectively. (**B**) Measurement of the change in isoform gene fraction relative to the cytosol and differences in CDS, 5' UTR and 3' UTR length and GC-content of differentially sedimenting isoforms relative to the dominant isoform in the cytosol. Kernel densities for all coding isoforms are drawn with a 0.2 threshold. Subplots in the bottom left of each plot summarize the relative abundance of observations in each quadrant of their respective main plot, colored by fraction. (**C**) HOMER-derived *de novo* sequence motif and known RBP

motif enrichment ratios in skipped exons enriched in subcellular fractions versus skipped exons not enriched in each given fraction. “M” and “L” refer to the Mono and LPR-associated FASE sets, while “HvM” and “HvL” refer to the HPR-associated FASE sets relative to the Mono and LPR fractions respectively. (**D**) Enrichment of motifs using the same approach as in (**C**), but in 30 nt windows of the CDS, 5’ UTR and 3’ UTR of isoforms exhibiting divergent sedimentation profiles across cell types. The target sets were made from isoforms that preferentially sediment with each given subcellular fraction in one cell type, and the background sets were made from isoforms exhibiting the same in the other cell type. The standard-scaled enrichment ratio colorbar is shared by (**C**) and (**D**).

To look beyond length and GC-content measurements, we identified sequence motifs that are associated with subcellular fractions. To do this, we took fraction-associated skipped exons (FASE) – exons that were determined to be significantly enriched in a subcellular fraction ($\Delta\Psi \geq 0.1$, Q-value ≤ 0.05 across subcellular fractions) relative to the cytosol – and sliced them into 30 nt windows. Each set of windows was complemented with a background set consisting of windows made from skipped exons that were not significantly enriched in their given subcellular fraction. The HPR fraction was tested for enrichment against the Mono and LPR fraction (HvM and HvL, respectively) to produce HPR FASE sets due to a dearth of HPR-enriched exons relative to the cytosol. The resulting sets of FASEs included about 415 and 240 exons on average for each subcellular fraction in ESC and NPC respectively.

Using HOMER [166] on each set of windows, we discovered 111 *de novo* motifs, in total, that were significantly enriched (p-value ≤ 0.05, FDR ≤ 0.2) in the target sequences over their respective background sets. We used Tomtom [167] to identify the best matches (p-value $\leq$ 0.05) between the *de novo* motifs and known RBP motifs in the Ray 2013 Homo sapiens dataset [168]. We next combined the set of motifs with the CISBP-RNA *Homo sapiens* RBP motif set [168] and used SEA [169] to measure their enrichment (p-value ≤ 0.05, enrichment ratio ≥ 1.1) in each set of FASEs (Figure 4C). Several motifs exhibited enrichment in at least one fraction, with motif enrichment bissecting into clusters of Mono and LPR FASE sets, and HPR-associated FASE sets regardless of cell type.

We applied the same approach to the CDS, 3' UTR and 5' UTR of divergently sedimenting isoforms between ESC and NPC to identify cell type-specific motifs that may underlie the differences in their sedimentation. Target sequences were generated from isoforms preferentially sedimenting with each given fraction in one cell type versus those preferentially sedimenting with the same fraction in the other cell type. Strikingly, motif enrichment in these sets of sequences cluster more distinctly by cell type than by fraction, and most motifs are exclusively enriched in one cell type and not the other. We acknowledge, however, that motif analyses are limited by the fact that RBP binding specificities are often multivalent and difficult to predict. Nonetheless, we report the presence of statistically significant sequence motifs enriched in FASEs, and those that are differentially enriched and utilized between divergently sedimenting isoforms. Altogether, 19 of the 43 known RBPs identified as fraction- or cell type-specific have been previously implicated in translational control or observed to associate with polyribosomes (Supplemental Tables 10, 11) [170–180]. Additional studies are necessary to test the role of these potential factors in ASTC. Because we were specifically interested in motifs related to ribosome association, we did not perform motif analysis on introns. As a whole, these results suggest that ribosome association is impacted by the composition of intrinsic transcript features; likely with combinatorial effects.

DISCUSSION

Here, we report the first integration of long read RNA sequencing with a translatomic method, which we call LR Frac-seq, and we describe an approach to integrate long read and short read Frac-seq to characterize the translated transcriptome in human ESC and NPC. We took a complementary approach to capitalize on the major strength of long read sequencing in capturing complete transcript structures, while leveraging short read sequencing's significantly higher throughput for accurate quantification. Many examples of hybrid

sequencing approaches have previously been applied to complex biological problems by other groups (Reese et al. 2023; Puglia et al. 2020). For the long reads, we employed the R2C2 method to generate high-confidence consensus sequences with high base calling accuracy and well-defined transcript start and end sites. From these, we performed *de novo* transcriptome assembly to generate the set of full-length transcripts detected in the system, deemed the long read-derived transcriptome. Indeed, the long read-derived transcriptome does not comprehensively capture the entirety of the expressed transcriptome in ESC and NPC, as indicated by short read transcript-level mapping rates: on average, 87% of short reads mapped to the genome, while 42% mapped to the long read-derived transcriptome. The high quality of the short reads suggests that the lower transcriptomic mapping rate is due to the incomprehensive nature of the long read-derived transcriptome, which can likely be improved by deeper sequencing; ideally at 1 million or greater reads per long read library. To account for transcripts potentially missed by long read sequencing, we merged the long read-derived transcriptome annotation with GENCODE's GRCh38.p13 Release 41 primary assembly annotation [150] to produce a non-redundant, "comprehensive" transcriptome annotation for downstream analyses. The much deeper fractionated short read libraries were utilized to quantify the comprehensive transcriptome across the gradient, consisting of: the cytosol, monosome, light polyribosome (2-4 ribosomes), and heavy polyribosome (≥ 5 ribosomes) fractions. Highlighting one of the major benefits of long read sequencing, we found 3,281 transcripts with either novel combinations of known splice sites or ≥ 1 novel splice sites; accounting for 8.7% of the expressed (≥ 1 CPM) comprehensive transcriptome.

We compared transcript abundances in subcellular fractions to their cognate cytosolic fractions to identify transcripts with enrichment in particular fractions relative to the cytosol, which represents the raw output of the nucleus. We found that 7.5% and 6.8% of transcripts,

in ESC and NPC respectively, preferentially associate with subcellular fractions and that the proportion of productive transcripts associated with a given fraction directly correlates with ribosome association (Figure 5). Isoforms observed to preferentially sediment in subcellular fractions accounted for 13% and 9.8% of transcripts in multi-isoform genes. We trained RFCs to select features at the transcript level and we found that the number of exons, CDS, 5' UTR and 3' UTR length along with 5' UTR and 3' UTR GC-content were the most important features in our feature set for the accurate prediction of transcript polyribosome profiles in our dataset.

Among multi-isoform genes expressed in both ESC and NPC, gene-linked differences in isoform sedimentation relative to the cytosol were largely cell type-specific, although patterns of intrinsic transcript feature differences between fractions were similar between cell types (Supplemental Figures 2, 5 and 6). We found that isoforms with a shorter CDS and longer 3' UTR relative to the dominant isoform in the cytosol corresponded most clearly to Mono and LPR sedimentation. Additionally, motif analyses revealed potential RBP motifs in fraction-associated skipped exons and in divergently sedimenting isoforms. Interestingly, these motifs cluster by fraction and by cell type respectively (Figure 4C,D). In total, binding sites for 43 unique RBPs exhibited fraction-specific enrichment and nearly half (44%) have previously established roles in translational control or demonstrated association with polyribosomes (Supplemental Tables 10, 11). For example, proteomic analysis of polyribosomes revealed numerous splicing factors, including hnRNPC, SRSF10, and SRSF7 as polyribosome-associated [172]. Intriguingly, many of these factors have distinct sedimentation profiles across sucrose gradients, an observation that is consistent with the fraction-specific enrichment of RBP binding sites observed here. As a whole, our results present intrinsic feature measurements and potential RBP motifs that likely enact

combinatorial effects on translation, providing both previously reported and novel insights into the underlying mechanisms of ASTC. Because the most predictive intrinsic features were rather broad, we hypothesize that inter-isoform differences in length and GC-content more likely vaguely encapsulate changes to the isoform-specific *cis*-regulatory landscape. Several factors may affect an mRNA's translational output, including: intrinsic and *trans*-acting influences to mRNA stability, post-transcriptional modifications and combinatorial interactions with multiple RBPs. Therefore, it may be difficult to distill trends in transcript-level features across polyribosome fractions without also measuring transcriptome-wide mRNA half-life and capturing RBP-mRNA interactions, for example.

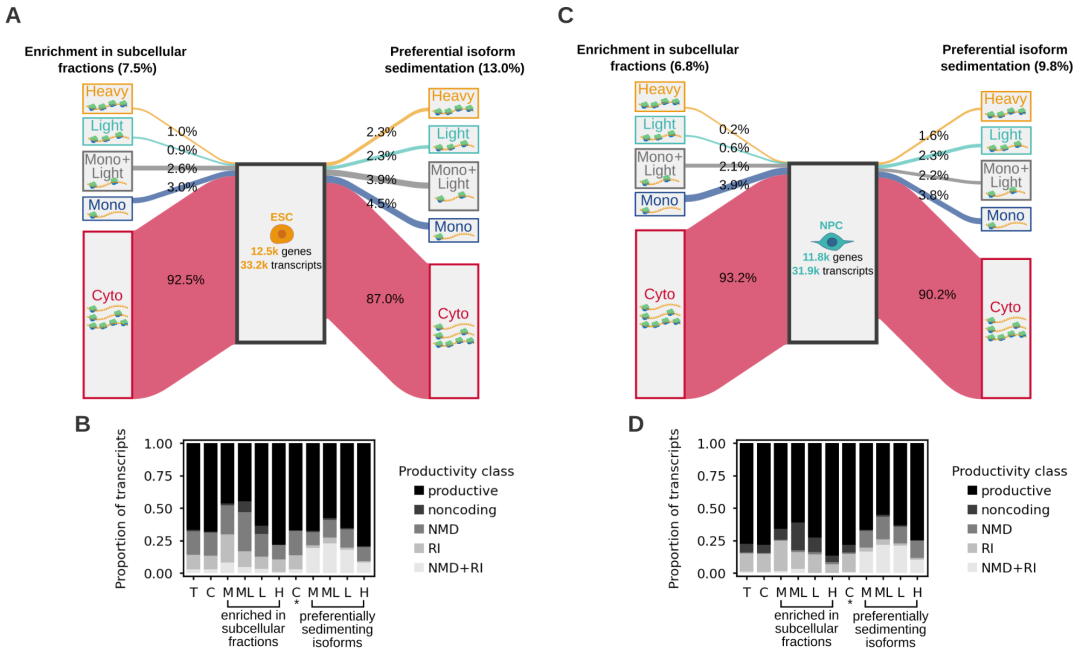


Figure 5. Summary of transcriptomic ribosome association profiles. (**A, C**) We identified thousands of transcripts in ESC and NPC whose expression was significantly higher ($log_2FC \geq 1.0$, p-value ≤ 0.05) in a subcellular fraction relative to the cytosol. We also identified thousands of isoforms in multi-isoform genes whose gene fraction was significantly higher (Δ Gene fraction ≥ 0.1, Q-value ≤ 0.05) in a subcellular fraction relative to its gene fraction in the cytosol. (**B, D**) The proportion of productive and unproductive transcript classes in each subpopulation of transcripts enriched and/or preferentially sedimenting in subcellular fractions. "T" represents the whole expressed transcriptome in the given cell type, "C" represents the cytosol, "M", "ML", "L" and "H" represent the Mono, Mono+Light, LPR and

HPR fractions respectively. “C*” represents the subset of transcripts comprising multi-isoform genes specifically, in the cytosol.

Because the light and heavy polyribosome fractions were pooled sets of individual polyribosome fractions, we could not assess features in the context of ribosome density. To be clear, LR Frac-seq can be performed without pooling individual polyribosome fractions, which would enable ribosome density-level analyses. We note that the heavy polyribosome fraction is likely composed of both efficiently and inefficiently translated transcripts depending on their ribosome density, and that trends of feature length and GC-content are subject to exceptions in each subcellular fraction. Additionally, Frac-seq differs from ribosome profiling methods in that it doesn’t capture single-nucleotide resolution ribosome footprints. Rather, it stratifies the translated transcriptome in terms of the number of ribosomes associated with full-length mRNAs. Therefore, it is not intended to replace ribosome profiling methods and is instead an alternative approach that benefits from retaining UTRs. We recommend LR Frac-seq for the study of translational control in cases where complete isoform structures and detection of novel isoforms is desired.

A major implication of LR Frac-seq in the field of translatomics is that its library preparation can be modified to enable direct RNA sequencing after fractionation to detect post-transcriptional modifications which are understood to significantly influence translation. For instance, RNA methylation, specifically N6-methyladenosine (m6A), can alter translation efficiency. Pseudouridylation can affect translation dynamics by influencing ribosome stalling and pausing during protein synthesis. RNA editing events, such as adenosine-to-inosine (A-to-I) editing, can modify regulatory sequences, altering the fate of mRNAs. These post-transcriptional modifications exemplify some of the multifaceted ways in which RNA modifications can impact translational control. By coupling accurate positions of post-transcriptional modifications with polyribosome profiles at isoform resolution, LR

Frac-seq could enable more direct correlation of modifications with their effects on translation. Because we used R2C2, which is a cDNA method, to strengthen the confidence of isoform structures, we did not capture modification information beyond RNA editing events. But future adopters of LR Frac-seq can employ direct RNA sequencing methods after fractionation to gain that additional layer of data.

In conclusion, LR Frac-seq enables polyribosome profiling at isoform resolution, retaining complete information about UTRs and novel transcript structures. We tested this method in the context of neuronal differentiation, revealing thousands of transcripts enriched in subcellular fractions relative to the cytosol and largely cell type-specific patterns of isoforms-specific sedimentation between ESC and NPC. Our results present intrinsic transcript features and known and novel RBP motifs that may be important determinants of ribosome association, and this work presents a promising new approach to study translational control without the information loss suffered by ribosome profiling and short read sequencing-based methods.

METHODS

H9 cell culture and differentiation to NPC

H9 cells in feeder-free culture were disaggregated using accutase and resuspended in hESC medium (StemMACS) containing 10 μM Rock inhibitor (Y27632). Cells were then seeded on a matrigel-coated 12-well plate at 50k live cells per well. Rock inhibitor was withdrawn the next day and the cells were cultured in hESC medium for 3 days. Neural differentiation was then induced over 7 days using KSR medium (for 500.5 mL stock: 415 mL KO-DMEM, 75 mL KSR, 100X Glutamax, 100X NEAA, 1000X bME, 10 μM SB431542, 100 nM LDN-193189). A subset of differentiated cells were stained for PAX6 to confirm neural differentiation.

Short read Frac-seq

Cytosolic extracts from monolayer-cultured H9 cells and H9-derived NPCs, both in triplicate, were separated on sucrose gradients as described in the original Frac-seq publication [145]. From these, the monosome fraction (RNAs associated with 1 ribosome), light polyribosome fraction (2-4 ribosomes) and heavy polyribosome fraction (≥5 ribosomes) were isolated using the Gradient Station (Biocomp Inc). RNA was extracted with TRIzol, polyA selected, and converted to directional RNA Seq libraries (BIOO Scientific qRNA) from these fractions in addition to total cytosolic RNA. Biological and technical replicates were sequenced using Hiseq 4000 PE150 (50-100M reads per library).

Long read Frac-seq

From the same fractionated mRNA used prior for Illumina sequencing, full-length cDNA was prepared using the Rolling Circle Amplification to Concatemeric Consensus (R2C2) method [147]. Libraries were pooled and sequenced on an ONT PromethION, generating 12.11M reads with read length N50 of 17.6Kb.

De novo transcriptome assembly from long reads

R2C2 long reads were basecalled with Bonito v0.0.1 (https://github.com/nanoporetech/bonito). Subsequent polyA tail and adapter trimming followed by definition of high-confidence isoform consensus sequences was carried out using Mandalorion v4.0.0 [148] with all sample FASTAs (from ESC and NPC, all subcellular and cytosolic fractions in duplicate) as input. The resultant transcriptome was filtered for redundant transcripts using GFFCompare v0.12.6 [181] against the GRCh38.p13 Release 41 primary assembly annotation [150], and then further filtered and annotated using SQANTI3 v5.1.1 and IsoAnnot Lite v.2.7.3 [149]. SQANTI3 filtering was done using the machine learning filter with a training set proportion of 80% and a correct classification probability threshold of 70%. The final, filtered long read transcriptome was then merged with the GRCh38.p13 Release 41 primary assembly annotation [150], producing a "comprehensive transcriptome", to account for transcripts that were potentially missed by long read sequencing.

Short read data analysis

Short reads were adapter-trimmed with cutadapt, then mapped to the GRCh38.p13 primary assembly genome with the comprehensive transcriptome annotation using STAR v2.7.8a [39]. Transcript-level quantification was performed from the alignments using Salmon v1.9.0 [84] in alignment-based mode. Differential expression analysis at the gene, transcript, and isoform level were carried out using tappAS v1.0.7 [149], which utilizes maSigPro v1.72.0 with the following analysis parameters: polynomial degree of 3, significance level of 0.05, R^2 cutoff of 0.7, fold change of 2, and 9 K clusters. Differential expression analyses were performed for each subcellular fraction against its cognate cytosolic fraction (all in triplicate) for

each cell type, and between subcellular and cytosolic fractions across ESC and NPC. Pathway analyses were done using Metascape [91].

Alternative splicing (AS) analysis was performed using junctionCounts [182], which identifies and quantifies binary splicing events from RNA-seq data, including: alternative 5' and 3' splice sites (A5SS and A3SS), alternative first and last exons (AFE and ALE), skipped exons (SE), retained introns (RI), and mutually exclusive exons (MXE). AS events were then statistically tested by comparing the dispersions of junction support for their included and excluded forms using DEXSeq v1.46.0 [49]. Events were considered significant if they had $0.1 \leq \Psi \leq 0.9$ and adjusted p-value ≤ 0.05 when assessing splicing within a condition, or $|\Delta\Psi| \geq 0.1$ and Q-value ≤ 0.05 when assessing changes in splicing across conditions.

Feature analysis

Transcript features were collected from the transcriptome IsoAnnot Lite annotation and by using custom python scripts, including length measurements of: transcript, CDS, upstream open reading frames (uORF), 5' and 3' UTRs. Total counts of: 5' UTR (TOP and UNR_BS) and 3' UTR (BRD-BOX, CPE, DMRT1_RE, GY-BOX, K-BOX, MBE and UNR_BS) motifs, uORFs and repeat sequences (DNA/hAT-Charlie, DNA/TcMar-Tigger, LINE/L1, LINE/L2, low complexity, LTR/ERVL-MaLR, retroposon/SVA, simple repeat, SINE/Alu, SINE/MIR and srpRNA). Binary features: coding/noncoding, proximal/distal polyA tail usage, predicted nonsense-mediated decay (NMD)/no NMD and intron retention/no intron retention. And lastly, codon frequencies and GC-content of the transcript, CDS, and 5' and 3' UTRs. Feature selection for binary classification between transcripts belonging to subcellular fractions was performed using the Random Forest Classifier (RFC) method from the sklearn.ensemble module of scikit-learn v1.2.2

(https://scikit-learn.org/stable) and evaluated using permutation importance from the sklearn.inspection module. RFC models were generated with the interest of identifying predictive features of ribosome association and were limited by the relatively small subsets of transcripts classed as associated with a particular subcellular fraction.

Motif analysis was performed using HOMER v4.11 [166]. Target sequences were produced by slicing fraction-associated skipped exons (in the monosome relative to cytosol, the light polyribosome fraction relative to cytosol, and the heavy polyribosome relative to the monosome and the light polyribosome separately) into 30 nt windows. Each set was subjected to *de novo* motif discovery against background sets of 30 nt windows produced from skipped exons that were not enriched in their given fraction. Significant motifs (p-value ≤ 0.05, FDR ≤ 0.2) plus a set of known RBP motifs – CISBP-RNA *Homo sapiens* [168] – were then tested for enrichment across all sets of windows in each fraction using SEA v5.5.4 [169]. Motif enrichment scores were filtered for p-value ≤ 0.05. *De novo* motifs enriched in at least one set of windows were compared to RNA-binding protein motifs in the Ray 2013 *Homo sapiens* dataset ([168]) for potential matches using Tomtom v5.5.4 ([167]). The best RBP motif match (p-value ≤ 0.05, Q-value ≤ 0.2) for each *de novo* motif was assigned accordingly.

The same approach to motif analysis was taken with transcripts exhibiting divergent isoform sedimentation between cell types. 30 nt windows were generated for the CDS, 3’ UTR and 5’ UTR of each such isoform. *De novo* motif discovery and enrichment was performed on windows from sets of isoforms preferentially sedimenting with each fraction in each cell type versus those preferentially

sedimenting with the same fraction in the other cell type. These isoforms had inverse sedimentation profiles: meaning that those that preferentially sediment with a fraction in ESC show the opposite sedimentation in the same fraction in NPC.

Luciferase reporter assays

Luciferase reporters designed to test translational control by alternative first exon sequences were assembled from gene blocks (IDTDNA) and cloned into pLightSwitch 5' UTR report (Switchgear Genomics). HEK293 cells, grown on 6 well plates in DMEM supplemented with 10% FCS, were transfected with 2.5 µg pLightswitch reporter plasmid and pMIR (Ambion). 24 hours post-transfection, cells were lysed with Passive Lysis Buffer and analyzed by dual luciferase assay (Promega). Experiments were performed in triplicate. Relative luciferase activity (Renilla vs. Firefly) was plotted in Graphpad and analyzed by paired T-test.

DATA ACCESS

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE244655.

COMPETING INTEREST

The authors declare no competing interests.

ACKNOWLEDGEMENTS

This research was made possible by a grant from the California Institute for Regenerative Medicine (GCIR-06673-A). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of CIRM or any other agency

of the State of California. This work was also supported by grants from the National Institutes of Health R35GM130361.

AUTHOR CONTRIBUTIONS

AJR performed data processing, bioinformatic analyses, and wrote the paper; JMD designed and performed experiments; JRS conceptualized the project and designed experiments; CV helped JMD with performing R2C2 library preparation and data analysis.

Chapter 5: Future Directions and Other Works

5.1 Future Direction: The Next Iteration of *junctionCounts*

Ideally, *junctionCounts* will advance to characterize non-binary event types. For instance, if three or more overlapping events within a gene were detected, junction reads could be assigned by expectation maximization to quantify Ψ values for a single multi-isoform event rather than multiple binary events. Such instances could further involve the assignment of canonical event types as attributes to each isoform to clarify the moving parts involved in the event for easier interpretation.

Additionally, *cdsInsertion* was validated for NMD event predictions, but its coding-to-noncoding switch and NSD event predictions have not been assessed. These classes of events could be tested by comparing the translational output of predicted coding-to-noncoding switch and NSD substrates to their coding and non-NSD counterpart isoforms respectively *in vitro*. *cdsInsertion* could also be expanded to include characterization of codon optimality, especially to highlight the presence of rare codons associated with consequences to mRNA stability and/or translation.

5.2 Future Direction: Towards a Mechanistic Understanding of ASTC

While work from many groups, including our work on LR Frac-seq, has provided some insights into transcript-level features that correlate with specific sedimentation profiles, there's still a lot of room for discovering the specific mechanisms that control isoform-specific ribosome association. Additionally, most studies, including ours, have focused on steady-state sedimentation profiles without perturbation conditions. Two ideas for future studies of ASTC that could make use of Frac-seq are to associate the landscape of RNA modifications with sedimentation, and to correlate mRNA secondary structure with

sedimentation. Moreover, it would be interesting to parse between isoform-specific sedimentation due simply to transcript half-life versus other factors.

5.3 Other Work: Internship at Genentech, Inc.

During my Summer Computational Biology internship at Genentech, Inc. I had the pleasure of working on the development of a long read transcript discovery and quantification tool called *Isosceles*, which is currently in pre-print [183]. This tool importantly outperforms available tools in terms of sensitivity and quantification accuracy across single-cell, pseudo-bulk and bulk resolution levels. My specific contribution to the project was to benchmark *Isosceles* against Bambu [184], FLAIR [185], LIQA [186] and NanoCount [187] using downsampled simulated datasets. This experience serendipitously primed me for the benchmarking experiments I would later conduct with *junctionCounts*.

5.4 Other Work: The Role of IGF2BP3 in B-cell Acute Lymphoblastic Leukemia

Despite decades of research and much progress, certain subtypes of leukemia remain highly resistant to treatment. One recently discovered determinant of the aggressive behavior of leukemia is a protein that regulates post-transcriptional gene expression, insulin-like growth factor 2 mRNA binding protein 3 (IGF2BP3 or IMP3). This oncofetal RNA-binding protein (RBP) is undetectable in most adult tissues but is strongly expressed in embryos and diverse tumor types [188]. IMP3 is known to regulate genes that are related to proliferation, migration, and signaling – which are important in fetal development – but also in cancer. Concordant with this gene regulatory function, IMP3 is overexpressed in a wide range of malignancies (approximately 15% of all cancers), including acute leukemia, and portends a poor prognosis when highly-expressed [189].

Recent studies conducted by our lab in collaboration with the Rao Lab at University of California, Los Angeles have revealed that IMP3 binding sites are enriched in the 3' untranslated region (UTR) of target mRNAs to regulate their stability via a mechanism that involves the RNA-induced silencing complex (RISC). Using novel, murine models of IMP3 deficiency, we have discovered that IMP3 is required for the development of a fully-penetrant, lethal leukemia *in vivo*. Together, our extensive prior work provides a mechanistic framework for IMP3's function and a solid foundation for its importance in disease.

To fully understand the nature of IMP3's effect on RISC-mRNA association and to understand its role in cancer, I performed global characterization of RISC-associated mRNA transcripts in the presence and absence of IMP3 using conditional enhanced cross-linking immunoprecipitation sequencing (eCLIP-seq). Next, our lab plans to employ bioinformatic analyses aimed at discovering features in mRNA 3' UTRs that are necessary for IMP3-RISC allostery. Following that, we hope to experimentally dissect the mechanisms by which IMP3 enables or obstructs RISC association by manipulating the accessibility and stability of RISC target sites in mRNA 3' UTRs.

References

1. Michelini F, Jalihal AP, Francia S, Meers C, Neeb ZT, Rossiello F, et al. From “Cellular” RNA to “Smart” RNA: Multiple Roles of RNA in Genome Stability and Beyond. Chem Rev. 2018;118: 4365–4403.

2. Moazed D. Small RNAs in transcriptional gene silencing and genome defence. Nature. 2009;457: 413–420.

3. Alshaer W, Zureigat H, Al Karaki A, Al-Kadash A, Gharaibeh L, Hatmal MM, et al. siRNA: Mechanism of action, challenges, and therapeutic approaches. Eur J Pharmacol. 2021;905: 174178.

4. Ergin K, Çetinkaya R. Regulation of MicroRNAs. 2022. pp. 1–32.

5. Iwasaki YW, Siomi MC, Siomi H. PIWI-Interacting RNA: Its Biogenesis and Functions. Annu Rev Biochem. 2015;84: 405–433.

6. Allen SE, Nowacki M. Roles of Noncoding RNAs in Ciliate Genome Architecture. J Mol Biol. 2020;432: 4186–4198.

7. Furrer DI, Swart EC, Kraft MF, Sandoval PY, Nowacki M. Two Sets of Piwi Proteins Are Involved in Distinct sRNA Pathways Leading to Elimination of Germline-Specific DNA. Cell Rep. 2017;20: 505–520.

8. Hoehener C, Hug I, Nowacki M. Dicer-like Enzymes with Sequence Cleavage Preferences. Cell. 2018;173: 234–247.e7.

9. Sandoval PY, Swart EC, Arambasic M, Nowacki M. Functional Diversification of Dicer-like Proteins and Small RNAs Required for Genome Sculpting. Dev Cell. 2014;28: 174–188.

10. Cheloufi S, Dos Santos CO, Chong MMW, Hannon GJ. A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. Nature. 2010;465: 584–589.

11. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R. Control of translation and mRNA degradation by miRNAs and siRNAs: Table 1. Genes Dev. 2006;20: 515–524.

12. Morales L, Oliveros JC, Fernandez-Delgado R, tenOever BR, Enjuanes L, Sola I. SARS-CoV-Encoded Small RNAs Contribute to Infection-Associated Lung Pathology. Cell Host Microbe. 2017;21: 344–355.

13. Perez JT, Varble A, Sachidanandam R, Zlatev I, Manoharan M, Garcia-Sastre A, et al. Influenza A virus-generated small RNAs regulate the switch from transcription to replication. Proceedings of the National Academy of Sciences. 2010;107: 11525–11530.

14. Shi J, Sun J, Wang B, Wu M, Zhang J, Duan Z, et al. Novel microRNA‐like viral small regulatory RNAs arising during human hepatitis A virus infection. The FASEB Journal. 2014;28: 4381–4393.

15. Weng K-F, Hung C-T, Hsieh P-T, Li M-L, Chen G-W, Kung Y-A, et al. A cytoplasmic

RNA virus generates functional viral small RNAs and regulates viral IRES activity in mammalian cells. Nucleic Acids Res. 2014;42: 12789–12805.

16. Meng F, Siu GK-H, Mok BW-Y, Sun J, Fung KSC, Lam JY-W, et al. Viral MicroRNAs Encoded by Nucleocapsid Gene of SARS-CoV-2 Are Detected during Infection, and Targeting Metabolic Pathways in Host Cells. Cells. 2021;10: 1762.

17. Cheng Z, Cheng L, Lin J, Lunbiao C, Chunyu L, Guoxin S, et al. Verification of SARS-CoV-2-Encoded small RNAs and contribution to Infection-Associated lung inflammation. bioRxiv. 2021; 2021.05.16.444324.

18. Pawlica P, Yario TA, White S, Wang J, Moss WN, Hui P, et al. SARS-CoV-2 expresses a microRNA-like small RNA able to selectively repress host genes. Proc Natl Acad Sci U S A. 2021;118. doi:10.1073/pnas.2116668118

19. Withers JB, Mondol V, Pawlica P, Rosa-Mercado NA, Tycowski KT, Ghasempur S, et al. Idiosyncrasies of Viral Noncoding RNAs Provide Insights into Host Cell Biology. Annual Review of Virology. 2019;6: 297–317.

20. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell. 2020;181: 914–921.e10.

21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550.

22. Uren PJ, Bahrami-Samani E, Burns SC, Qiao M, Karginov FV, Hodges E, et al. Site identification in high-throughput RNA–protein interaction data. Bioinformatics. 2012;28: 3013–3020.

23. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A. 2004;101: 7287–7292.

24. Kruger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res. 2006;34: W451–W454.

25. Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA. 2004;10: 1507–1517.

26. DeDiego ML, Nieto-Torres JL, Regla-Nava JA, Jimenez-Guardeno JM, Fernandez-Delgado R, Fett C, et al. Inhibition of NF- B-Mediated Inflammation in Severe Acute Respiratory Syndrome Coronavirus-Infected Mice Increases Survival. J Virol. 2014;88: 913–924.

27. Nelemans T, Kikkert M. Viral Innate Immune Evasion and the Pathogenesis of Emerging RNA Virus Infections. Viruses. 2019;11: 961.

28. Xiong Y, Liu Y, Cao L, Wang D, Guo M, Jiang A, et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. Emerg Microbes Infect. 2020;9: 761–770.

29. Chen Y, Liu Q, Guo D. Emerging coronaviruses: Genome structure, replication, and pathogenesis. J Med Virol. 2020;92: 418–423.

30. Li C-X, Chen J, Lv S-K, Li J-H, Li L-L, Hu X. Whole-Transcriptome RNA Sequencing Reveals Significant Differentially Expressed mRNAs, miRNAs, and lncRNAs and Related Regulating Biological Pathways in the Peripheral Blood of COVID-19 Patients. Mediators Inflamm. 2021;2021: 1–22.

31. Mirzaei R, Mahdavi F, Badrzadeh F, Hosseini-Fard SR, Heidary M, Jeda AS, et al. The emerging role of microRNAs in the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. Int Immunopharmacol. 2021;90: 107204.

32. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu A-L, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res. 2008;18: 610–621.

33. Cazalla D, Yario T, Steitz JA. Down-regulation of a host microRNA by a Herpesvirus saimiri noncoding RNA. Science. 2010;328: 1563–1566.

34. Ebert MS, Sharp PA. MicroRNA sponges: progress and possibilities. RNA. 2010;16: 2043–2050.

35. Zhou Q, Liu L, Zhou J, Chen Y, Xie D, Yao Y, et al. Novel Insights Into MALAT1 Function as a MicroRNA Sponge in NSCLC. Front Oncol. 2021;11: 758653.

36. Xiao J, Lin L, Luo D, Shi L, Chen W, Fan H, et al. Long noncoding RNA TRPM2-AS acts as a microRNA sponge of miR-612 to promote gastric cancer progression and radioresistance. Oncogenesis. 2020;9: 29.

37. Bartoszewski R, Dabrowski M, Jakiela B, Matalon S, Harrod KS, Sanak M, et al. SARS-CoV-2 may regulate cellular responses through depletion of specific host miRNAs. American Journal of Physiology-Lung Cellular and Molecular Physiology. 2020;319: L444–L455.

38. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359.

39. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21.

40. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14: 128.

41. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44: W90–W97.

42. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene Set Knowledge Discovery with Enrichr. Current Protocols. 2021;1. doi:10.1002/cpz1.90

43. Chen EY. Enrichr. [cited 1 Dec 2021]. Available: https://maayanlab.cloud/Enrichr

44. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012;28: 1919–1920.

45. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.

46. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12: 996–1006.

47. Stepanowsky P, Levy E, Kim J, Jiang X, Ohno-Machado L. Prediction of MicroRNA Precursors Using Parsimonious Feature Sets. Cancer Inform. 2014;13: 95–102.

48. Kerpedjiev P, Hammer S, Hofacker IL. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics. 2015;31: 3377–3379.

49. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Res. 2012;22: 2008–2017.

50. Martinez-Nunez RT, Wallace A, Coyne D, Jansson L, Rush M, Ennajdaoui H, et al. Modulation of nonsense mediated decay by rapamycin. Nucleic Acids Res. 2017;45: 3448–3459.

51. Field AR, Jacobs FMJ, Fiddes IT, Phillips APR, Reyes-Ortiz AM, LaMontagne E, et al. Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes. Stem Cell Reports. 2019;12: 245–257.

52. Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc Natl Acad Sci U S A. 2003;100: 189–192.

53. Ilouz R, Lev-Ram V, Bushong EA, Stiles TL, Friedmann-Morvinski D, Douglas C, et al. Isoform-specific subcellular localization and function of protein kinase A identified by mosaic imaging of mouse brain. Elife. 2017;6. doi:10.7554/eLife.17681

54. Wang X, Hou J, Quedenau C, Chen W. Pervasive isoform‐specific translational regulation via alternative transcription start sites in mammals. Mol Syst Biol. 2016;12: 875.

55. Rosenfeld MG, Lin CR, Amara SG, Stolarsky L, Roos BA, Ong ES, et al. Calcitonin mRNA polymorphism: peptide switching associated with alternative RNA splicing events. Proc Natl Acad Sci U S A. 1982;79: 1717–1721.

56. Jiang W, Chen L. Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. Comput Struct Biotechnol J. 2021;19: 183–195.

57. El Marabti E, Younis I. The Cancer Spliceome: Reprograming of Alternative Splicing in

Cancer. Front Mol Biosci. 2018;5: 80.

58. Su C-H, D D, Tarn W-Y. Alternative Splicing in Neurogenesis and Brain Development. Front Mol Biosci. 2018;5: 12.

59. Ren P, Lu L, Cai S, Chen J, Lin W, Han F. Alternative Splicing: A New Cause and Potential Therapeutic Target in Autoimmune Disease. Front Immunol. 2021;12: 713540.

60. Tang SJ, Shen H, An O, Hong H, Li J, Song Y, et al. Cis- and trans-regulations of pre-mRNA splicing by RNA editing enzymes influence cancer development. Nat Commun. 2020;11: 799.

61. Bradley RK, Anczuków O. RNA splicing dysregulation and the hallmarks of cancer. Nat Rev Cancer. 2023;23: 135–155.

62. Urbanski LM, Leclair N, Anczuków O. Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. Wiley Interdiscip Rev RNA. 2018;9: e1476.

63. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol. 2017;18: 437–451.

64. Zhang X, Chen MH, Wu X, Kodani A, Fan J, Doan R, et al. Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. Cell. 2016;166: 1147–1162.e15.

65. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12: 671–682.

66. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7: 562–578.

67. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33: 290–295.

68. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8: 1494–1512.

69. Nagy E, Maquat LE. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci. 1998;23: 198–199.

70. Lytle JR, Steitz JA. Premature termination codons do not affect the rate of splicing of neighboring introns. RNA. 2004;10: 657–668.

71. Sterne-Weiler T, Weatheritt RJ, Best AJ, Ha KCH, Blencowe BJ. Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. Mol Cell. 2018;72: 187–200.e6.

72. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47: D766–D773.

73. Andreadis A, Nadal-Ginard B. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. Annual review of. 1987. Available: https://www.annualreviews.org/doi/pdf/10.1146/annurev.bi.56.070187.002343

74. Sammeth M, Foissac S, Guigó R. A general definition and nomenclature for alternative splicing events. PLoS Comput Biol. 2008;4: e1000147.

75. Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, Yan BY, et al. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep. 2012;1: 167–178.

76. Stovner EB, Sætrom P. PyRanges: efficient comparison of genomic intervals in Python. Bioinformatics. 2020;36: 918–919.

77. Alamancos GP, Pagès A, Trincado JL, Bellora N, Eyras E. Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA. 2015;21: 1521–1531.

78. Vigevani L, Gohr A, Webb T, Irimia M, Valcárcel J. Molecular basis of differential 3' splice site sensitivity to anti-tumor drugs targeting U2 snRNP. Nat Commun. 2017;8: 2100.

79. Vaquero-Garcia J, Aicher JK, Jewell S, Gazzara MR, Radens CM, Jha A, et al. RNA splicing analysis using heterogeneous and large RNA-seq datasets. Nat Commun. 2023;14: 1230.

80. Wang Y, Xie Z, Kutschera E, Adams JI, Kadash-Edmondson KE, Xing Y. rMATS-turbo: an efficient and flexible computational tool for alternative splicing analysis of large-scale RNA-seq data. Nat Protoc. 2024;19: 1083–1104.

81. Kahles A, Ong CS, Zhong Y, Rätsch G. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. Bioinformatics. 2016;32: 1840–1847.

82. Frazee AC, Jaffe AE, Langmead B, Leek JT. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics. 2015;31: 2778–2784.

83. Zhang R, Lahens NF, Ballance HI, Hughes ME, Hogenesch JB. A circadian gene expression atlas in mammals: implications for biology and medicine. Proc Natl Acad Sci U S A. 2014;111: 16219–16224.

84. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14: 417–419.

85. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005;437: 69–87.

86. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, et al. Comparative and demographic analysis of orang-utan genomes. Nature. 2011;469: 529–533.

87. Zimin AV, Cornish AS, Maudhoo MD, Gibbs RM, Zhang X, Pandey S, et al. A new rhesus macaque assembly and annotation for next-generation sequencing analyses. Biol Direct. 2014;9: 20.

88. Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG, et al. Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation. Genome Res. 2018;28: 1029–1038.

89. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100.

90. Clustering large applications (program CLARA). Finding Groups in Data. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2008. pp. 126–163.

91. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10: 1523.

92. Tranchevent L-C, Aubé F, Dulaurier L, Benoit-Pilven C, Rey A, Poret A, et al. Identification of protein features encoded by alternative exons using Exon Ontology. Genome Res. 2017;27: 1087–1097.

93. Fritz SE, Ranganathan S, Wang CD, Hogg JR. An alternative UPF1 isoform drives conditional remodeling of nonsense-mediated mRNA decay. EMBO J. 2022;41: e108898.

94. Kuderna LFK, Ulirsch JC, Rashid S, Ameen M, Sundaram L, Hickey G, et al. Identification of constrained sequence elements across 239 primate genomes. Nature. 2023;625: 735–742.

95. Mazin PV, Jiang X, Fu N, Han D, Guo M, Gelfand MS, et al. Conservation, evolution, and regulation of splicing during prefrontal cortex development in humans, chimpanzees, and macaques. RNA. 2018;24: 585–596.

96. Razzaq A, Robinson IM, McMahon HT, Skepper JN, Su Y, Zelhof AC, et al. Amphiphysin is necessary for organization of the excitation–contraction coupling machinery of muscles, but not for synaptic vesicle endocytosis in Drosophila. Genes Dev. 2001;15: 2967–2979.

97. Floyd SR, Porro EB, Slepnev VI, Ochoa GC, Tsai LH, De Camilli P. Amphiphysin 1 binds the cyclin-dependent kinase (cdk) 5 regulatory subunit p35 and is phosphorylated by cdk5 and cdc2. J Biol Chem. 2001;276: 8104–8110.

98. Anggono V, Koç-Schmitz Y, Widagdo J, Kormann J, Quan A, Chen C-M, et al. PICK1 interacts with PACSIN to regulate AMPA receptor internalization and cerebellar long-term depression. Proc Natl Acad Sci U S A. 2013;110: 13976–13981.

99. Dumont V, Lehtonen S. PACSIN proteins in vivo: Roles in development and physiology. Acta Physiol . 2022;234: e13783.

100. Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. Curr Protoc Bioinformatics. 2012;Chapter 1: 1.4.1–1.4.33.

101. Boutz PL, Stoilov P, Li Q, Lin C-H, Chawla G, Ostrow K, et al. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes Dev. 2007;21: 1636–1652.

102. Bhat VD, Jayaraj J, Babu K. RNA and neuronal function: the importance of post-transcriptional regulation. Oxf Open Neurosci. 2022;1: kvac011.

103. Wang Y, Liu X, Biederer T, Südhof TC. A family of RIM-binding proteins regulated by alternative splicing: Implications for the genesis of synaptic active zones. Proc Natl Acad Sci U S A. 2002;99: 14464–14469.

104. Pálinkás HL, Rácz GA, Gál Z, Hoffmann OI, Tihanyi G, Róna G, et al. CRISPR/Cas9-Mediated Knock-Out of dUTPase in Mice Leads to Early Embryonic Lethality. Biomolecules. 2019;9. doi:10.3390/biom9040136

105. Ladner RD, Caradonna SJ. The human dUTPase gene encodes both nuclear and mitochondrial isoforms. Differential expression of the isoforms and characterization of a cDNA encoding the mitochondrial species. J Biol Chem. 1997;272: 19072–19080.

106. Frade JM, Ovejero-Benito MC. Neuronal cell cycle: the neuron itself and its circumstances. Cell Cycle. 2015;14: 712–720.

107. Weyn-Vanhentenryck SM, Feng H, Ustianenko D, Duffié R, Yan Q, Jacko M, et al. Precise temporal regulation of alternative splicing during neural development. Nat Commun. 2018;9: 2189.

108. Lev-Maor G, Sorek R, Shomron N, Ast G. The birth of an alternatively spliced exon: 3’ splice-site selection in Alu exons. Science. 2003;300: 1288–1291.

109. Zhou J, Zhao S, Dunker AK. Intrinsically Disordered Proteins Link Alternative Splicing and Post-translational Modifications to Complex Cell Signaling and Regulation. J Mol Biol. 2018;430: 2342–2359.

110. Buljan M, Chalancon G, Dunker AK, Bateman A, Balaji S, Fuxreiter M, et al. Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. Curr Opin Struct Biol. 2013;23: 443–450.

111. Klauer AA, van Hoof A. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. Wiley Interdiscip Rev RNA. 2012;3: 649–660.

112. Lee PJ, Yang S, Sun Y, Guo JU. Regulation of nonsense-mediated mRNA decay in neural development and disease. J Mol Cell Biol. 2021;13: 269–281.

113. Dhamija S, Menon MB. Non-coding transcript variants of protein-coding genes - what are they good for? RNA Biol. 2018;15: 1025–1031.

114. Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, et al. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. Nat Commun. 2021;12: 463.

115. Suzuki JMNGL, Osterhoudt K, Cartwright-Acar CH, Gomez DR, Katzman S, Zahler AM. A genetic screen in C. elegans reveals roles for KIN17 and PRCC in maintaining 5' splice site identity. PLoS Genet. 2022;18: e1010028.

116. Cartwright-Acar CH, Osterhoudt K, Suzuki JMNGL, Gomez DR, Katzman S, Zahler AM. A forward genetic screen in C. elegans identifies conserved residues of spliceosomal proteins PRP8 and SNRNP200/BRR2 with a role in maintaining 5' splice site identity. Nucleic Acids Res. 2022;50: 11834–11857.

117. Draper JM, Philipp J, Neeb ZT, Thomas R, Katzman S, Salama S, et al. Isoform-specific translational control is evolutionarily conserved in primates. bioRxiv. 2023. doi:10.1101/2023.04.21.537863

118. Hunter O, Talkish J, Quick-Cleveland J, Igel H, Tan A, Kuersten S, et al. Broad variation in response of individual introns to splicing inhibitors in a humanized yeast strain. bioRxiv. 2023. doi:10.1101/2023.10.05.560965

119. Osterhoudt K, Bagno O, Katzman S, Zahler AM. Spliceosomal helicases DDX41/SACY-1 and PRP22/MOG-5 both contribute to proofreading against proximal 3' splice site usage. RNA. 2024;30: 404–417.

120. Witten JT, Ule J. Understanding splicing regulation through RNA splicing maps. Trends Genet. 2011;27: 89–97.

121. Fong N, Kim H, Zhou Y, Ji X, Qiu J, Saldi T, et al. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes Dev. 2014;28: 2663–2676.

122. Zhang P, He D, Xu Y, Hou J, Pan B-F, Wang Y, et al. Genome-wide identification and differential analysis of translational initiation. Nat Commun. 2017;8: 1749.

123. Wang Y, Liu J, Huang BO, Xu Y-M, Li J, Huang L-F, et al. Mechanism of alternative splicing and its regulation. Biomed Rep. 2015;3: 152–158.

124. Konarska MM. Recognition of the 5' splice site by the spliceosome. Acta Biochim Pol. 1998;45: 869–881.

125. Wu S, Romfo CM, Nilsen TW, Green MR. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. Nature. 1999;402: 832–835.

126. Hellen CU, Sarnow P. Internal ribosome entry sites in eukaryotic mRNA molecules. Genes Dev. 2001;15: 1593–1612.

127. Morris DR, Geballe AP. Upstream open reading frames as regulators of mRNA translation. Mol Cell Biol. 2000;20: 8635–8642.

128. Weber R, Ghoshdastider U, Spies D, Duré C, Valdivia-Francia F, Forny M, et al. Monitoring the 5'UTR landscape reveals isoform switches to drive translational

efficiencies in cancer. Oncogene. 2023;42: 638–650.

129. Ciolli Mattioli C, Rom A, Franke V, Imami K, Arrey G, Terne M, et al. Alternative 3' UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. Nucleic Acids Res. 2019;47: 2560–2573.

130. Mayya VK, Duchaine TF. Ciphers and Executioners: How 3'-Untranslated Regions Determine the Fate of Messenger RNAs. Front Genet. 2019;10: 6.

131. Otsuka H, Fukao A, Funakami Y, Duncan KE, Fujiwara T. Emerging Evidence of Translational Control by AU-Rich Element-Binding Proteins. Front Genet. 2019;10: 332.

132. Weidensdorfer D, Stöhr N, Baude A, Lederer M, Köhn M, Schierhorn A, et al. Control of c-myc mRNA stability by IGF2BP1-associated cytoplasmic RNPs. RNA. 2009;15: 104–115.

133. Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, Cooper TA, et al. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet. 2008;40: 1416–1425.

134. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40: 1413–1415.

135. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456: 470–476.

136. Lian Z, Wang L, Yamaga S, Bonds W, Beazer-Barclay Y, Kluger Y, et al. Genomic and proteomic analysis of the myeloid differentiation program. Blood. 2001;98: 513–524.

137. Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, et al. Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. Mol Cell Proteomics. 2002;1: 323–333.

138. Cox B, Kislinger T, Emili A. Integrating gene and protein expression data: pattern analysis and profile mining. Methods. 2005;35: 303–314.

139. Schmidt MW, Houseman A, Ivanov AR, Wolf DA. Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe. Mol Syst Biol. 2007;3: 79.

140. Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, et al. Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics. 2009;10: 161.

141. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324: 218–223.

142. Wang T, Cui Y, Jin J, Guo J, Wang G, Yin X, et al. Translating mRNAs strongly

correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. Nucleic Acids Res. 2013;41: 4743–4754.

143. Reynoso MA, Juntawong P, Lancia M, Blanco FA, Bailey-Serres J, Zanetti ME. Translating Ribosome Affinity Purification (TRAP) followed by RNA sequencing technology (TRAP-SEQ) for quantitative assessment of plant translatomes. Methods Mol Biol. 2015;1284: 185–207.

144. Heiman M, Kulicke R, Fenster RJ, Greengard P, Heintz N. Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP). Nat Protoc. 2014;9: 1282–1291.

145. Sterne-Weiler T, Martinez-Nunez RT, Howard JM, Cvitovik I, Katzman S, Tariq MA, et al. Frac-seq reveals isoform-specific recruitment to polyribosomes. Genome Res. 2013;23: 1615–1623.

146. Steijger T, Abril JF, Engström PG, Kokocinski F, RGASP Consortium, Hubbard TJ, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013;10: 1177–1184.

147. Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, et al. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. Proc Natl Acad Sci U S A. 2018;115: 9726–9731.

148. Volden R, Schimke KD, Byrne A, Dubocanin D, Adams M, Vollmers C. Identifying and quantifying isoforms from accurate full-length transcriptome sequencing reads with Mandalorion. Genome Biol. 2023;24: 167.

149. de la Fuente L, Arzalluz-Luque Á, Tardáguila M, Del Risco H, Martí C, Tarazona S, et al. tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. Genome Biol. 2020;21: 119.

150. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22: 1760–1774.

151. Li X, Feng R, Huang C, Wang H, Wang J, Zhang Z, et al. MicroRNA-351 regulates TMEM 59 (DCF1) expression and mediates neural stem cell morphogenesis. RNA Biol. 2012;9: 292–301.

152. Floor SN, Doudna JA. Tunable protein synthesis by transcript isoforms in human cells. Elife. 2016;5. doi:10.7554/eLife.10921

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

154. Königs V, de Oliveira Freitas Machado C, Arnold B, Blümel N, Solovyeva A, Löbbert S, et al. SRSF7 maintains its homeostasis through the expression of Split-ORFs and nuclear body assembly. Nat Struct Mol Biol. 2020;27: 260–273.

155. Maquat LE, Kinniburgh AJ, Rachmilewitz EA, Ross J. Unstable beta-globin mRNA in mRNA-deficient beta o thalassemia. Cell. 1981;27: 543–553.

156. Nickless A, Bailis JM, You Z. Control of gene expression through the nonsense-mediated RNA decay pathway. Cell Biosci. 2017;7: 26.

157. Celik A, Baker R, He F, Jacobson A. High-resolution profiling of NMD targets in yeast reveals translational fidelity as a basis for substrate selection. RNA. 2017;23: 735–748.

158. Zhang P, Zhou C, Jing Q, Gao Y, Yang L, Li Y, et al. Role of APR3 in cancer: apoptosis, autophagy, oxidative stress, and cancer therapy. Apoptosis. 2023. doi:10.1007/s10495-023-01882-w

159. Ding X, Chen Y, Han L, Qiu W, Gu X, Zhang H. Apoptosis related protein 3 is a lysosomal membrane protein. Biochem Biophys Res Commun. 2015;460: 915–922.

160. Zhao D, Hamilton JP, Hardigan M, Yin D, He T, Vaillancourt B, et al. Analysis of Ribosome-Associated mRNAs in Rice Reveals the Importance of Transcript Size and GC Content in Translation. G3 . 2017;7: 203–219.

161. Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 2003;100: 3889–3894.

162. Lyu X, Yang Q, Zhao F, Liu Y. Codon usage and protein length-dependent feedback from translation elongation regulates translation initiation and elongation speed. Nucleic Acids Res. 2021;49: 9404–9423.

163. Rogers DW, Böttcher MA, Traulsen A, Greig D. Ribosome reinitiation can explain length-dependent translation of messenger RNA. PLoS Comput Biol. 2017;13: e1005592.

164. Courel M, Clément Y, Bossevain C, Foretek D, Vidal Cruchez O, Yi Z, et al. GC content shapes mRNA storage and decay in human cells. Elife. 2019;8. doi:10.7554/eLife.49708

165. Litterman AJ, Kageyama R, Le Tonqueze O, Zhao W, Gagnon JD, Goodarzi H, et al. A massively parallel 3’ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. Genome Res. 2019;29: 896–906.

166. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38: 576–589.

167. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8: R24.

168. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013;499:

172–177.

169. Bailey TL, Grant CE. SEA: Simple Enrichment Analysis of motifs. bioRxiv. 2021. p. 2021.08.23.457422. doi:10.1101/2021.08.23.457422

170. Sévigny M, Bourdeau Julien I, Venkatasubramani JP, Hui JB, Dutchak PA, Sephton CF. FUS contributes to mTOR-dependent inhibition of translation. J Biol Chem. 2020;295: 18459–18473.

171. Paronetto MP, Zalfa F, Botti F, Geremia R, Bagni C, Sette C. The nuclear RNA-binding protein Sam68 translocates to the cytoplasm and associates with the polysomes in mouse spermatocytes. Mol Biol Cell. 2006;17: 14–24.

172. Aviner R, Hofmann S, Elman T, Shenoy A, Geiger T, Elkon R, et al. Proteomic analysis of polyribosomes identifies splicing factors as potential regulators of translation during mitosis. Nucleic Acids Res. 2017;45: 5945–5957.

173. Jin J, Jing W, Lei X-X, Feng C, Peng S, Boris-Lawrie K, et al. Evidence that Lin28 stimulates translation by recruiting RNA helicase A to polysomes. Nucleic Acids Res. 2011;39: 3724–3734.

174. Anisimova AS, Kolyupanova NM, Makarova NE, Egorov AA, Kulakovskiy IV, Dmitriev SE. Human Tissues Exhibit Diverse Composition of Translation Machinery. Int J Mol Sci. 2023;24. doi:10.3390/ijms24098361

175. Rizzotto D, Zaccara S, Rossi A, Galbraith MD, Andrysik Z, Pandey A, et al. Nutlin-Induced Apoptosis Is Specified by a Translation Program Regulated by PCBP2 and DHX30. Cell Rep. 2020;30: 4355–4369.e6.

176. Smart F, Aschrafi A, Atkins A, Owens GC, Pilotte J, Cunningham BA, et al. Two isoforms of the cold-inducible mRNA-binding protein RBM3 localize to dendrites and promote translation. J Neurochem. 2007;101: 1367–1379.

177. Markus MA, Morris BJ. RBM4: a multifunctional RNA-binding protein. Int J Biochem Cell Biol. 2009;41: 740–743.

178. Zhang W, Sun Y, Bai L, Zhi L, Yang Y, Zhao Q, et al. RBMS1 regulates lung cancer ferroptosis through translational control of SLC7A11. J Clin Invest. 2021;131. doi:10.1172/JCI152067

179. Ueno T, Taga Y, Yoshimoto R, Mayeda A, Hattori S, Ogawa-Goto K. Component of splicing factor SF3b plays a key role in translational control of polyribosomes on the endoplasmic reticulum. Proc Natl Acad Sci U S A. 2019;116: 9340–9349.

180. Svitkin YV, Yanagiya A, Karetnikov AE, Alain T, Fabian MR, Khoutorsky A, et al. Control of translation and miRNA-dependent repression by a novel poly(A) binding protein, hnRNP-Q. PLoS Biol. 2013;11: e1001564.

181. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res. 2020;9. doi:10.12688/f1000research.23297.2

182. Ritter A, Wallace A, Ronaghi N, Sanford JR. The evolutionary dynamics of alternative splicing during primate neuronal differentiation. bioRxiv. 2024. p. 2024.02.20.581203. doi:10.1101/2024.02.20.581203

183. Kabza M, Ritter A, Byrne A, Sereti K, Le D, Stephenson W, et al. Accurate long-read transcript discovery and quantification at single-cell resolution with Isosceles. bioRxiv. 2023. p. 2023.11.30.566884. doi:10.1101/2023.11.30.566884

184. Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, et al. Context-aware transcript quantification from long-read RNA-seq data with Bambu. Nat Methods. 2023;20: 1187–1195.

185. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun. 2020;11: 1438.

186. Hu Y, Fang L, Chen X, Zhong JF, Li M, Wang K. LIQA: long-read isoform quantification and analysis. Genome Biol. 2021;22: 182.

187. Gleeson J, Leger A, Prawer YDJ, Lane TA, Harrison PJ, Haerty W, et al. Accurate expression quantification from nanopore direct RNA sequencing with NanoCount. Nucleic Acids Res. 2022;50: e19.

188. Ennajdaoui H, Howard JM, Sterne-Weiler T, Jahanbani F, Coyne DJ, Uren PJ, et al. IGF2BP3 Modulates the Interaction of Invasion-Associated Transcripts with RISC. Cell Rep. 2016;15: 1876–1883.

189. Lochhead P, Imamura Y, Morikawa T, Kuchiba A, Yamauchi M, Liao X, et al. Insulin-like growth factor 2 messenger RNA binding protein 3 (IGF2BP3) is a marker of unfavourable prognosis in colorectal cancer. Eur J Cancer. 2012;48: 3405–3413.