Title
Transcriptomic studies on the production and cytosolic fate of alternative isoforms in mammalian cells

Permalink
https://escholarship.org/uc/item/3ht8g2sq

Author
Wallace, Andrew

Publication Date
2019

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA
SANTA CRUZ

TRANSCRIPTOMIC STUDIES ON THE PRODUCTION AND CYTOSOLIC FATE OF ALTERNATIVE ISOFORMS IN MAMMALIAN CELLS

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

Doctor of Philosophy

in

CHEMISTRY

by

Andrew Wallace

December 2019

The Dissertation of Andrew Wallace is approved:

____________________________________________
Professor Michael Stone, Chair

____________________________________________
Professor Jeremy Sanford

____________________________________________
Professor Todd Lowe

____________________________________________
Professor Christopher Vollmers

____________________________________________
Quentin Williams
Acting Vice Provost and Dean of Graduate Studies
# Table of Contents

List of Figures ........................................ vi
List of Tables ....................................... xii
Abstract ........................................... xiii
Dedication ........................................... xv
Acknowledgments ..................................... xvi

1 Introduction ........................................... 1

2 RNA-seq utilities junctionCounts and cdsInsertion ............. 7
  2.1 junctionCounts ........................................ 8
      2.1.1 Rationale for junctionCounts ....................... 8
      2.1.2 Alternative event definition in junctionCounts ....... 10
      2.1.3 Alternative event classification in junctionCounts .... 11
      2.1.4 Evaluation of junctionCounts event identification and classification .... 16
      2.1.5 Alternative event quantification in junctionCounts .... 18
      2.1.6 Evaluation of junctionCounts quantification .......... 20
      2.1.7 Discussion and Future Directions .................... 30
  2.2 cdsInsertion and findSwitchEvents ........................ 32
      2.2.1 Rationale for cdsInsertion .......................... 32
      2.2.2 Rationale for findSwitchEvents ...................... 34
      2.2.3 Approach to CDS insertion ........................... 35
      2.2.4 Assessment of implications for alternative events using findSwitchEvents 36
      2.2.5 Evaluation of NMD switch events identified using cdsInsertion and findSwitchEvents 37
      2.2.6 Discussion and Future Directions .................... 41
3 Measurement and interpretation of isoform-specific polysome co-sedimentation using Frac-seq

<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
</tr>
<tr>
<td>3.1.1 Brief introduction to translatomics and the rationale for Frac-seq</td>
</tr>
<tr>
<td>3.1.2 Previous work involving Frac-seq and similar techniques</td>
</tr>
<tr>
<td>3.1.3 Considerations in the identification and interpretation of isoform-specific sedimentation</td>
</tr>
<tr>
<td>3.2 Materials and Methods</td>
</tr>
<tr>
<td>3.2.1 Mapping Illumina reads</td>
</tr>
<tr>
<td>3.2.2 Identification of differentially expressed genes</td>
</tr>
<tr>
<td>3.2.3 Identification of putative CDS regions with/without PTCs in unannotated transcripts</td>
</tr>
<tr>
<td>3.2.4 Classification of NMD events</td>
</tr>
<tr>
<td>3.2.5 Identification of alternative splicing/processing events</td>
</tr>
<tr>
<td>3.2.6 Identification of alternative event differences across fractions/cell types</td>
</tr>
<tr>
<td>3.3 Results</td>
</tr>
<tr>
<td>3.3.1 A substantial fraction of alternative events exhibit differential sedimentation in both ESC and NPC</td>
</tr>
<tr>
<td>3.3.2 Fraction trend clusters and characteristics of differentially sedimenting events</td>
</tr>
<tr>
<td>3.3.3 NMD event cluster association is not fully explained by CDS lengths</td>
</tr>
<tr>
<td>3.3.4 Changes in differential sedimentation between H9 ESC and NPC</td>
</tr>
<tr>
<td>3.3.5 UPF3B and UPF1 knockdown induce modest down-polysome shift in PTC isoforms</td>
</tr>
<tr>
<td>3.3.6 NMD isoforms decrease in relative abundance after neurodifferentiation and NMD factor knockdown</td>
</tr>
<tr>
<td>3.4 Conclusions and Discussion</td>
</tr>
<tr>
<td>3.4.1 Interpretation of differential sedimentation</td>
</tr>
<tr>
<td>3.4.2 NMD status contributes to NMD switch event sedimentation</td>
</tr>
<tr>
<td>3.4.3 NMD switch event sedimentation in H9 NPC is consistent with stabilization of NMD targets</td>
</tr>
</tbody>
</table>

4 Conservation of alternative splicing dynamics in primate in vitro neurodifferentiation

<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
</tr>
<tr>
<td>4.2 Materials and Methods</td>
</tr>
<tr>
<td>4.2.1 Mapping Illumina reads</td>
</tr>
<tr>
<td>4.2.2 Choice, pre-processing, and expansion of transcriptome annotations</td>
</tr>
<tr>
<td>4.2.3 Identification of differentially expressed genes</td>
</tr>
<tr>
<td>4.2.4 Identification of putative CDS regions with/without PTCs in unannotated transcripts</td>
</tr>
<tr>
<td>4.2.5 Identification of alternative splicing/processing events</td>
</tr>
<tr>
<td>4.2.6 Identification of orthologous alternative splicing/processing events</td>
</tr>
</tbody>
</table>
4.2.7 Quantification of alternative splicing/processing events .......................... 106
4.2.8 Identification of alternative splicing differences across conditions/species 107
4.2.9 Unsupervised clustering of gene expression and \( \Psi \) value time series .. 107
4.2.10 Gene ontology analysis of clustered genes and events ......................... 108
4.2.11 Classification of NMD events ..................................................... 108

4.3 Results .................................................................................. 109
4.3.1 Primate neurodifferentiation is accompanied by extensive AS changes . 109
4.3.2 Dynamic event alternative regions occur primarily in coding regions of protein coding genes ............................................................... 125
4.3.3 Orthologous dynamic events generally exhibit conserved temporal behavior ................................................................. 126
4.3.4 Alternative event orthologs involving CDS alterations behave more similarly than those that do not ........................................... 142
4.3.5 Conservation of alternative event dynamics increases as a function of PPI degree ................................................................. 148
4.3.6 NMD event dynamics are generally conserved and result in the relative increase of the NMD isoform inclusion ......................... 153

4.4 Conclusions and Discussion ................................................................ 159
4.4.1 In vitro neurodifferentiation is accompanied by an extensive alternative event dynamics ............................................................. 159
4.4.2 Orthologous events exhibit similar dynamics through neurodifferentiation 160
4.4.3 Coding region alterations and protein-protein interactions are associated with more conserved event dynamics ..................... 161
4.4.4 Increased relative abundance of putative PTC-containing transcripts with differentiation is conserved ................................. 162

Bibliography .......................................................... 164
## List of Figures

1.1 Central Dogma ................................................. 3  
1.2 Alternative event example ........................................ 4  
2.1 Schematic illustrating the overall procedure of junctionCounts ............... 10  
2.2 Alternative event classifications in junctionCounts .......................... 16  
2.3 Comparison of junctionCounts $\Psi$ estimates to published RT-PCR quantifications ............................................. 23  
2.4 Comparison of junctionCounts $\Delta\Psi$ estimates to published RT-PCR quantifications ............................................. 24  
2.5 Comparison of junctionCounts $\Delta\Psi$ estimates of simulated data to ground truth $\Delta\Psi$ across event types ............................................. 26  
2.6 Comparison of junctionCounts $\Delta\Psi$ estimates of simulated data to ground truth $\Delta\Psi$ across event types ............................................. 27  
2.7 Comparison of junctionCounts $\Delta\Psi$ estimates of simulated data to ground truth $\Delta\Psi$ ............................................. 28  
2.8 Comparison of junctionCounts $\Delta\Psi$ estimates of simulated data to ground truth $\Delta\Psi$ ............................................. 29  
2.9 Transcript with unannotated retained intron ........................................ 33  
2.10 Transcript with unannotated retained intron (zoomed) ............................ 34  
2.11 Schematic illustrating the overall procedure of cdsInsertion and findSwitchEvents ............................................. 35  
2.12 Counts of events with positive and negative $\Delta\Psi$ after UPF1 or UPF2 knockdown (lower threshold) ............................................. 39  
2.13 Counts of events with positive and negative $\Delta\Psi$ after UPF1 or UPF2 knockdown (higher threshold) ............................................. 40  
2.14 Fraction of predicted events with expected $\Delta\Psi$ sign after UPF1 or UPF2 knockdown as a function of threshold ............................................. 41  
3.1 Differentially sedimenting event counts in h9 ESC ........................................ 53  
3.2 Differentially sedimenting event counts in h9 NPC ........................................ 54  
3.3 Alternative first event in ATP5MD ............................................. 55  
3.4 Skipped exon event in PDZD11 ............................................. 55  
3.5 Alternative first event in FDPS ............................................. 56  
3.6 Cluster trends of differentially sedimenting events in h9 ESC ............................................. 57
3.7 Cluster trends of differentially sedimenting events in h9 NPC ............... 58
3.8 CDS-length ratios of events by cluster in h9 ESC .......................... 59
3.9 CDS-length ratios of events by cluster in h9 NPC .......................... 60
3.10 3′-UTR-length ratios of events by cluster in h9 ESC ....................... 61
3.11 3′-UTR-length ratios of events by cluster in H9 NPC ..................... 62
3.12 5′-UTR-length ratios of events by cluster in h9 ESC ....................... 63
3.13 5′-UTR-length ratios of events by cluster in H9 NPC ..................... 64
3.14 Count of putative NMD switch events by cluster in h9 ESC ............... 66
3.15 Count of putative NMD switch events by cluster in H9 NPC .......... 66
3.16 Skipped exon NMD event in NOVA1 ................................. 68
3.17 Skipped exon NMD event in AUH ................................ 68
3.18 Retained intron NMD event in DHRS1 ................................ 69
3.19 Retained intron NMD event in BDH2 ................................ 69
3.20 Included-excluded isoform CDS length ratios in all putative NMD switch events ............................. 70
3.21 Density of Cyto vs Mono ΔΔΨ between non-NMD proxies and NMD included form events in h9 ESC (0 threshold and additional CDS filtering) ............. 73
3.22 Count of event pairs with given sign of Cyto vs Mono ΔΔΨ between non-NMD proxies and NMD included form events in h9 ESC (0 threshold and additional CDS filtering) ............................. 74
3.23 Density of Cyto vs Mono ΔΔΨ between non-NMD proxies and NMD excluded form events in h9 ESC (0 threshold and additional CDS filtering) ............. 75
3.24 Count of event pairs with given sign of Cyto vs Mono ΔΔΨ between non-NMD proxies and NMD excluded form events in h9 ESC (0 threshold and additional CDS filtering) ............................. 76
3.25 CDS length ratios of proxies and NMD included form events in h9 ESC (0 threshold) with CDS filtration ........................................ 77
3.26 CDS length ratios of proxies and NMD excluded form events in h9 ESC (0 threshold) with CDS filtration ........................................ 78
3.27 Number of comparisons between proxies and NMD events in H9 ESC with expected ΔΔΨ sign (0 threshold and additional CDS filtration) ............. 79
3.28 Number of comparisons between proxies and NMD events in H9 NPC with expected ΔΔΨ sign (0 threshold and additional CDS filtration) ............. 80
3.29 Volcano plot of differentially expressed genes between ESC and NPC .......... 81
3.30 Included form NMD switch event min-max normalized Ψ values in ESC and NPC ........................................ 83
3.31 Excluded form NMD switch event min-max normalized Ψ values in ESC and NPC ........................................ 84
3.32 Count of NMD events that exhibit highest NMD isoform relative inclusion in a given fraction in H9 ESC and H9 NPC ............................. 85
3.33 Count of fraction pairs in which NMD events exhibit highest NMD isoform relative inclusion in H9 ESC and H9 NPC ............................. 86
3.34 Count of NMD events that exhibit highest NMD isoform relative inclusion in a given fraction in UPF3B knockdown and control cells ............................. 88
3.35 Count of NMD events that exhibit highest NMD isoform relative inclusion in a given fraction in UPF1 knockdown and control cells .................................................. 89
3.36 Count of fraction pairs in which NMD events exhibit highest NMD isoform relative inclusion in UPF3B knockdown and control cells .................................................. 90
3.37 Count of fraction pairs in which NMD events exhibit highest NMD isoform relative inclusion in UPF1 knockdown and control cells .................................................. 91
3.38 Count of NMD event transition direction in UPF3B knockdown and control cells .................................................. 92
3.39 Count of NMD event transition direction in UPF1 knockdown and control cells .................................................. 93
3.40 Number of NMD switch events exhibiting the expected $\Delta \Psi$ sign .................................................. 95

4.1 Schematic of Field et al experiment and analysis .......................................................... 110
4.2 Gene expression trends and enriched regulatory factors in human ........................................ 112
4.3 Counts of dynamic alternative events in all species .................................................. 114
4.4 Trends of differentiation-associated alternative event dynamics in human ................................ 115
4.5 Monotonicity bar plot .......................................................... 116
4.6 Monotonicity density .......................................................... 118
4.7 Differential expression of dynamic events .......................................................... 120
4.8 Maximum absolute log-fold change distribution of genes with dynamic events .................................. 121
4.9 Dynamics of MX event in MAPK8 .................................................. 122
4.10 XPO1 skipped exon event sashimi plot .......................................................... 123
4.11 Alternative events in regulatory genes heatmap .................................................. 124
4.12 Counts of dynamic alternative events in coding and non-coding genes ................................ 125
4.13 Counts of dynamic alternative events involving putative CDS alterations .................................. 126
4.14 Heatmap of ortholog $\Psi$ values events with high magnitude temporal changes .................................. 128
4.15 Orthologs generally exhibit similar $\Delta \Psi$ values in the same timepoint comparison in human and macaque cells .......................................................... 129
4.16 Orthologs generally exhibit similar $\Delta \Psi$ values in the same timepoint comparison in human and macaque cells .......................................................... 130
4.17 Orthologs generally exhibit the same sign $\Delta \Psi$ in the same timepoint comparison in human and macaque cells .......................................................... 131
4.18 Orthologs generally exhibit the same sign $\Delta \Psi$ in the same timepoint comparison in chimpanzee and orangutan cells .......................................................... 132
4.19 Orthologs generally exhibit $\Delta \Psi$ values in SC-W5 comparison in human and macaque cells .......................................................... 133
4.20 Orthologs generally exhibit the same sign $\Delta \Psi$ in the SC-W5 comparison in human and macaque cells .......................................................... 134
4.21 Orthologs generally exhibit similar $\Delta \Psi$ values in the SC-W5 comparison in chimpanzee and orangutan cells .......................................................... 135
4.22 Orthologs generally exhibit the same sign $\Delta \Psi$ in the SC-W5 comparison in chimpanzee and orangutan cells .......................................................... 136
4.23 Count of events with zero and non-zero $\text{max}|\Delta \Delta \Psi_{\text{inner}}|$ values in human-macaque .................................................. 137
4.24 Density of $\text{max}|\Delta \Delta \Psi_{\text{inner}}|$ values in human-macaque .................................................. 137
4.56 Density of Cyto vs Mono ΔΔΨ between non-NMD proxies and NMD included form events in h9 ESC (0 threshold) ........................................ 199
4.57 Count of event pairs with given sign of Cyto vs Mono ΔΔΨ between non-NMD proxies and NMD included form events in h9 ESC (0 threshold) .......... 200
4.58 Density of Cyto vs Mono ΔΔΨ between non-NMD proxies and NMD excluded form events in h9 ESC (0 threshold) ........................................ 201
4.59 Count of event pairs with given sign of Cyto vs Mono ΔΔΨ between non-NMD proxies and NMD excluded form events in h9 ESC (0 threshold) .......... 202
4.60 CDS length ratios of proxies and NMD included form events in h9 ESC (0 threshold) .......................................................... 203
4.61 CDS length ratios of proxies and NMD excluded form events in h9 ESC (0 threshold) .......................................................... 204
4.62 Number of comparisons between proxies and NMD events in h9 ESC with expected ΔΔΨ sign (0 threshold) ........................................ 205
4.63 Number of NMD switch events exhibiting the expected ΔΨ sign .......... 206
4.64 Chimpanzee gene expression cluster trends ..................................... 207
4.65 Chimpanzee gene expression cluster enriched ontologies ................. 208
4.66 Orangutan gene expression cluster trends ...................................... 209
4.67 Orangutan gene expression cluster enriched ontologies .................... 210
4.68 Macaque gene expression cluster trends ......................................... 211
4.69 Macaque gene expression cluster enriched ontologies ...................... 212
4.70 Chimpanzee alternative event cluster trends and enriched ontologies. .... 213
4.71 Orangutan alternative event cluster trends and enriched ontologies. ...... 214
4.72 Macaque alternative event cluster trends and enriched ontologies. ...... 215
4.73 Human-macaque PhastCons and temporal conservation of A3 events ..... 215
4.74 Human-macaque PhastCons and temporal conservation of A5 events ..... 216
4.75 Chimpanzee-orangutan PhastCons and temporal conservation of A3 events. 216
4.76 Chimpanzee-orangutan PhastCons and temporal conservation of A5 events. 217
4.77 Human-macaque phyloP and temporal conservation of SE/MS events .... 217
4.78 Human-macaque phyloP and temporal conservation of A3 events .......... 218
4.79 Human-macaque phyloP and temporal conservation of A5 events .......... 218
4.80 Chimpanzee-orangutan phyloP and temporal conservation of SE/MS events. 219
4.81 Chimpanzee-orangutan phyloP and temporal conservation of A3 events. .... 219
4.82 Chimpanzee-orangutan phyloP and temporal conservation of A5 events. .... 220
4.83 Human-macaque PhastCons and temporal conservation of SE/MS events ..... 220
4.84 Human-macaque PhastCons and temporal conservation of A3 events ...... 221
4.85 Human-macaque PhastCons and temporal conservation of A5 events ...... 221
4.86 Human-macaque PhastCons and temporal conservation of A5 events ...... 222
4.87 Chimpanzee-orangutan PhastCons and temporal conservation of SE/MS events. 222
4.88 Chimpanzee-orangutan PhastCons and temporal conservation of A3 events. 223
4.89 Chimpanzee-orangutan PhastCons and temporal conservation of A5 events. 223
4.90 Human-macaque phyloP and temporal conservation of SE/MS events ...... 224
4.91 Human-macaque phyloP and temporal conservation of A3 events .......... 224
4.92 Human-macaque phyloP and temporal conservation of A5 events . . . . . . . . . 225
4.93 Chimpanzee-orangutan phyloP and temporal conservation of SE/MS events . 225
4.94 Chimpanzee-orangutan phyloP and temporal conservation of A3 events . . . . 226
4.95 Chimpanzee-orangutan phyloP and temporal conservation of A5 events . . . . 226
4.96 Events with zero or non-zero $|\Delta \Delta \Psi_{inner}|$ by CDS-overlap status in human-macaque227
4.97 Events with zero or non-zero $|\Delta \Delta \Psi_{inner}|$ by CDS-overlap status in chimpanzee-
orangutan . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 228
4.98 $|\Delta \Delta \Psi_{inner}|$ by CDS-overlap status in human-macaque . . . . . . . . 229
4.99 $|\Delta \Delta \Psi_{inner}|$ by CDS-overlap status in chimpanzee-orangutan . . . . . . . . 230
List of Tables

4.1  Differential events in human and macaque  . . . . . . . . . . . . . . . . . . . . 142
4.2  Differential events in chimpanzee and orangutan  . . . . . . . . . . . . . . . . . 143
Abstract

Transcriptomic studies on the production and cytosolic fate of alternative isoforms in mammalian cells

by

Andrew Wallace

The vast majority of human genes are now known to (at least occasionally) give rise to multiple alternative isoforms. Alternative isoforms of the same gene are now known to have the potential to exhibit vast differences in RNA localization, stability, translational efficiency, and even function of their encoded protein products. Despite more than four decades of research in this area, however, the extent to which the properties of alternative isoforms are known remains minimal. The advent of next-generation sequencing technologies heralded a new era in the study of alternative isoforms. The increasing power and accessibility of technologies these techniques has allowed them to play a substantial role in pulling back the curtains and enable thorough cataloging efforts. Beyond the ability to simply identify and quantify transcripts in a high-throughput fashion across a host of different tissue and cell types, next-generation sequencing has enabled the creation of technologies that allow the measurement of (proxies of) translational efficiency and RNA binding protein binding profiles. In this work we describe our efforts in these directions. First, we introduce software we developed for the identification, quantification, and contextual elaboration of events that distinguish alternative isoforms. We also highlight two efforts towards leveraging and increasing the power of next-generation sequencing to illuminate the properties of alternative isoforms: 1) the application and charac-
terization of Frac-seq applied to early neurodifferentiation, and its implications for studying isoform-specific translation efficiency and nonsense-mediated decay, and 2) characterization of RNA-seq applied to in vitro models of human and close primate brain development, and its implications for alternative event dynamics through differentiation.
This work is dedicated to Elizabeth Otto, without whose support and love it would never have been completed.
Acknowledgments

I would like to thank current and former members of the Sanford Laboratory, with particular mention of Aishwarya Jacob, Julia Philipp, and Neda Ronaghi. Jolene Draper performed the Frac-seq experiments described in chapter 3 and made the initial observations that putative NMD substrates are predominantly light fraction-enriched and exhibit down-gradient shifts in NPC.

I would also like to thank my committee. Michael Stone, Todd Lowe, and Christopher Vollmers were extremely helpful to me both within and outside of science, and I am extremely grateful to them. Further, I would like to thank Sol Katzman, whose mentorship was invaluable.
Chapter 1

Introduction

The central dogma of molecular biology is a widely stated paradigm that characterizes the principle information flow executed by life’s macromolecules (Crick, [20], [21]). Stated simply, the central dogma posits that DNA passes information to itself (via replication) and to RNA (via transcription), which in turn passes information into proteins via translation (figure 1.1). It is worth noting that even in 1970, the additional pathways of RNA passing information to itself (via RNA replication by RNA-dependent RNA polymerases) and to DNA (via reverse transcription) were considered to be plausible and even contextually likely, though rare in most cellular circumstances [21]). While innumerable complicating details have been distilled in the 6 decades since its original statement, the central dogma does remain fundamentally representative of the principle information flow that gives rise to the sequence of macromolecules in the cell (with the exception of the dotted line indicating DNA translation, which is possible with restrictions under specific in vitro settings [63], but unlikely to occur in the cell due to structural requirements for a 2′-OH in certain steps [35]). It is nevertheless important to highlight
the general nature of those complicating details to gain an appropriate appreciation of those complicating details. Each of the arrows in figure 1.1 are modulated on a case-by-case basis by other macromolecules. For instance transcription of DNA into RNA is modulated by by protein transcription factors (reviewed in e.g. Lambert et al [50] in human) and the post-translational modification state of chromatin histone proteins (reviewed in e.g. Lawrence et al [52]). Translation of mRNA into protein is modulated not only by the concentration of the mRNA in question, but also by the availability and post-transcriptional modification state of the tRNAs needed to decode its codons, the identity and concentration of RNA-binding proteins (RBPs) that interact with it, and the identity and concentration of miRNAs that could mediate its degradation or suppress its translation [46]. Thus, while proteins may never get “reverse translated” into RNA or DNA sequences that encode them, they certainly impact all of the transitions in the central dogma. RNA is also now known to play a pan-regulatory role including as a catalyst in tRNA maturation [36], protein translation [17], and mRNA splicing [78]. It essentially has begun to appear as though everything regulates everything, and a central dogma diagram annotated with all regulatory relationships would rapidly become an inscrutable tangle.
What remains clear, however, is that the identity, concentrations, and spatial distribution of macromolecules in the cell determine, or at least substantially contribute to its ability to maintain homeostasis under varied conditions, and to enact its specialized function. It is thus of critical importance to understand the entirety of the transcriptome, and in turn the entirety of the proteome. Understanding the entirety of the transcriptome, particularly the metazoan transcriptome, is known to involve appreciating the production of alternative transcripts [81]. Alternative transcripts can be produced in eukaryotes by alternative transcription start
sites in which unique transcription start sites give rise to transcripts with distinct 5′-termini, alternative splicing, in which transcript exons can be stitched together in different combinations, and alternative polyadenylation, in which unique transcription termination sites can give rise to transcript with distinct 3′-termini. These three mechanisms can occur alone or in combination with one another, and are now known to impact a large proportion of the human transcriptome [71]. The differentiating features of pairs or groups of alternative transcripts are often referred to as alternative events (figure 1.2, and consist of sub-transcripts that typically contain a region or regions unique to one or more transcripts and the exons that flank those regions. Alternative events are typically viewed as easier to quantify and as most relevant to the determinants that regulate their production compared with the full-length transcripts from which they are derived.

![Transcripts and Events Diagram](image)

Figure 1.2: Examples of alternative events (bottom) and the full-length transcripts from which they are derived. The events are derived from the sub-transcripts that are boxed with the same color and line style.

The consequences of alternative transcripts are far from globally characterized, but
are known at least to be manifold. Alternative transcripts are known to have different half-lives [53], intracellular localizations (e.g. Ilouz et al [41]), translation efficiencies [95], and most obviously, can produce different protein isoforms (e.g. Rosenfeld et al [73]). It is thus not enough to know that a given gene is expressed: one must also know which isoforms of that gene are being produced. The advent of next generation sequencing has greatly accelerated the discovery of alternative transcripts and the identification of the context in which they are produced [7]. A number of studies have applied next generation sequencing (in this case RNA sequencing, or RNA-seq) to characterize and highlight the tissue-specificity of many alternative isoforms [94], [66]. Next generation sequencing has also paved the way for studies on alternative isoform regulation: CLIP-seq has illuminated the regulatory patterns of a number of RNA binding proteins [56]. It further has provided new approaches to elucidate the cytoplasmic fate of transcripts [95].

Despite these recent advances, much work remains necessary in order to fully characterize the identity, context, functional importance, and behavior of alternative isoforms in the transcriptome. Techniques of next-generation sequencing have advanced dramatically over the nearly two decades since their inception. Application of these techniques in novel biological comparisons can provide indirect insight into the appropriate context and functional importance of alternative isoforms. Moreover, techniques to investigate transcript fates are still in their relative infancy.

In this work we describe two efforts to address the gap in our understanding of alternative isoforms. In chapter 3 we apply Frac-seq, a technique in which velocity gradient-fractionated cytoplasm is subject to RNA-seq [82], to human embryonic stem cells and human
neural precursor cells. We consider the ability of Frac-seq and related techniques to quantify isoform-specific translation efficiency, and we characterize differentially sedimenting events in those two cell types. We further highlight the behavior of putative nonsense-mediated decay (NMD) substrates in these experiments, and posit the utility of Frac-seq in characterizing NMD.

Efforts to understand the transcriptome in its entirety must include interrogations of isoform-specific translation efficiency. Since the feasibility of quantifying isoforms from velocity gradient fractions was first demonstrated by Sterne-Weiler et al [82], a handful of efforts have built on on the approach (discussed in more detail in chapter 3).

In chapter 4 we analyze RNA-seq data derived from in vitro-developing cortical neurospheres in human and three close primate species. We assess the dynamics of alternative events along with the conservation of their behavior among the four species. We further report characteristics of events that appear to be more, or less dynamically conserved, and highlight some instances of poor conservation.

Through these efforts, we attempt to contribute to our global understanding of alternative isoform production and significance.
Chapter 2

RNA-seq utilities junctionCounts and cdsInsertion

Our transcriptomic studies are fundamentally reliant on two faculties: 1) the identification and quantification of alternative processing events, and 2) prediction of functional consequences implied by a change in the relative inclusion of an alternative event. While many tools exist for the former, none had the combinations of characteristics we sought, namely the identification and simple, junction read-based quantification of binary alternative events of arbitrary type. While a small number of tools exist for the latter task, they come with significant limitations that render them unsuitable for our needs. To address these issues, we developed junctionCounts and cdsInsertion.

junctionCounts identifies pairwise alternative splicing events in transcriptome annotations and quantifies them using exon-exon and exon-intron junction reads in RNA-seq data. cdsInsertion annotates putative CDS regions based on in-silico translation of transcripts from
overlapping annotated start codons. It then uses these CDS annotations to identify transcript characteristics such as UTR length and presence of a PTC. An included utility findSwitchEvents uses this information to generate hypotheses as to the functional consequences of changes in alternative splicing outcomes. For instance, cases in which the included form of an event is consistent with PTC-containing transcripts (and thus putative NMD substrates) while the excluded form of an event is consistent with non-PTC-containing transcripts. A change in such an event might imply regulation of NMD.

2.1 junctionCounts

2.1.1 Rationale for junctionCounts

While many tools for the identification and quantification of alternative splicing exist, we were in need of one with all of the following characteristics:

1. The ability to identify events of arbitrary type:

   • In preliminary examinations of data under investigation (see subsequent chapters) we identified a number of alternative splicing events with a clear binary shift in relative abundance whose properties did not clearly match those of canonical event types (e.g. DLEU2, OGDH, DLG3 - SHOW FIGURES). While I am aware of at least one example of software with the ability to identify such events, ASTALAVISTA [31], its output format would require extensive processing to be used in my desired context (i.e. it does not output a standard GTF/GFF3 file as has become the norm). A number of other tools perform canonical event identification (e.g.
SUPPA, MISO tool, SplAdder etc), but most of them are limited in the types of events they will identify (typically only canonical events and often excluding events involving transcript termini) and/or cannot be run independent of other steps in a larger pipeline.

2. Ability to quantify arbitrary pairwise events in a reasonable amount of time using junction reads whenever possible:

- Given the availability of annotations (e.g. generated via the above functionality) I also required a module capable of quantifying them. In particular, I required a method capable of tallying counts of junction reads consistent with these events that a) handles the possibility of splice junctions occurring in an overlapping region of paired-end reads, b) handles the possibility of multiple junctions present in a single fragment (and takes advantage of this scenario to reduce ambiguity of read assignment), c) does not discard large numbers of reads due to the presence of (a reasonable number of) mismatches/indels, d) does not require that both mates in a paired-end fragment have the same length, e) produces readily interpretable output that can be analyzed by common statistical methods, f) calculates a $\Psi$ estimate for each combination of included and excluded junctions, and g) runs in a reasonable amount of time. To the best of my knowledge, no extant method fulfills all of these criteria while handling pairwise events of arbitrary definition.
2.1.2 Alternative event definition in junctionCounts

In junctionCounts, alternative events are simply 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 common are separated by any two combinations of distinct exon coordinates. Examples of these scenarios are illustrated in figure (FIGURE). Note that cases in which the aforementioned pairs are separated by more than two sets of distinct exon coordinates, all pairwise combinations of those sets will be considered distinct alternative events. In cases where more two or more events defined in this way share the same
splicing structure, these events will be collapsed into a single representative event in which the most proximal outer exon boundaries are used (e.g. figure FIGURE). As discussed below, this approach to event identification encompasses standard alternative event types and further incorporates identifies non-standard events of arbitrary exon structure.

2.1.3 Alternative event classification in junctionCounts

The majority of event types in junctionCounts correspond to types either implicitly or explicitly defined elsewhere (e.g. Breitbart et al 1987 [13]). junctionCounts also adopts previous usage of the term ”complex” (e.g. Sammeth et al (2008) [74], Huelga et al [40]) to refer to non-standard event types, distinguishing between internal (CO), 5′-terminal (CF), and 3′-terminal (CL) contexts. Full descriptions of the criteria for each event type are provided below:

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

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

**A3 - alternative 3′-splice site**: an event in which each isoform consists of two exons, and the upstream exon is common to both isoforms. Further, the 3′-boundary of the downstream exon is also common to both isoforms, but critically, the 3′-splice site is distinct. The form with the
most upstream of the alternative 3′-splice sites is the included form.

**A5 - alternative 5′-splice site:** an event in which each isoform consists of two exons, and the downstream exon is common to both isoforms. Further, the 5′-boundary of the upstream exon is also common to both isoforms, but critically, the 5′-splice site is distinct. The form with the most downstream of the alternative 5′-splice sites 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. In this case, the only requirement for the two exons being distinct is that they do not share either splice site. Consequently, 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 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.

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

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

**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 to the terminal exons is common to both isoforms. The form with the most downstream 3′-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.

**AT** - *alternative transcription start site tandem UTR*: an event in which 5′-terminal exons in two isoforms have distinct 5′-termini but common 5′-splice sites. The form with the most upstream 5′-terminus is the included form.

**AP** - *alternative polyadenylation tandem UTR*: an event in which 3′-terminal exons in two isoforms have distinct 3′-termini but common 3′-splice sites. The form with the most downstream 3′-terminus 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 (i.e. the sum of the lengths of the involved exons) 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.

**UF - unique single 5′-terminal exon:** in this event one isoform is distinguished by a 5′-terminal exon. The exon immediately downstream in the same isoform is common to both isoforms, but serves as the 5′-terminal exon in the other isoform. Essentially, there is the implicit requirement that the 5′-boundary of the common exon serves as a transcription start site in one isoform and a 3′-splice site in the other. It is unknown if this situation is supported by other evidence or if these events are just artifacts of existing transcriptome annotations. The form with the most upstream 5′-terminus is the included form.

**UL - unique single 3′-terminal exon:** in this event one isoform is distinguished by a 3′-terminal exon. The exon immediately upstream in the same isoform is common to both isoforms, but serves as the 3′-terminal exon in the other isoform. Essentially, there is the implicit requirement that the 3′-boundary of the common exon serves as a polyadenylation site in one isoform and
a 5′-splice site in the other. It is unknown if this situation is supported by other evidence or if these events are just artifacts of existing transcriptome annotations. The form with the most downstream 3′-terminus is the included form.

For instance, MS events can have any non-unitary number of alternative exons in their included form. Similarly, MR events can have any non-unitary number of retained introns in their included form. MF and ML events can have any non-unitary number of alternative exons in either isoform. CO, CF, and CL events are broad catch-all classes referring to internal, 5′-terminal, and 3′-terminal events respectively that do not meet the criteria for other more specific classes.
2.1.4 Evaluation of junctionCounts event identification and classification

As the ground truth of transcriptome content is not known, evaluation of event identification is not strictly possible. However, we can establish whether junctionCounts-identified events are consistent with those identified by other tools, in event types that are handled by both.
Generally speaking, it is expected that junctionCounts should identify a superset of events identified by other programs that operate using similar event definitions. We expect that junctionCounts should, for instance, identify the same set of skipped exon events as any other method given the same input transcriptome annotation. junctionCounts, however, will inevitably contain unique events due to its handling of less common and arbitrary event types (i.e. MS, CO etc).

In order to perform this assessment, we generated alternative events from the GENCODE human transcriptome annotation (v30, [39]) using SUPPA2 (Trincado et al [90]), SplAdder (Ratsch et al [47]), and junctionCounts. We then matched events between junctionCounts and the other two utilities, considering events with the same set of exon-exon and exon-intron junctions in each isoform to be identical. We find that all three programs identify a similar set of events, with the greatest similarity in SE event definition. junctionCounts identifies a strict superset of the SE events identified by SplAdder and SUPPA2, with SUPPA2 identifying 13041 of the 13224 SE events identified by junctionCounts and SplAdder in turn identifying 7980 of the events identified by both SUPPA2 and junctionCounts. Differences were observed both here, and in RI events due to the fact that SUPPA2 appears to require that the flanking exons have identical outer coordinates. Large overlaps were observed in other event types, with many events unique to particular methods. In the case of junctionCounts and SUPPA2, differences in event type definition clearly reduced the size of the overlap in the cases of AF, AL, A3, and A5 events. In order for an event to be identified as an AF, rather than an A5 event, for instance, junctionCounts only requires that the isoforms contain a unique splice site and a unique putative transcription start site. SUPPA2 in contrast requires AF and AL events to possess non-
overlapping alternative exons. We believe that merits can be argued for either approach. As discussed above junctionCounts further identifies a host of event types not considered by the other approaches. SplAdder identified in general far fewer events than the other two methods, for reasons that remain unclear. It also uniquely identified certain events that, upon manual inspection, did not appear to be supported by junctions found in the input transcriptome. It is unclear why this was the case, insertion of new exons and introns was disabled in our comparison. Broadly our inspection these results gave us confidence in the events generated by junctionCounts.

2.1.5 Alternative event quantification in junctionCounts

junctionCounts employs a simple junction read-centric approach to alternative event quantification. For each read, junctionCounts considers matches between splice junctions and event splice junctions, as well as overlaps between contiguous mapped read sequence and informative exon-intron junctions. Informative exon-intron junctions are those that are overlapped by an exon in the alternative isoform. Reads overlapping such an exon-intron junction are therefore consistent with the alternative isoform. Key examples of this occur in the excluded isoform of RI events, which are overlapped by the monolithic exon of the included form.

Having established all of 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 (whenever possible). In this way junctionCounts goes beyond simple junction-by-junction read counting performed in some methods (e.g. SplAdder [47]). junctionCounts does not at present implement any statistical methods to fully disam-
biguate assignment of reads containing only junctions consistent with more than one isoform. Nonetheless, it should be noted that 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. Subsequently, a $\Psi$ value is calculated for all pairwise combinations of included and excluded junction counts (i.e. the ratio between a given included form junction’s counts and the sum of those counts and a given excluded form junction’s counts):

$$\Psi_k = \frac{n_i}{n_i + m_j}$$ (2.1)

Where $n_i$ is the number of reads assigned to included form junction $i$ and $m_j$ is the number of reads assigned to the excluded form junction $j$. In this way, a set of $\Psi$ values established for each sample. Taking the minimum:

$$\Psi_{\text{min}} = \min \Psi_k$$ (2.2)

and the maximum:

$$\Psi_{\text{max}} = \max \Psi_k$$ (2.3)

of these values enables the calculation of a $\Psi_{\text{span}}$ as follows:

$$\Psi_{\text{span}} = \Psi_{\text{max}} - \Psi_{\text{min}}$$ (2.4)
The $\Psi_{span}$ can thus serve as a rough measure of within-sample uncertainty. junctionCounts further report those values of $n$ and $m$ that gave rise to the calculation of both $\Psi_{min}$ and $\Psi_{max}$, so a user may take advantage of either or both in a count-based statistical method.

As a further optional method of assessing within-sample uncertainty, junctionCounts offers the possibility of bootstrap quantification, in which the program repeats a user-specified number of rounds of bootstrap read selection and re-quantification. For each bootstrap round junctionCounts reports all of the same aforementioned information, in addition to the initial non-resampled quantification.

### 2.1.6 Evaluation of junctionCounts quantification

#### 2.1.6.1 Methods

**Data sources**

junctionCounts quantifications are based on published RNA-seq data from Shen et al (2014) [77] and Zhang et al (2014) [103]. RT-PCR data matching the RNA-seq from Shen et al was derived from the same work, while RT-PCR data matching Zhang et al was derived from Vaquero-Garcia et al (2016) [92] who replicated the experiments in Zhang et al.

**Mapping and transcriptome assembly**

Data were mapped with STAR (v 2.7 [23]) and the resulting BAM files were used for transcriptome assembly using StringTie (v 1.3.4d [68]). A consensus transcriptome was created using StringTie merge.
Alternative event identification and quantification

Alternative events were identified from the consensus transcriptome GTF and quantified using splice and exon-intron junction reads with junctionCounts (v 0.1.0). junctionCounts identified events were matched (unambiguously in almost all cases) with event junction coordinates provided by Shen et al and Vaquero-Garcia et al. It is worth noting that events in Vaquero-Garcia et al are defined differently and do not correspond strictly to events defined in junctionCounts. However, the Vaquero-Garcia et al events (termed local splice variants) corresponded one-to-one with a single included and excluded form splice junction of a junctionCounts SE event. One event quantified by Vaquero-Garcia et al could not be identified by junctionCounts. $R^2$ values were calculated between the mean of junctionCounts $\Psi$ midpoints and the mean of RT-PCR $\Psi$ values. A similar calculation was performed for $\Delta \Psi$.

Simulation of RNA-seq experiments

Simulated RNA-seq dataset TPM values were derived from kallisto (Bray et al [12]) quantification of the Zhang et al [103] and Shen et al [77] data. Transcript read counts were back-calculated from the TPM values, and used as input to Polyester (Frazee et al [33]). Reads were simulated with default end bias and RNA fragmentation assumptions. In this way (at least somewhat) realistic TPM values were used as ground-truth input for simulating RNA-seq data. Ground-truth $\Psi$ values were calculated with the following equation:

$$\Psi = \frac{\Sigma_{i=1}^n TPM_i}{\Sigma_{i=1}^n TPM_i + \Sigma_{j=1}^m TPM_j}$$  \hspace{1cm} (2.5)

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. $\Psi$ and $\Delta\Psi$ values were calculated using custom Python scripts, leveraging the event-transcript associations output in the junctionCounts IOE file.

2.1.6.2 Results

**Comparison of junctionCounts $\Psi$ values to RT-PCR $\Psi$ values**

In order to assess the correspondence of junctionCounts-quantified RNA-seq data to alternative non-NGS-based event quantifications, we took advantage of the existence of a number of published RNA-seq datasets with corresponding RT-PCR quantifications of alternative events either from the same experiment (Shen et al [77]) or a replicated experiment (Vaquero-Garcia et al [92] replicated Zhang et al [103] to generate RT-PCR quantifications). We mapped the data, generated an augmented transcriptome annotation, and finally identified and quantified alternative events using junctionCounts. We were able to establish a correspondence between all but one of the events quantified by RT-PCR by either Shen et al or Vaquero-Garcia et al, despite the difference in event definition in Vaquero-Garcia et al. We find an overall high correspondence between junctionCounts $\Psi$ values and the reported RT-PCR $\Psi$ values ($R^2 = 0.91$ for both experiments combined, figure ??).
We further find an event better correspondence between junctionCounts-determined \( \Delta \Psi \) values and reported RT-PCR-derived \( \Delta \Psi \) values (\( R^2 = 0.97 \) for both experiments combined, figure 2.4).
Figure 2.4: Comparison of junctionCounts $\Delta \Psi$ estimates to published RT-PCR quantifications.

**Comparison of junctionCounts $\Psi$ values ground truth-simulated $\Psi$ values**

In order to gain a broader understanding of junctionCounts’ quantification performance across event types and expression levels, we simulated RNA-seq data using Polyester
[33]. We then analyzed the simulated data with our standard approach, using junctionCounts to quantify alternative events. We compared the junctionCounts quantifications to the ground truth quantifications, and we find that junctionCounts-derived $\Delta \Psi$ values broadly agree with ground truth $\Delta \Psi$ values. This is particularly true of events with junctionCounts-derived $|\Delta \Psi_{\text{inner}}| > 0$ (which is to say that none of the $\Psi$ values among the two conditions overlap). A linear model fit between these values and the ground truth $\Delta \Psi$ values has an $R^2 = 0.85$ (figure 2.5).
Figure 2.5: Comparison of junctionCounts $\Delta \Psi$ estimates of simulated data to ground truth $\Delta \Psi$ based on Zhang et al [103] data

Events with higher magnitude $|\Delta \Psi_{inner}|$ values show even greater agreement (e.g. $R^2 = 0.91$ among events with $|\Delta \Psi_{inner}| \geq 0.1$ (figure 2.6)).
We further established that $\Delta \Psi$ value correlation is generally high across event types, but is typically higher for internal events than terminal events. Internal event types exhibit a minimum $R^2$ of 0.66 (corresponding to A3 events) given $|\Delta \Psi_{inner}| > 0$ and 0.82 given $|\Delta \Psi_{inner}| \geq 0.1$ (figure 2.8). Meanwhile terminal events exhibit a minimum $R^2$ of 0.62 and 0.77 at the two...
thresholds corresponding to MF and CL event types respectively.

Correlation of junctionCounts to simulated data with known ground truth $\Psi$

Given junctionCounts inner $|\Delta \Psi| > 0$ and $\geq 20$ junction reads

<table>
<thead>
<tr>
<th></th>
<th>A3</th>
<th>A5</th>
<th>AF</th>
<th>AL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.66</td>
<td>0.83</td>
<td>0.67</td>
<td>0.7</td>
</tr>
<tr>
<td>$p$</td>
<td>2.8e-32</td>
<td>6e-27</td>
<td>2.3e-33</td>
<td>1.8e-33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CF</th>
<th>CL</th>
<th>CO</th>
<th>MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.71</td>
<td>0.69</td>
<td>0.95</td>
<td>0.62</td>
</tr>
<tr>
<td>$p$</td>
<td>3.3e-24</td>
<td>4.5e-6</td>
<td>1.8e-17</td>
<td>9.5e-05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>MS</th>
<th>MX</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.9</td>
<td>0.94</td>
<td>0.98</td>
<td>0.87</td>
</tr>
<tr>
<td>$p$</td>
<td>6.8e-48</td>
<td>2.8e-50</td>
<td>1e-17</td>
<td>2.8e-46</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.92</td>
</tr>
<tr>
<td>$p$</td>
<td>8.700060e-318</td>
</tr>
</tbody>
</table>

Figure 2.7: Comparison of junctionCounts $\Delta \Psi$ estimates of simulated data to ground truth $\Delta \Psi$ based on Zhang et al [103] data
Figure 2.8: Comparison of junctionCounts $\Delta \Psi$ estimates of simulated data to ground truth $\Delta \Psi$ based on Zhang et al [103] data
2.1.7 Discussion and Future Directions

2.1.7.1 junctionCounts is an accessible method for identification and accurate quantification of alternative events

Broadly we find that junctionCounts is a useful approach for identifying both canonical and non-canonical alternative events. Moreover, junctionCounts exhibits good performance in recapitulating both RT-PCR-derived $\Psi$ values and ground-truth $\Psi$ values from simulated data.

2.1.7.2 The utility of the binary event-centric perspective

The concept of the binary alternative event pervades the alternative splicing literature, and has been a primary object of quantification in transcriptomic experiments focused on alternative splicing for more than a decade. Key arguments for their use include the following: 1) the facility with which they can be accurately quantified relative to full-length transcripts, 2) the greater likelihood that they accurately represent (a subset of) transcript structure as compared with full-length transcripts, and 3) their exclusion of gene segments not relevant to the regulation of the event (i.e. introns and exons outside and distal to the event).

Arguments 1 and 2, while logically premised, have only been tested in a handful of software-specific evaluations (e.g. Sterne-Weiler et al 2018 [83], Lin and Krainer 2019 [57]). Further, it is likely the case that this argument will decrease in validity as improving long-read sequencing approaches provide better and better representations of the ground-truth expressed transcriptome.

Argument 3 can in principle withstand technological improvements, though it is worth
noting that the contribution of factors not necessarily local to the event itself have been reported
to be important to alternative event regulation (discussed in review by Witten et al [98], e.g.
transcription elongation [32]). Nonetheless, focusing on the site of the alternative event 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, and future efforts to understand alternative
event regulation will likely benefit from considering this perspective.

We note, however, that a key weakness of the traditional binary event is the known
existence of loci in which more than two alternative sub-transcripts overlap and are subject to
simultaneous change in relative abundance. While such non-binary events could (and typically
are) 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 recent efforts, such as MAJIQ [92] and Whippet [83]
have attempted to address this issue with several approaches. Future efforts must continue
to embrace and expand upon non-binary events in order to properly represent the regulatory
context associated with processing decisions.

2.1.7.3 A proposal for a unified binary and non-binary event identification approach

The approach to event definition taken by junctionCounts can be modified to permit
the inclusion of non-binary events. We suggest the possibility of defining events largely as in
section 2.1.2, but adding the requirement that the event includes segments of all overlapping
transcripts on the same strand, and permitting as many unique intervening exonic sequences as
there are segments. These criteria allow for traditional (and non-traditional) binary events in
largely the same form as they are usually presented, while still encompassing events of greater complexity. Moreover, events defined in this way could be simplified by the removal of segments with measured relative abundance, and re-classified after trimming away any outer exonic sequence newly common to all segments. Future iterations of junctionCounts will build in these ideas.

2.2 cdsInsertion and findSwitchEvents

2.2.1 Rationale for cdsInsertion

Reference genome-guided transcriptome assembly is a widely used step in the analysis of RNA-seq experiments. In this process, transcript structure is predicted from the data with or without the use of a reference transcriptome annotation as a template. This process allows analysts to consider putatively novel transcripts that may be important to the biological phenomena under study. Unfortunately, common tools for this task (e.g. Cufflinks [88]) and StringTie [68]) are unable to provide any information on the presence and nature of any ORFs that may be contained within these predicted transcripts (e.g. figure 2.9).

There is at least one available tool, Transdecoder (Haas et al [37]) that somewhat addresses this limitation. Transdecoder, however, was developed for use with Trinity (also described in Haas et al [37]) which is intended for completely de novo transcriptome assembly in the total 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 cases of immediate interest to us, however, exist-
ing ORF annotations are available for the majority of genes in which new transcripts might be identified. In these cases it is of primary importance to examine novel coding sequences that begin with high-confidence annotated start codons. This prediction approach is useful not only because of the fact that it informs on potential novel peptides, but also because it has the ability to identify the presence of premature termination codons (PTCs) (e.g. figure 2.10), 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 the host transcripts via nonsense-mediated decay and non-stop decay respectively. ORF prediction tools such as Transdecoder are based on the assumption that translation of the sequence will yield a reasonable and stable protein with some extent of similarity to other known proteins, and will thus likely miss many examples of ORFs that have important consequences for their host transcripts despite never yielding an abundant or functional protein.

Figure 2.9: Transcript with unannotated retained intron identified by StringTie.
2.2.2 Rationale for findSwitchEvents

Beyond the prediction of coding features and their implications, we have need of the ability to infer the potential functional implications of binary alternative splicing events that may be consistent with or partly/completely defined by novel transcripts defined above. One such implication is nonsense-mediated decay. For instance, if all transcripts consistent with the included form (i.e. all transcripts of whom the included form is a strict subset) of the alternative event contain a predicted PTC, while none of the transcripts consistent with the excluded form contain a PTC, then a change in relative inclusion of the event might indicate a modulation of NMD. Other situations might involve a coding-to-noncoding switch, or a long-to-short UTR switch, or the inclusion of a rare codon. findSwitchEvents thus serves to bridge the gap between full transcript- and event-level analysis, allowing one to take advantage of the potential superior quantification accuracy (e.g. [83]) and regulatory interpretation of event-level analysis while still leveraging information that can only be derived from full-length transcripts.
2.2.3 Approach to CDS insertion

To address these limitations we wrote cdsInsertion, which translates provided putatively transcripts in silico from user-provided overlapping start codons, and determines resulting transcript characteristics such as UTR lengths, putative protein sequences, the presence of PTCs, PTC distances from downstream splice sites, and more.

Figure 2.11: Schematic illustrating the overall procedure of cdsInsertion and findSwitchEvents.

For a given codon and transcript, cdsInsertion first checks whether the genomic coor-
dinates of the start codon’s first position overlap the genomic coordinates of a transcript’s exons. If they overlap, it is then checked the putative start codon is in fact AUG in the spliced transcript’s sequence. At this time non-AUG initiation is not supported. If the start codon sequence is AUG, cdsInsertion will translate the spliced transcript in silico by looking for an in-frame downstream stop codon (i.e. one of UAA, UGA, UAG). If a downstream stop codon is found, the resulting CDS is recorded, and associated information such as CDS length, CDS sequence, putative protein sequence, UTR lengths and sequences, PTC presence and PTC distance are all, among others, associated to the transcript. If more than one CDS is found within the transcript, additional CDS features are associated to the transcript as distinct CDS features. No attempt is made to prioritize one feature over another. If no in-frame downstream stop codon is found, the transcript is recorded as a possible non-stop decay substrate.

After repeating this process for all transcripts, cdsInsertion outputs a table with summary information about each transcript, as well as putative protein sequences for each CDS, and a separate GTF file for putative non-PTC, PTC, and non-stop transcript-CDS combinations. In the GTF files, a separate transcript record is written for every CDS-transcript combination. Optionally, cdsInsertion will additionally output bigGenePred files which enable codon visualization if viewed on the UCSC Genome Browser. cdsInsertion further outputs a pickled Python dictionary containing all of the aforementioned information associated with each transcript.

2.2.4 Assessment of implications for alternative events using findSwitchEvents

A primary drawback of event-centric analysis is the fact that it complicates the consideration of an event’s functional consequences. For instance, it is typically not possible to
discern whether a skipped exon contains a PTC or not, because that information depends on a) whether the exon overlaps a CDS and b) the reading frame of the overlapping CDS. For the most part, this information is not (solely) local to the event itself, which is to say that the CDS in the example is unlikely to be confined entirely to the skipped exon and its flanking exons. To access this information, one must therefore consider the transcripts of which the event isoforms are subsets. In order to do this, one must have an approach to associate an event to all of its transcripts and their properties.

findSwitchEvents is such an approach, which takes advantage of IOE files (a format originally introduced in SUPPA, Alamancos et al [1]) generated by junctionCounts to associated events with transcripts, and a pickled Python dictionary output by cdsInsertion containing information on transcript CDS and associated details.

2.2.5 Evaluation of NMD switch events identified using cdsInsertion and findSwitchEvents

In order to simultaneously evaluate the success of NMD switch event identification by cdsInsertion and findSwitchEvents, we leveraged ENCODE ([24], [22]) experiments in which NMD factors UPF1 and UPF2 were knocked down in K562 cells (experiments ENCSR174OYC, ENCSR251ABP, ENCSR438MDN, and ENCSR810FHY). Although we cannot a priori know the true status of a given set of transcripts as NMD targets, we expect that the majority of the time the relative abundance of NMD target isoforms will increase upon depletion of NMD factors. In order to assess whether this was the case, we identified and quantified alternative events in the ENCODE RNA-seq data after performing reference aided transcriptome assembly.
with StringTie [68] using GENCODE as input [39]. We then identified putative NMD switch events using cdsInsertion and findSwitchEvents, and assessed their behavior. We observe these putative NMD switch events to exhibit increases in the relative inclusion of the putative PTC-containing isoforms in both the UPF1 and UPF2 knockdown experiments (figures 2.12 and 2.13), and the extent to which this is the case increases as a function of the magnitude of the $\Delta \Psi$ and the minimum number of junction reads in the comparison (figure 2.14).
Figure 2.12: Counts of events with positive and negative $\Delta \Psi$ after UPF1 or UPF2 knockdown given $|\Delta \Psi| > 0$ and greater than 15 junction counts for any sample in the comparison.
Figure 2.13: Counts of events with positive and negative $\Delta \Psi$ after UPF1 or UPF2 knockdown given $|\Delta \Psi| > 0.1$ and greater than 15 junction counts for any sample in the comparison.
2.2.6 Discussion and Future Directions

In an ideal case, studies which intend to consider translation and its implications on a transcriptome-wide scale would include an experimental technique to define CDS regions or start codons at the least empirically. For example, ribosome profiling or its later variant, TI-seq (i.e. ribosome profiling in the presence of harringtonine or similar inhibitors that prevent the ribosome’s first translocation but do not otherwise interfere with elongation, described in detail by Zhang et al [102]) 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 com-
plexity of these approaches, tools such as 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 their behavior. Its partner tool, findSwitchEvents, attempts to infer alternative event characteristics from those of its constituent transcripts. For instance, it predicts whether or not an event can be considered an NMD switch event. We evaluated those particular predictions using ENCODE data and found broadly that our results support the assertion that findSwitchEvents is predicting NMD switch events, and is thus a useful approach for the interpretation of alternative event behavior.
Chapter 3

Measurement and interpretation of isoform-specific polysome co-sedimentation using Frac-seq

3.1 Introduction

Alternative transcripts of a given gene differ in a number of key properties. One of the possible functional implications of a change in isoform composition is a difference in translation efficiency.

3.1.1 Brief introduction to translatomics and the rationale for Frac-seq

It has long been reported that measured gene expression values are imperfectly correlated to measured protein concentrations (e.g Schwanhausser et al [76]). While a host of technical artifacts conceivably contribute to the reported discrepancies, and the degree of the
discrepancy varies considerably among reports, one known contributing factor is the translation efficiency of the gene’s transcripts. Consequently, the development of methods to directly measure translation efficiency in a high-throughput fashion, aided by the advent of next-generation sequencing, has been pursued in recent years. One such method, known as ribosome profiling, involves the isolation and sequencing of ribosome protected fragments of mRNA [42]. A proxy for translation efficiency (often referred to directly as translation efficiency within ribosome profiling papers (e.g. [42]) can be calculated by dividing the number of sequencing depth and CDS length-normalized ribosome-protected fragments in a given gene’s CDS by the sequencing depth and transcript length-normalized number of mRNA-seq reads in the same gene. It is worth noting that this metric is unable to account for variation in ribosome elongation rate. For instance, a transcript with a slow average translation elongation rate relative to its translation initiation rate, and a relatively long half-life might accumulate many elongating ribosomes despite supporting a relatively low rate of protein production. Fortunately, existing data suggest that to be a relatively rare situation (e.g. Riba et al (2019) [72] figure 2D). Despite its limitations, it has been widely acclaimed for its utility in establishing translation efficiency and regulation thereof [11].

Another key limitation associated with ribosome profiling is the fact that it is limited to interrogating occupied CDS sequences directly. Consequently, it is unable to differentiate among alternative transcripts with different UTRs (except possibly in the muddying case where those UTRs contain upstream or downstream ORFs). Further, its ability to distinguish even among alternative coding sequences is hampered (though not eliminated - e.g. Weatheritt et al [96]) by the relatively short length of the fragment sequences (29 nt). Efforts to address these
issues and allow for transcriptomic interrogation of alternative isoform translation began with [82] and have centered around the sequencing of velocity gradient fractionated polysomes.

3.1.2 Previous work involving Frac-seq and similar techniques

Next-generation sequencing of polysome fractions without the RNase digestion step common to ribosome profiling protocols allows the full use of substantially longer read lengths, as well as the direct assessment of UTR sequences. Efforts toward the application of NGS to polysome fractions for the purpose of identifying alternative isoforms with distinct translation efficiencies began with Sterne-Weiler et al (2013) [82] who dubbed their technique Frac-seq. We note that while subsequent efforts using similar approaches introduced alternative monikers (e.g. TrIP-seq), for convenience in this manuscript we will refer to all approaches that apply NGS to polysome fractions as Frac-seq. Sterne-Weiler et al demonstrated the possibility of quantifying the relative abundance (i.e. percent spliced-in (PSI or Ψ)) of alternative isoforms in both the cytosol and (pooled) polysome fractions, and highlighted the large fraction of alternative events that exhibit differential relative abundance (i.e. ΔΨ) between the two fractions. Subsequently, several research groups pursued similar directions. Floor et al (2016) [30] took a highly similar approach, but radically increase the number of individual polysome fractions collected such that monosomes and polysomes of two through seven ribosomes are sequenced individually, along with a separate fraction for polysomes consisting of eight or more ribosomes. In so doing, they highlighted the importance of characterizing the distribution of mRNA abundances in different polysome fractions. Simply looking at pooled polysome-associated counts has the potential to hide a substantial amount of information. Wang et al (2016) [95]
pursued a similar multi-fraction design but applied CAGE-seq instead of standard RNA-seq in order to assess differential translation efficiency of isoforms differing in 5’-UTR sequence due to the existence and usage of alternative transcription start sites. Both Floor et al and Wang et al also performed a critically important comparison between translation efficiency values determined via Frac-seq and translation efficiency values determined via ribosome profiling, and to proteomics-derived protein abundances. Comparisons between Frac-seq and ribosome profiling, as well as Frac-seq and protein abundances, demonstrated non-trivial correlations. These comparisons are necessary in order to rule out the possibility that mRNA sedimentation is broadly caused by association with heavy non-polysome complexes, or that sedimentation is caused by mRNA association with stalled or otherwise non-productive ribosomes.

3.1.3 Considerations in the identification and interpretation of isoform-specific sedimentation

Previous work leveraging Frac-seq (or a similar variant) approached the question of identification and interpretation of differential sedimentation in a number of different ways. As described above, Sterne-Weiler et al (2013) [82] identify differentially sedimenting events as those exhibiting a $\Delta \Psi$ of a given non-zero magnitude between the polysome and total cytoplasm fractions. Wahba et al (2017) [93], who applied Frac-seq to cells pre- and post-radiation stress, pursued a similar approach, and further assessed the change in differential sedimentation in response to radiation using the $\Delta \Delta \Psi$. There are conceptual difficulties, however, with the interpretation of $\Psi$ values with respect to the relative translation efficiency of transcripts in alternative events. For instance, one might imagine an event exhibiting high $\Psi$ values in
fractions corresponding to low ribosome numbers and low $\Psi$ values in fractions corresponding to intermediate ribosome numbers, and again high $\Psi$ values in fractions corresponding to high ribosome numbers. This configuration defies ready interpretation regarding the relative translation efficiency of the two isoforms. To properly assess this, one must consider the relative amount of both isoforms that is associated with the various fractions. If, for instance the total amount of both isoforms in the low-ribosome fractions is small, the data may indicate higher relative translation efficiency of the included isoform (though the bimodality exhibited my the $\Psi$ values may indicate some interesting biology in this case). To get at the per-fraction RNA amount, one must not only have access to standard within- and cross-library normalized metrics of transcript abundance: one must also have means of estimating the relative amount of mRNA associated with each fraction. One method of inferring the latter information involves the addition of spike-in RNA to the entirety of each sucrose fraction prior to RNA isolation. Wang et al [95] performed this step, whereas Floor et al [30] did not, which may explain the superior agreement with ribosome profiling in the Wang et al measurements. Finally, it is important to account for the length of coding sequences (CDS), as was done by both Floor et al and Wang et al, but not by the aforementioned papers using $\Psi$ values and their differences as a primary means of quantification.

In this study, we analyze Frac-seq data derived from human H9 embryonic stem cells (H9 ESC) and human H9 neural precursor cells (H9 NPC). These data were also collected without the benefit of spike-in RNA, and thus a direct calculation of an Ingolia-style translation efficiency metric cannot be performed. However, we nonetheless see value in evaluating the relationship between transcript properties and event $\Psi$ values across the gradient. From such
information we can establish the nature and extent of the utility of Frac-seq datasets collected in this way. Our analysis highlights the fact that many events exhibit differential sedimentation across the gradient, and the fact that event $\Psi$ value trajectories are typically consistent with heavier polysome association of transcripts with longer CDS. Further, we find that NMD substrates are preferentially relatively enriched in lighter fractions, chiefly the 80S fraction, and that this enrichment is best explained by their putative NMD status rather than solely by their relative CDS length. Finally, we highlight the fact that while few events exhibit substantial changes in differential sedimentation between the two cell types, NMD isoforms do exhibit a general up-gradient shift in H9 NPCs, consistent with previous reports of NMD downregulation as a function of neurodifferentiation.

3.2 Materials and Methods

3.2.1 Mapping Illumina reads

RNA-seq data were pre-processed to remove adapter sequences, and reads mapping to repetitive sequences were removed by mapping to RepeatMasker sequences (Smit et al., RepeatMasker Open-4.0 at http://repeatmasker.org) using Bowtie2 [51]. The remaining reads were mapped to the human genome (hg19) using STAR [23], and PCR duplicates were removed collapsing fragments with common CIGAR strings, start and end positions, and pairs of unique molecular identifiers using custom scripts. The mapped reads were then lifted to GRCh38 also using STAR (version 2.7 [23]).
3.2.2 Identification of differentially expressed genes

Count tables of reads consistent with annotated genes were used as input for differential expression testing using DESeq2 [60]. Differential expression across the time dimension was established in each species by testing the inclusion of the time term. Genes with $|\log_2(\text{fold change})| \geq 1$ and FDR $\leq 0.05$ were required for a gene to be classed as differentially expressed.

3.2.3 Identification of putative CDS regions with/without PTCs in unannotated transcripts

Consensus coding sequence (CCDS) start codons were translated in silico. If an in-frame stop codon was encountered, it was checked for a) whether it occurred upstream of an exon-exon junction, and b) whether the distance from the final exon-exon junction was $\geq 55$ nt according to convention [64]. Consequently, within this approach, some transcripts may be annotated with more than one ORF, and a number of transcripts may also be annotated as putative nonstop decay substrates due to the absence of an in-frame stop codon. This process was performed using cdsInsertion (described above).

3.2.4 Classification of NMD events

NMD switch events were classified based on the following criteria: 1) all of the transcripts consistent with one (but not both) of the event isoforms are predicted to contain a PTC (stop codon $\geq 55$ nt upstream of the final exon-exon junction) when translated from any overlapping CCDS start codon; 2) assuming criterion 1 is satisfied, none of the transcripts from the
opposite form can be predicted to contain a PTC when translated from any overlapping CCDS start codon.

3.2.5 Identification of alternative splicing/processing events

Alternative splicing events were identified by pairwise comparison of all transcripts sharing at least one exon-intron junction. An alternative event was identified by the set of exons unique to either transcript bounded on both sides either by two exon-intron junctions common to both transcripts or by one exon-intron junction common to both transcripts and the transcript termini. The resulting events were then classified into the following types: skipped exon, multiple skipped exon, mutually exclusive exons, alternative donor, alternative acceptor, alternative first, alternative last, multiple alternative first, multiple alternative last, retained intron, multiple retained intron, complex alternative first, complex alternative last, complex. This approach is described in more detail in chapter 2.

3.2.6 Identification of alternative event differences across fractions/cell types

Within each cell type, the differential sedimentation status of alternative events was assessed by comparing Ψ values in all possible pairwise comparisons of fractions. To be considered for analysis, an event was required to at least a non-zero $|ΔΨ_{inner}|$ whose sign is the same regardless of which pair of Ψ values is used for the assessment (i.e. min(Ψ) or max(Ψ) as described above from any pair of replicates). For certain analyses, such as characterizing the overall number of (detected) differentially sedimenting events by event type, a threshold of $|ΔΨ_{inner}| ≥ 0.1$ was used. A 0.1 threshold has been used in a number of previous studies to
identify changes of appreciable magnitude. For other analyses, such as the assessment of the role of NMD status in differential sedimentation and the changing differential sedimentation of NMD events in different cell types, a threshold of $|\Delta \Psi_{\text{inner}}| > 0$ was used to improve statistical power. Although small differences may be of less clear biological significance at the level of an individual event, in aggregate they were considered to be important for assessing the significance of pan-event phenomena. Both thresholds were found to exhibit an empirical FDR of well below 0.05 in simulated data. For cross-cell type comparisons, the primary utilized metric was the $|\Delta \Delta \Psi_{\text{inner}}|$, such that differences in differential sedimentation, rather than base levels of inclusion, were reflected.

3.3 Results

3.3.1 A substantial fraction of alternative events exhibit differential sedimentation in both ESC and NPC

In order to assess differential sedimentation of alternative isoforms we performed Frac-seq on H9 ESCs and H9 NPCs. In line with more recent Frac-seq efforts, we increased the number of fractions sequenced separately: we sequenced a cytoplasmic sample (cyto), along with a monosome (mono) sample (i.e. a sample corresponding to mRNAs associating with a single ribosome or other complex of similar molecular weight), a light polysome (poll) sample (i.e. two-to-four ribosomes) and a heavy polysome (polh) sample (i.e. five or more ribosomes).

We defined differentially sedimenting events as those exhibiting a $|\Delta \Psi| \geq 0.1$ in at least one fraction-to-fraction comparison in a given cell type, and with a minimum of 20 sup-
porting junction reads in all samples of the two fractions. A total of (991, 7.8%) events meet these criteria in H9 ESC and (671, 4.9%) in H9 NPC. We compared the types of differentially sedimenting events to those non-differentially sedimenting events who could have met the criteria (i.e. $0.1 \geq \Psi \geq 0.9$ in at least one fraction and a minimum of 20 supporting junction reads in all samples of that fraction and at least one other fraction), and found the distribution of event types to be substantially different in both H9 ESC (P-value $\leq 2.2 \times 10^{-16}$, figure 3.1) and H9 NPC (P-value $\leq 2.2 \times 10^{-16}$, figure 3.2). The differentially sedimenting events are dominated by retained intron (RI) and skipped exon (SE) events in both cell types. For instance, RI events constitute 33% of the differential events in H9 ESC and 35% in H9 NPC, but only 19% and 17% of the non-differential events in those respective cell types. SE and alternative first exon (AF) events are similarly enriched in both cell types. Note that despite constituting the third most numerous category of differential events in both cell types, complex 5'-terminal (CF) events are actually depleted in relative abundance compared to non-differential events.
Figure 3.1: Count of non-differentially (left) and differentially (right) sedimenting events by event type in h9 ESC. In this case differential status is assigned when an inner $|\Delta \Psi| \geq 0.1$ is observed between any two fractions, and a minimum number of 20 junction count reads supports the measurement in all samples.
Figure 3.2: Count of non-differentially (left) and differentially (right) sedimenting events by event type in h9 NPC. In this case differential status is assigned when an inner $|\Delta \Psi| \geq 0.1$ is observed between any two fractions, and a minimum number of 20 junction count reads supports the measurement in all samples.
Figure 3.3: Alternative first event in ATP5MD.

Figure 3.4: Skipped exon event in PDZD11.
3.3.2 Fraction trend clusters and characteristics of differentially sedimenting events

In order to identify clusters of event behavior we first min-max normalized $\Psi$ values within fraction and cell type. Min-max normalization facilitates a focus on the relative enrichment of a particular isoform in a given fraction by eliminating the influence of global $\Psi$ magnitude on the clustering. We clustered the trends using CLARA (Kaufman et al [48]) and established the number of clusters using the gap statistic (Tibshirani et al [86]). This approach identified 6 clusters of events in H9 ESC (figure 4.45) and 8 in H9 NPC (figure 4.46) among events satisfying the aforementioned criteria for differential sedimentation. We plotted the mean trends of each cluster to illustrate overall behavior (figures 3.6 and 3.7), and we find that broadly they fall into two categories: monosome/light polysome-enriched (henceforth “light” clusters - clusters 1, 3, and 4 in H9 ESC, clusters 1, 3, 4, 5, and 7 in H9 NPC) or cytoplasm/heavy polysome enriched (henceforth “heavy” clusters - clusters 2, 5, and 6 in H9 ESC, clusters 2, 6, and 8 in H9 NPC).
Figure 3.6: Cluster trends of differentially sedimenting events in h9 ESC. As above differential status is assigned when an inner $|\Delta\Psi| \geq 0.1$ is observed between any two fractions, and a minimum number of 20 junction count reads supports the measurement in all samples. Clustering is performed with the Clara algorithm using the Manhattan distance.
Figure 3.7: Cluster trends of differentially sedimenting events in h9 NPC. As above differential status is assigned when an inner $|\Delta \Psi| \geq 0.1$ is observed between any two fractions, and a minimum number of 20 junction count reads supports the measurement in all samples. Clustering is performed with the CLARA algorithm using the Manhattan distance.

We further assessed certain properties of events found in each cluster. Of particular interest are the CDS and UTR lengths of the transcripts involved in the events, for their possible and putative relationships with sedimentation both independent of and dependent on translation.
efficiency respectively. We found that the CDS length ratios of events (i.e. \( \frac{\text{lengths(CDS_{\text{included}})}}{\text{lengths(CDS_{\text{excluded}})}} \)) - the mean is taken to handle cases where the event isoforms are consistent with CDS of multiple lengths) are typically low in light clusters and high in heavy clusters (figures 3.8 and 3.9 for H9 ESC and H9 NPC respectively - not that in the figures it is the ratio \( \log_{10} \) that is shown).

![H9 ESC event CDS length ratios by cluster](image)

Figure 3.8: CDS-length ratios of differentially sedimenting events by cluster in h9 ESC. CDS lengths for event isoforms with more than one consistent CDS were averaged.
Figure 3.9: CDS-length ratios of differentially sedimenting events by cluster in h9 NPC. CDS lengths for event isoforms with more than one consistent CDS were averaged.

Further, we find that the 3′-UTR length ratios are typically higher (i.e. longer included form 3′-UTRs) for light clusters and roughly equal (i.e. median 1) for the heavy clusters (figures 3.10 and 3.11 for H9 ESC and H9 NPC respectively).
Figure 3.10: 3′-UTR-length ratios of differentially sedimenting events by cluster in h9 ESC. 3′-UTR lengths for event isoforms with more than one consistent CDS were averaged.
Figure 3.11: 3′-UTR-length ratios of differentially sedimenting events by cluster in H9 NPC. 3′-UTR lengths for event isoforms with more than one consistent CDS were averaged.

In contrast, 5′-UTR length ratios on the whole roughly equal (i.e. median 1, figures 3.10 and 3.11 for H9 ESC and H9 NPC respectively). Some light clusters (4 and 3 in H9 ESC and H9 NPC respectively) do deviate slightly lower, and some high clusters (6 in H9 ESC, 2 and 4 in H9 NPC) deviate slightly higher than 1.
Figure 3.12: 5'-UTR-length ratios of differentially sedimenting events by cluster in h9 ESC. 5'-UTR lengths for event isoforms with more than one consistent CDS were averaged.
Another key property of isoforms that may impact their observed relative sedimentation properties is their susceptibility to translation-associated decay processes, such as nonsense-mediated decay (NMD). In order to assess the behavior of putative NMD substrates, we identified NMD switch events, in which one event isoform is consistent only with transcripts con-
aining a putative premature termination codon (PTC), and the other event isoform is consistent only with transcripts lacking a putative PTC. These NMD switch events therefore have a particular "NMD isoform" that is putatively an NMD substrate. We find that among these NMD switch events, events in which the included form is the NMD isoform are overwhelmingly associated with the light clusters, while events in which the excluded form is the NMD isoform are overwhelmingly associated with the heavy clusters (figures 3.14 and 3.15 for H9 ESC and H9 NPC respectively). One exception to this observation is heavy cluster 7 in H9 NPC, in which the excluded NMD isoform event counts still exceed the included NMD isoform event counts but only by a single event.
Figure 3.14: Count of putative NMD switch events by cluster in h9 ESC. NMD switch events are those in which one (and only one) isoform is consistent exclusively with putatively PTC-containing transcripts. This is referred to as the NMD isoform.
Figure 3.15: Count of putative NMD switch events by cluster in H9 NPC. NMD switch events are those in which one (and only one) isoform is consistent exclusively with putatively PTC-containing transcripts. This is referred to as the NMD isoform.

We highlight a handful of examples of NMD events. For instance, PTC-containing exons in NOVA1 and AUH exhibit preferential monosome enrichment (figures 3.16 and 3.17 respectively).
Retained intron events in DHRS1 and BDH2 exhibit the same monosome enrichment, and induce a putative PTC by frameshift and direct inclusion respectively (figures 3.18 and 3.19).
3.3.3 NMD event cluster association is not fully explained by CDS lengths

The biased association of included form NMD events with light clusters shown in figure 3.14 is consistent with predominant association of events with shorter included than excluded CDS lengths to these same clusters visible in figure 3.8. This consistency is expected given the similar bias in CDS length ratios inherent to most NMD events visible in figure 3.20. A bias towards the remaining clusters is observed in excluded form NMD events (figure 3.14),
which matches the similar bias of events with shorter excluded than included CDS lengths (figure 3.8).

![Figure 3.20: Included-excluded isoform CDS length ratios in all putative NMD switch events included in the utilized annotation.](image)

However, it is notable that these cluster associations are also consistent with the expectation that NMD isoforms will be degraded after the pioneer round of translation [44] and
thus be underrepresented in heavier fractions relative to their non-NMD counterparts. Consequently we endeavored to determine whether NMD event behavior could be distinguished in aggregate from that of non-NMD events with similar CDS lengths. In order to do this, we identified for each differentially sedimenting NMD event a non-NMD event with included and excluded form CDS lengths that are close to the NMD event’s CDS lengths in euclidean space. We sought matches for included and excluded form NMD events only among non-NMD events with included-excluded CDS length ratio $< 1$ or $> 1$ respectively. As this process still resulted in a fairly different distribution of CDS length ratios among the non-NMD events, we further filtered the matches such that the appropriate matched event isoform had a shorter CDS length than its NMD isoform counterpart, and such that the matched event CDS length ratio was either less than (in the case of included NMD events) or greater than (in the case of excluded form NMD events) that of the NMD event. In so doing, we aimed to err on the side of identifying matched non-NMD events whose CDS lengths might be expected to have an even more extreme impact on the event’s behavior than those of the NMD events.

Following the selection of matching events, a $\Delta \Delta \Psi_{\text{NMD - non-NMD}}$ was calculated for each comparison in h9 ESC (figures 3.21 and 3.23) and H9 NPC (figures). The number of comparisons with a particular $\Delta \Delta \Psi_{\text{NMD - non-NMD}}$ sign was counted for both h9 ESC (figures 3.22 and 3.24) and H9 NPC (figures). Further, the number of comparisons in which the $\Delta \Delta \Psi_{\text{NMD - non-NMD}}$ sign matched the expected sign under the hypothesis that an NMD isoform will be more restricted to the monosome and lower polysome fractions than match events with similar CDS lengths and CDS length ratios (figure 3.27 shows H9 ESC results and figure 3.28 shows H9 NPC results). For included form NMD events $\Delta \Delta \Psi_{\text{NMD - non-NMD}}$ sign expectations are positive.
for cytoplasm versus monosome and cytoplasm vs light polysome, and negative for the remainder. The expectations are reverse for excluded form NMD events. We find that the observed bias towards the expected sign to be significantly greater than even (P-value = $1 \times 10^{-7}$ for H9 ESC and $1.43 \times 10^{-5}$ for H9 NPC, exact binomial test with $p = 0.5$).
Figure 3.21: Density of Cyto vs Mono $\Delta\Delta\Psi$ between non-NMD proxies and NMD included form events given $|\Delta\Psi| > 0$ in at least one comparison. Further, the ratio of the means of the included and excluded form CDS lengths must be greater in each NMD event than in its matched non-NMD event, and the included form CDS length must be greater in each NMD event than in its matched non-NMD event.
Figure 3.22: Count of event pairs with given sign of Cyto vs Mono $\Delta\Delta\Psi$ between non-NMD proxies and NMD included form events given $|\Delta\Psi| > 0$ in at least one comparison. The CDS lengths are also restricted as above.
Figure 3.23: Density of Cyto vs Mono $\Delta \Delta \Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta \Psi| > 0$ in at least one comparison. Further, the ratio of the means of the included and excluded form CDS lengths must be shorter in each NMD event than in its matched non-NMD event, and the excluded form CDS length must be shorter in each NMD event than in its matched non-NMD event.
Figure 3.24: Count of event pairs with given sign of Cyto vs Mono $\Delta\Delta\Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta\Psi| > 0$ in at least one comparison. The CDS lengths are also restricted as above.
Figure 3.25: CDS length ratios of proxies and NMD included form events in h9 ESC (0 threshold). These are event pairs with given sign of Cyto vs Mono $\Delta\Delta\Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta\Psi| > 0$ in at least one comparison. Event pairs are further filtered such that the included-excluded form CDS length ratio is greater in the NMD event than in its non-NMD match, and such that the NMD event included form CDS length is greater in the NMD event than in its non-NMD match.
Figure 3.26: CDS length ratios of proxies and NMD excluded form events in h9 ESC (0 threshold). These are event pairs with given sign of Cyto vs Mono $\Delta\Delta\Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta\Psi| > 0$ in at least one comparison. Event pairs are further filtered such that the included-excluded form CDS length ratio is smaller in the NMD event than in its non-NMD match, and such that the NMD event excluded form CDS length is greater in the NMD event than in its non-NMD match.
A greater number of comparisons are consistent with NMD-dependent effect in H9 ESC with expected \(\Delta \Delta \Psi\) sign (0 threshold and additional CDS filtration).

Figure 3.27: Number of comparisons between proxies and NMD events in H9 ESC with expected \(\Delta \Delta \Psi\) sign (0 threshold and additional CDS filtration).
A greater number of comparisons are consistent with NMD-dependent effect in H9 NPC with expected $\Delta \Delta \Psi$ sign (0 threshold and additional CDS filtration).

Figure 3.28: Number of comparisons between proxies and NMD events in H9 NPC with expected $\Delta \Delta \Psi$ sign (0 threshold and additional CDS filtration).
3.3.4 Changes in differential sedimentation between H9 ESC and NPC

3.3.4.1 Assessment of differential gene expression and splicing occurring as a function of differentiation from ESC to NPC

Figure 3.29: Volcano plot showing differential expression of genes between ESC and NPC.

3.3.4.2 A small number of events exhibit cell type-specific differential sedimentation

In order to assess the possibility of cell type-specific regulation (for instance, cell type specific relative translation efficiency), we identified instances of cell type-specific differential sedimentation. In order to do so we identified events instances of a fraction-to-fraction $|\Delta \Delta \Psi|$ $\geq 0.1$ with a minimum number of 20 junction reads in any sample involved in the comparison.
Only a small handful of events (13) meet these criteria. Almost exclusively, these events exhibit differential sedimentation in one cell type, and not in the other. In the latter cell type, it is exclusively the case that the events exhibit near complete inclusion or exclusion in all fractions.

3.3.4.3 **Considered in aggregate, NMD events exhibit cell type-specific differential sedimentation consistent with an NPC-associated down-gradient shift**

Despite the lack of events exhibiting AS-independent large changes in isoform-specific sedimentation, we investigated the behavior of NMD switch events through differentiation (i.e. between the two cell types) both due to our observations on NMD switch sedimentation (above), and due to the reported decrease in NMD efficiency as a whole throughout neurodifferentiation ([58], [5], [45]). In observing the min-max normalized Ψ values of the same events in both H9 ESC and H9 NPC, it is relatively apparent that the NMD isoform seems to shift down the gradient to heavier fractions. In figures 3.30 and 3.31 this can be seen as relative increase or decrease respectively in the "h9npc_poll" column.
Figure 3.30: Included form NMD switch event min-max normalized $\Psi$ values in ESC and NPC. Given $|\Delta \Psi_{inner}| \geq 0.1$ and $\geq 20$ junction reads in at least one comparison.
Figure 3.31: Excluded form NMD switch event min-max normalized $\Psi$ values in ESC and NPC. Given $|\Delta\Psi_{\text{inner}}| \geq 0.1$ and $\geq 20$ junction reads in at least one comparison.

We counted the number of instances in which a given fraction exhibits an NMD switch event’s NMD isoform maximal inclusion (i.e. $\max \Psi$ in the case of included NMD isoforms and $\min \Psi$ in the case of excluded NMD isoforms). We find overwhelmingly NMD switch events exhibit their NMD isoform’s maximum relative inclusion in either the monosome or
In H9 NPC, however, a much greater proportion of events exhibit a light polysomal maximum than in H9 ESC (figure 3.32, $P = 2.11 \times 10^{-2}$, Chi-squared test).

Moreover, we find that among the NMD switch events exhibiting maximal NMD isoform inclusion in a different fraction in each cell type, the observed transition is overwhelmingly "down-gradient" from the monosome to light polysome fraction (figure 3.33).

Figure 3.32: Count of NMD events that exhibit highest NMD isoform relative inclusion in a given fraction in each cell type. A chi-squared test supports the distributions being different in the two cell types.
Figure 3.33: Count of fraction pairs in which NMD events exhibit highest NMD isoform relative inclusion in H9 ESC and H9 NPC. Up- and down-gradient refer to the direction the NMD isoform’s maximal relative inclusion appears to shift upon differentiation (i.e. towards lighter or heavier complexes respectively).
3.3.5 UPF3B and UPF1 knockdown induce modest down-polysome shift in PTC isoforms

In an effort to assess whether the observed NMD isoform relative down-gradient shifts in H9 NPC were indeed a result of attenuated NMD efficiency, we endeavored to observe the behavior of NMD switch events in cell lines depleted of NMD factors. We utilized Gen1C IPSC cells [61] with guides targeting either UPF3B or UPF1, and performed Frac-seq in knockdown or control (with or without doxycycline, respectively, to induce Cas9 expression) conditions. We repeated the analysis described above, and identified the fraction in which NMD switch events exhibit maximal NMD isoform inclusion before and after UPF3B or UPF1 knockdown. In both cases, we observe a higher number of events exhibiting maximal inclusion in the light polysome fraction (figures 3.34 and 3.35 for UPF3B and UPF1 knockdown respectively). We note that while only the UPF3B distributions are significantly different pre and post-knockdown (P = 3.84 × 10^{-2}), the UPF1 experiment exhibits the same trend.
Figure 3.34: Count of NMD events that exhibit highest NMD isoform relative inclusion in a given fraction in control (left) and UPF3B knockdown (right) cells. A chi-squared test supports the distributions being different in the two cell types.
Figure 3.35: Count of NMD events that exhibit highest NMD isoform relative inclusion in a given fraction in each cell type. A chi-squared test supports the distributions being different in the two cell types.

In examining the direction of maximal fraction transitions (figures 3.36 and 3.37), we observe that the majority of transitioning events move from a lighter to a heavier fraction in both the UPF3B and UPF1 knockdown experiments (figures 3.38 and 3.39), as was the case with ESC to NPC.
Figure 3.36: Count of fraction pairs in which NMD events exhibit highest NMD isoform relative inclusion in control and UPF3B knockdown cell lines. Up- and down-gradient refer to the direction the NMD isoform’s maximal relative inclusion appears to shift upon differentiation (i.e. towards lighter or heavier complexes respectively).
Figure 3.37: Count of fraction pairs in which NMD events exhibit highest NMD isoform relative inclusion in control and UPF1 knockdown cell lines. Up- and down-gradient refer to the direction the NMD isoform’s maximal relative inclusion appears to shift upon differentiation (i.e. towards lighter or heavier complexes respectively).
Figure 3.38: Count of NMD event transition direction in UPF3B knockdown and control cells, considering only events that underwent a transition. Up- and down-gradient refer to the direction the NMD isoform’s maximal relative inclusion appears to shift upon differentiation (i.e. towards lighter or heavier complexes respectively).
Figure 3.39: Count of NMD event transition direction in UPF1 knockdown and control cells, considering only events that underwent a transition. Up- and down-gradient refer to the direction the NMD isoform’s maximal relative inclusion appears to shift upon differentiation (i.e. towards lighter or heavier complexes respectively).
3.3.6 NMD isoforms decrease in relative abundance after neurodifferentiation and NMD factor knockdown

We also assessed the relative inclusion of NMD switch event NMD isoforms in the cytoplasmic fractions of all cell lines. While the gradient shifts described above are consistent with decreasing NMD efficiency, one would expect that a relative increase in isoform abundance in the cytoplasm as a whole should be at least somewhat apparent. To assess whether or not this is the case, we determined the cytoplasmic $\Delta \Psi$ values for NMD switch events, and then assessed whether the observed sign of the values was consistent with the hypothesis that NMD isoforms should increase in relative abundance upon NMD factor knockdown or differentiation from ESC to NPC. At our standard thresholds, only a handful of NMD switch events exhibited a change in each pair of conditions. In order to observe any trend in aggregate, we examined the number of events consistent with the above hypothesis given $|\Delta \Psi_{\text{inner}}| > 0$, and we find that in all pairs of conditions, the majority of events are consistent with the hypothesis (figure 3.40). We examined whether or not this was the case at a series of even less stringent thresholds and found it to remain true (figure 4.63).
3.4 Conclusions and Discussion

3.4.1 Interpretation of differential sedimentation

We observe H9 ESC and NPC differential sedimentation trends consisting primarily of either monosome/light polysome enrichment or cytoplasm/heavy polysome enrichment as mentioned above (figures 3.6 and 3.6). We further observe that the former category largely consist of events with shorter included than excluded mean ORF lengths, and that the latter consist of events with longer included than excluded mean ORF lengths (figures 3.8 and 3.9). Broadly, this observation is consistent with the fact that isoforms with longer ORFs will, given identical rates of translation initiation, elongation, and termination, host larger numbers of ribosomes during steady-state translation. However, another possibility is that those isoforms preferentially enriched in the monosome/light-polysome fraction (i.e. included isoforms in clusters with Ψ-value maxima in those fractions, excluded isoforms in clusters with Ψ-value minima in those

Figure 3.40: Number of NMD switch events exhibiting the expected ΔΨ sign.
fractions) are restricted to those fractions due to their surveillance by translation-dependent decay processes such as NMD. Indeed, it is certainly the case that we observe NMD isoform relative enrichment in light fractions (figures 3.14 and 3.15 for H9 ESC and H9 NPC respectively). However, even when such events are removed, a modest correlation between CDS length ratio and cluster remains (figures 4.47 and 4.48). We discussed earlier the difficulty in assessing the implications of Frac-seq results for Ingolia translation efficiency in the absence of fraction-specific estimates of absolute mRNA abundance, but this correlation of CDS length ratio with cluster trend provides some suggestion that a naive interpretation may have some validity at least in aggregate. Events whose isoforms are translated with equal efficiency are expected to differentially sediment in accordance with their relative CDS lengths: the isoform with the shorter CDS should associate proportionately more often with lighter fractions. In turn, a reasonable hypothesis for the differential sedimentation of events whose isoforms bear equal CDS lengths is differential Ingolia translation efficiency. In the absence of substantial differences in elongation rate one can reasonably hypothesize that isoforms with greater relative abundance in heavier fractions are translated more efficiently than their alternatives. However, we emphasize that an experimental design that enables the determination of absolute mRNA abundance in each fraction, such as one in which spike-in RNA is added to the sucrose post-fractionation, is strongly preferred. We do note that even with the above design examination of Ψ values fraction-by-fraction may illuminate instances of bimodal translation efficiency that would be masked by the exclusive use of Ingolia translation efficiency to summarize results.
3.4.2 NMD status contributes to NMD switch event sedimentation

A clear feature of differentially sedimenting events is the strong preference exhibited by NMD switch events for light-fraction enrichment of the NMD isoform (figures 3.14, 3.15, 3.30, and 3.31). This pattern has been noted by others (e.g. [30]), and is consistent with the aforementioned CDS ratio-associated sedimentation behavior. In addition to the fact that the NMD isoform of an event overwhelmingly has a shorter CDS (figure 3.20), the ribosome-triggered decay process itself is likely a strong factor in its restriction to lighter fractions.

In order to assess whether CDS lengths alone were likely to explain the sedimentation of NMD events we sampled non-NMD events with a CDS-ratio at least as extreme as a given NMD event. We found that in the NMD-non-NMD event pairs, the light fraction enrichment of the NMD isoform was more extreme than that of the corresponding isoform in the non-NMD event (figures 3.27, 3.28), suggesting that the NMD status of the events is a substantial contributor to the light fraction-enriched sedimentation of their NMD isoforms. Moreover, this suggests that Frac-seq may be a useful technique to identify and study putative NMD targets under minimal cellular perturbation.

3.4.3 NMD switch event sedimentation in H9 NPC is consistent with stabilization of NMD targets

Finally, we examined changes in differential sedimentation between H9 ESC and H9 NPC, and identified a trend in which NMD isoforms appear to modestly shift down-gradient to heavier fractions after differentiation (figures 3.30, 3.31, ). This behavior is potentially consistent with previous reports that NMD becomes less efficient (but still active) during and after
neurodifferentiation [59], [38]. Moreover, we also observe this phenomenon after knockdown of NMD factors UPF1 and UPF3B (albeit less consistently). If the decay process is fundamental in restricting NMD targets to light fractions (i.e. seldom being translated by more than a handful of ribosomes), then a decrease in decay rate should allow initiation by a greater number of ribosomes prior to decay. We note, however, that many events exhibiting this shift do not exhibit any clear increase in cytoplasmic abundance (e.g. the above-highlighted NMD event examples: figures 3.16, 3.17, 3.19, and 3.18). It is possible that some events are subject to alternative mechanisms of NMD isoform reduction. For instance, a reduction in inclusion of the NMD isoform at the level of splicing, or possibly increased nuclear retention [10]. However, we do note the fact that the majority of NMD events that exhibit a change in cytoplasmic Ψ after differentiation to NPC or knockdown of an NMD factor exhibit an increase in the relative abundance of the NMD isoform (figures 3.40 and 4.63). While further work (e.g. RNA-seq of nuclear fractions) is needed to fully deconvolute these issues, taken together, these elements suggest the possibility that Frac-seq may provide a novel method for the widespread assessment of NMD efficiency across experimental conditions.
Chapter 4

Conservation of alternative splicing dynamics
in primate in vitro neurodifferentiation

4.1 Introduction

Understanding neuronal cell fate specification and brain development more broadly is critical to a host of research questions in biomedicine and molecular biology (as highlighted by e.g. Tsai et al [91]). It is well established that precise changes in gene expression are essential to healthy brain development (and development more broadly, with a number of demonstrated species-specific variations (Ebisya et al [25]), but a complete temporal map of the molecular regulatory events and their consequences remains elusive. Such a roadmap would provide useful insight into the molecular basis of disease-related perturbations in early development, as well as clues as to the specific purposes and molecular consequences of those events, and directions to pursue in the evolving arena of regenerative medicine. The creation of strategies for inducing
in vitro neurodifferentiation, as well as in vitro cortical neurosphere development has enabled the use of RNA-seq to characterize the shifting neurodevelopmental transcriptomic landscape in early developmental stages not otherwise readily available for study due to a host of ethical and practical considerations (e.g. Field et al [29]; Fiddes et al [28]). A number of such experiments have highlighted dynamic transcriptional regulation of thousands of genes accompanying the vast morphological changes undergone by maturing neurons. Beyond transcriptional regulation, dynamic changes in signaling pathways, and epigenetic changes, post-transcriptional regulation has been established to be crucial to many aspects of brain development.

Alternative splicing (AS) is a phenomenon wherein competing splice sites may allow a single gene to produce multiple mature transcripts, and is a substantial contributor to transcriptome/proteome complexity in addition to being a major correlate of tissue identity (e.g. Yeo et al [101], Wang et al [94]). It is now well established that the majority (> 90%) of multi-exon human genes are subject to alternative splicing in one context or another. Moreover, neuronal and brain tissue generally are thought to exhibit the most complex pattern of alternative splicing (e.g. [94]), which is achieved through putatively precisely timed (careful here) developmental regulation mediated by key RNA binding proteins such as NOVA1 and RBFOX1 Weyn-Vanhentenryck et al [97], and AS defects have been established to cause a variety of deleterious neurodevelopmental phenotypes (e.g. Su et al [85]).

Despite extensive study in mice [e.g. Weyn-Vanhentenryck et al [97], Zhang et al [104]], relatively little is known about the extent, regulation, conservation, and consequences of neurodevelopmentally regulated alternative splicing in humans. Recent studies have shed
some light on the issue of conservation (e.g. Mazin et al [62], which focuses on peri/postnatal development), but the earliest developmental stages remain poorly characterized at the level of post-transcriptional control. Field et al [29] and Fiddes et al [28] have recently published data that may allow the bridging of this gap: RNA-seq data from in vitro cortical neurosphere differentiation from human and rhesus macaque embryonic stem cells (ESCs) as well as from chimpanzee and orangutan induced pluripotent stem cells (IPSCs). Moreover, Field et al [29] have put forth the notion that conservation of temporal gene expression dynamics across multiple species may suggest that precise timing regulation is important, and moreover sequence conservation itself is often considered suggestive of a genes functional importance (a relationship with a number of important caveats and considerations [19]). As a consequence, one might expect the sequence and dynamical conservation of alternative splicing events to be similarly indicative of heightened relevance, and so the establishment of those values should provide parameters for estimating not only the relative developmental distinguishment of particular splicing events but also clues as to the broader impacts of the developmental AS program as a whole.

While much work has been done to establish the importance of specific RBPs to AS regulation (both neurodevelopmental and otherwise), it is clear that the specific causality in many if not most of the cases remains unknown. Recent work by Xu et al [100] highlights the importance of interactions between RBPs and histone modifications, along with signaling pathways and cell cycle machinery. The extent to which these components contribute in a quantitative sense to any given AS event throughout neurodifferentiation is essentially totally unclear. Furthermore, most models entirely discount the possible impact of simple mass action, wherein the radical alteration of transcriptional state/distribution of a cell could have radical consequences
for splicing decisions. We propose that an examination of relative dynamical conservation, in concert with awareness of species-specific sequence variation, could shed considerable light on these regulatory unknowns. Finally, the functional consequences of most alternative splicing events remain completely unknown. A number of recent works (e.g. Tress et al [89]) have cast doubt the assertion AS variants result in extensive expansion of the proteome and (to the extent that they do) that those variants have substantively distinctive roles (e.g. [26]).

Here, we seek to expand our limited understanding of these facets of alternative splicing in the specific context of human in vitro cortical neurosphere development by analyzing RNA-seq timeseries data (generated by Field et al [29] and Fiddes et al [28]) derived from human, chimpanzee, orangutan, and rhesus macaque stem cells. In so doing, we uncover an extensive neurodevelopmental alternative splicing program that exhibits both sequence and dynamical conservation. The dynamic alternative splicing events span a variety of gene types and ways in which they alter transcript structure, and putatively affect a variety of protein features. We identify properties of events that are most likely to be conserved and find that those resulting in altered coding sequence and those whose host genes have a larger number of protein-protein interactions are more conserved. Moreover, we present evidence (based on data from Blair et al, Draper et al) that many of these isoforms are indeed translated, which along with their higher relevance in PPI and altered coding sequences suggests a role in rewiring the developmental proteome. We also demonstrate that the previously reported phenomenon of stabilization of nonsense-mediated decay substrates appears conserved among the tested species, and propose that it occurs simultaneous to (proposed compensatory mechanism). Further, we describe some species-specific developmentally regulated AS-events, and speculate as to their functional
consequences.

4.2 Materials and Methods

4.2.1 Mapping Illumina reads

RNA-seq data from Field et al [29], [8], and [99] were downloaded as gzipped fastq files from SRA, and were mapped to the appropriate genome (GRCh38, panTro4, ponAbe2, and rheMac8 for human, chimpanzee, orangutan, and macaque respectively) using STAR (version 2.7 [23]). PCR duplicates were removed (also using STAR) by collapsing fragments with identical start/end mapping positions.

4.2.2 Choice, pre-processing, and expansion of transcriptome annotations

GENCODE (v27, [39]) was used as a basis for CAT [27] to generate preliminary transcriptome 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 aligned RNA-seq reads for StringTie [68] to assemble unannotated transcripts. A final transcriptome for downstream analysis was generated for each species using the StringTie merge command. We refer to these transcriptome annotations as StringTie-augmented CAT-GENCODE annotations. A separate set of CAT-GENCODE-derived annotations including only transcripts included in the annotations of all four species was generated and used for cross-species differential expression analysis. We refer to these annotations below as common-transcript CAT-GENCODE annotations.
4.2.3 Identification of differentially expressed genes

The expression of transcripts in the StringTie-augmented CAT-GENCODE annotations of individual species was quantified using kallisto (version 0.45.0) [12] with 50 bootstrap replicates. Differential expression of genes across time was then assessed using sleuth (version 0.30.0) [69] in gene mode (i.e. with `gene_mode=TRUE` in `sleuth_prep`), such that count aggregation rather than p-value aggregation of the transcripts was performed. The latter decision was taken as alternative event dynamics (which may be indirectly identified by the p-value aggregation approach) were to be assessed separately. For each gene in each species the maximum absolute value of all pairwise timepoint-to-timepoint log2 fold-change values was determined.

4.2.4 Identification of putative CDS regions with/without PTCs in unannotated transcripts

Consensus coding sequence (CCDS) start codons were translated in silico. If an in-frame stop codon was encountered, it was checked for a) whether it occurred upstream of an exon-exon junction, and b) whether the distance from the final exon-exon junction was $\geq 55$ nt according to convention set by [64]. For a CCDS start codon-transcript combination to be tested, it was only required that the start codon overlap the transcript exon(s). Consequently, within this approach, some transcripts may be annotated with more than one ORF, and a number of transcripts may also be annotated as putative nonstop decay substrates due to the absence of an in-frame stop codon.
4.2.5 Identification of alternative splicing/processing events

Alternative splicing events were identified by pairwise comparison of all transcripts sharing at least one exon-intron junction. An alternative event was identified by the set of exons unique to either transcript bounded on both sides either by two exon-intron junctions common to both transcripts or by one exon-intron junction common to both transcripts and the transcript termini. The resulting events were then classified into the following types: skipped exon, multiple skipped exon, mutually exclusive exons, alternative donor, alternative acceptor, alternative first, alternative last, multiple alternative first, multiple alternative last, retained intron, multiple retained intron, complex alternative first, complex alternative last, complex. This approach is described in more detail in chapter 2.

4.2.6 Identification of orthologous alternative splicing/processing events

The whole genome sequences of human (GRCh38), chimpanzee (PanTro4), orangutan (ponAbe2), and macaque (rheMac8) were mapped to one another using minimap2 (version 2.11-r797) [54] with parameters --cs and -asm20. The resulting mappings were used to lift the coordinates of alternative event exons to other species with a slightly 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 re-assembled 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 then proposed as putative orthologs, and then checked against events natively identified in the target species to identify putative orthologous relationships. Non-one-to-one relationships
were removed from consideration. The code implementing this approach is available online at
https://github.com/ajw2329/multi_species_as_events.

4.2.7 Quantification of alternative splicing/processing events

Alternative events were quantified by counting exon-exon and exon-intron junction
fragments. Fragments were considered to be consistent with an exon-intron junction if the co-
ordinates of a contiguously mapped block overlapped the coordinates of the junction. Both a
python implementation [84] of Nested Containment Lists [3] and pysam (https://github.com/pysam-
developers/pysam), a python wrapper for SAMtools [55], were used for this determination.
Fragments were considered to be consistent with an exon-exon junction if the aligner (in this
case STAR) determined that they contained one with exon boundaries precisely matching the
genomic coordinates of exon boundaries in the event annotations derived from the StringTie-
augmented CAT-GENCODE transcript annotations. Both reads in a fragment were considered
together to eliminate the possibility of double-counting a junction in cases when read pairs over-
lap one another. Reads were uniquely assigned to an isoform whenever possible (i.e. in cases
where a junction, or combination of junctions, contained in a fragment was unique to a given
event isoform). Ψ values were calculated as the ratio of included form counts to total counts.
For each event, Ψ values were calculated for all pairs of included and excluded form junctions,
(resulting in, for instance, two possible Ψ values for a typical SE event).
4.2.8 Identification of alternative splicing differences across conditions/species

Within each species, dynamic alternative events (i.e. events whose inclusion changes as a function of time/differentiation status) were assessed by comparing $\Psi$ values in all possible pairwise comparisons of timepoints. To be considered for analysis, an event was required to at least a non-zero $|\Delta \Psi_{inner}|$ whose sign is the same regardless of which pair of $\Psi$ values is used for the assessment (i.e. $\min(\Psi)$ or $\max(\Psi)$ as described above from any pair of replicates).

For certain analyses, such as characterizing the overall number of (detected) changes by event type, a threshold of $|\Delta \Psi_{inner}| \geq 0.1$ was used. A 0.1 threshold has been used in a number of previous studies to identify changes of appreciable magnitude. For other analyses, such as the determination of the relationships between cross-species differences in dynamics and CDS alteration or putative PPI degree, a threshold of $|\Delta \Psi_{inner}| > 0$ was used to improve statistical power. Although small differences may be of less clear biological significance at the level of an individual event, in aggregate they were considered to be important for assessing the significance of pan-event phenomena. Both thresholds were found to exhibit an empirical FDR of well below 0.05 in simulated data. For cross-species comparisons, the primary utilized metric was the $|\Delta \Delta \Psi_{inner}|$, such that differences in dynamics, rather than base levels of inclusion, were reflected. Events were examined on the basis of their exhibition or lack of a non-zero $|\Delta \Delta \Psi_{inner}|$, and also on the magnitude thereof.

4.2.9 Unsupervised clustering of gene expression and $\Psi$ value time series

To cluster gene expression and alternative event $\Psi$ value time series, gene expression values (i.e. the sum of normalized reads-per-base for each transcript in a gene) or $\Psi$ values
of dynamic genes (i.e. genes with $q \leq 0.05$ and maximum $|\log_2(fold - change)| \geq 2$) and events (i.e. events with maximum $\Delta \Psi_{inner} \geq 0.1$) were individually subjected to min-max normalization. Subsequently, genes and events within each species were separately clustered using CLARA [48] with Manhattan distance with a series of possible center numbers (2 through 20). To obtain an unbiased estimate of the optimal number of clusters, the gap statistic [86] was utilized.

4.2.10 Gene ontology analysis of clustered genes and events

Following unsupervised clustering, set enrichment analysis was performed on the genes contained in each cluster with topGO [4], against Gene Ontology ([16], [6]) categories in biological processes. For enrichment analysis of alternative events, only unique genes within each cluster were considered. The default weight01 method was used and categories with $P \leq 0.05$ were considered enriched. P-values were unadjusted as the topGO manual recommends.

4.2.11 Classification of NMD events

NMD switch events were classified based on the following criteria: 1) all of the transcripts consistent with one (but not both) of the event isoforms are predicted to contain a PTC (stop codon $\geq 55$ nt upstream of the final exon-exon junction) when translated from any overlapping CCDS start codon; 2) assuming criterion 1 is satisfied, none of the transcripts from the opposite form can be predicted to contain a PTC when translated from any overlapping CCDS start codon.
4.3 Results

4.3.1 Primate neurodifferentiation is accompanied by extensive AS changes

In order to assess the primate neurodifferentiation-associated landscape of alternative pre-mRNA processing, we leveraged published RNA-seq data from human and rhesus macaque ESCs as well as chimpanzee and orangutan IPSCs (Field et al 2017 [29]). Field et al induced differentiation of the stem cells to cortical neurospheres. This approach has been used in a number of recent efforts as a model for prenatal brain development [15]. We then applied a bioinformatics pipeline designed for the quantification and interpretation of pairwise alternative events, including the tools junctionCounts and cdsInsertion described above. This pipeline further makes use of StringTie (Pertea et al [68]) to augment primate transcriptomes defined by mapping human-annotated events to primate genomes using Comparative Annotation Toolkit (CAT) (Fiddes et al [27]). A schematic illustrates the process in figure 4.1.
4.3.1.1 Thousands of genes exhibit differentiation-associated dynamics in each species

We first assessed both gene expression and and alternative processing changes occurring in individual species. We find tens-of-thousands of genes whose expression changes as a function of time in each species using kallisto (Bray et al [12]) for transcript quantification and sleuth (in gene mode, Pimentel et al [69]) for assessment of dynamic behavior. We assessed differential gene expression at a number of minimum fold-change and p-value thresholds and found thousands of genes to be classified as differential (i.e. changing as a function of time).
even with fairly stringent criteria (i.e. FDR ≥ 0.05, max |log₂(fold−change)| ≥ 2).

Using unsupervised clustering, we identify four broad trends in gene expression direction changes in human differentiation (figure 4.2): increasing slowly (cluster 2 in figure 4.2), increasing moderately (cluster 2 in figure 4.2), increasing quickly with a terminal decrease (cluster 1 in figure 4.2), and decreasing quickly (cluster 3 in figure 4.2). Broadly, these general trends occur in chimpanzee (figures 4.64 and 4.65), orangutan (figures 4.66 and 4.67), and macaque (figures 4.68 and 4.69). In order to further evaluate the functional implications of these clusters, we performed gene set enrichment analysis using topGO (Alexa et al [4]) and found that clusters with generally increasing trends were associated with neurodevelopmental ontologies while the single decreasing cluster was associated with a variety of non-neurodevelopmental categories (right panel of figure 4.2), notably positive regulation of cell proliferation.
Figure 4.2: Gene expression trends and enriched regulatory factors in human.
In order to infer regulatory networks driving the behavior of genes in specific clusters, we applied eXpression2Kinases Web (Clarke et al [18]), which attempts to infer a set of transcription factors underlying the differential expression of a list of genes, and in turn a set of kinases that target those transcription factors and other proteins with which they putatively interact. This approach identifies among others, REST, SUZ12, EZH2, AR, and SMAD4 consistently among those transcription factors enriched in up-regulated genes, and SOX2, NANOG, KLF4, and POU5F1 among those transcription factors enriched in down-regulated genes (figure 4.2, left panel). A number of kinases are consistently enriched in all clusters including GSK3B, MAPK14, MAPK1, and CDK2.

4.3.1.2 Hundreds of alternative events undergo differentiation-associated dynamics in all species

To assess differentiation-dependent alternative splicing, we applied junctionCounts (v0.1.0) to identify and quantify alternative events from the StringTie-augmented CAT transcriptomes described above. We identified events with differentiation-associated dynamics as those exhibiting an inner $|\Delta \Psi| \geq 0.1$ in at least one timepoint-to-timepoint comparison and a minimum of 15 junction reads in any sample within that comparison. This approach identifies hundreds of dynamic events in each species (figure 4.3). The distribution of dynamic event types is dominated by SE, AF, and RI events, and is largely similar to the background distribution of event type frequencies that could have exhibited such a change (i.e. $0.1 \leq \Psi_{\text{min}} \leq 0.9$ and $0.1 \leq \Psi_{\text{max}} \leq 0.9$ in all samples of at least one timepoint, and a minimum of 20 junction reads in all of those samples as well as all samples of at least one other timepoint).
We further applied unsupervised clustering to the min-max normalized $\Psi$ trends of the dynamic events to assess the overall dynamic trends. We find that the trends are broadly increasing or decreasing in each species with the suddenness of the change constituting the primary differences among them (left-most and middle panels of figure 4.4 for human trends,
Because changes in the relative abundance of a given form have no broadly interpretable meaning, we grouped the clusters further into early and later categories with respect to the timing of the dynamics in which each contains both a trend indicating the increase of the included isoform and a trend indicating the decrease of the included isoform (and therefore the increase of the excluded isoform).

Figure 4.4: Trends of differentiation-associated alternative event dynamics in humans. Trends identified by clustering min-max normalized $\Psi$ values. $|\Delta \Psi| \geq 0.1$ and minimum 20 junction reads in at least one timepoint-to-timepoint comparison comparison.

Because of the primarily increasing or decreasing behavior found in the cluster trends, we looked to establish the extent to which monotonic dynamics dominate individual event behavior. We first examined the number of dynamic events in each species which met or failed the
condition that all timepoint-to-timepoint comparisons with a nonzero $|\Delta \Psi|_{inner}$ have the same sign $|\Delta \Psi|_{inner}$. We established that the majority of events in all species meet these criteria at a number of different thresholds used to define dynamic events (figure 4.5). Moreover, the percentage of events that meet these criteria is greater at higher minimum $|\Delta \Psi|_{inner}$ thresholds.

![Monotonicity bar plot](image)

Figure 4.5: Monotonicity bar plot.
To examine the extent of non-monotonicity exhibited by events failing to meet the above criteria, we identified the maximum $|\Delta \Psi|_{inner}$ of each sign, and calculated the ratio of the smaller and larger absolute values of the two numbers. We plot the distributions of these values at a series of different $|\Delta \Psi|_{inner}$ thresholds, and we find that as the thresholds increase, the extent of counter-trend changes decreases (figure 4.6). This is to say that even among non-monotonic events (i.e. events which exhibit both positive and negative $\Delta \Psi_{inner}$ changes as a function of differentiation), the majority of those counter-trend changes are small relative to the on-trend changes, and don’t typically subtract substantially from the overall tonicity of the trend.
We applied gene set enrichment (GO biological processes) using topGO (Alexa et al [4]) to the genes hosting events found in these categories. To do this, we only considered unique genes hosted in each cluster, which is to say that genes in which multiple events change in the same way are not weighted higher than those in which single events exhibit dynamics.
We a mixture of ontologies enriched in both sets of clusters involving regulation of splicing and cytoskeletal reorganization, as well as signaling pathways (e.g. MAPK cascade enriched in the later clusters) (right-most panel of figure 4.4). We find far fewer ontologies unambiguously associated with neurodifferentiation than were enriched among differentially expressed genes. Broadly, we find similar types of categories enriched in other species (right-most panels of figures 4.70, 4.71, 4.72 for other chimpanzee, orangutan, and macaque respectively).

We further examined the extent of overlap of differentially expressed genes and alternative spliced genes in each species at a number of different thresholds. We find that while many genes are both differentially expressed and exhibit alternative splicing dynamics through differentiation, a large number exhibit alternative splicing dynamics independent of obvious differential expression. Over half the genes exhibiting alternative event dynamics at any given threshold are not classed as differentially expressed (in other words rejected as non-differentially expressed on the basis of $FDR \geq 0.05$, $\max |\log_2(fold - change)| \geq 1$) as shown in figure 4.7.
Figure 4.7: Differential expression status of dynamic alternative events. All events have at least 20 junction reads in the appropriate samples and are subject to the $\Delta \Psi$ thresholds indicated in the facet labels. Differentially expressed genes (x-axis label "True") exhibit FDR $\geq 0.05$, max $|\log_2(\text{fold-change})| \geq 1$.

Moreover, the distribution of maximum log fold-changes is not different across dynamic event magnitude thresholds, including in genes not thresholded by alternative event dynamics (figure 4.8, "-Inf" row).
Figure 4.8: Maximum absolute log-fold change distribution of genes with dynamic events. All events have at least 20 junction reads in the appropriate samples and are subject to the $\Delta \Psi$ thresholds indicated in the facet labels.

We also looked to see whether any of the putative regulators identified by expression2Kinases were alternatively spliced. Indeed we find that in human, 16 of the 262 (6.1%) identified regulators (i.e. transcription factors and kinases enriched with FDR $\geq 0.05$) and their putative intermediary nodes contain an alternative event with $|\Delta \Psi_{inner}| \geq 0.1$ in at least one timepoint comparison ($P = 0.0001452$, one-sided Fisher’s Exact Test for enrichment). In figure 4.2, regulators and nodes with alternative events satisfying the above threshold are highlighted.
in yellow). For instance, one of the enriched kinases, MAPK8, exhibits high-magnitude dynamics in the inclusion of mutually exclusive exons (figure 4.9) while also increasing modestly in overall expression \( (q = 2.23 \times 10^{-4}, \max log_2(fold-change) = 1.42). \)

**Figure 4.9: Dynamics of MX event in MAPK8.**

XPO1, which is a putative interacting partner of a number of the enriched kinases and transcription factors (e.g. MAPK14, E2F4) similarly undergoes a large increase in inclusion of a skipped exon (figure 4.10), which is predicted to result in a downstream PTC due to a frameshift. Expression of XPO1 also modestly decreases over differentiation \( (q = 1.11 \times 10^{-4}, \text{ESC-to-W5 } log_2(fold-change) = -0.93). \) It is also noteworthy that in our reanalysis of Blair et al [9], the behavior of the XPO1 skipped exon is similar, and moreover appears in cytoplasmic and polysomal fractions increasingly through neurodifferentiation.
Figure 4.10: XPO1 skipped exon event sashimi plots and Ψ values.

We note also that the temporal behavior of the highlighted XPO1 event, and indeed all of the alternative events in the aforementioned putative regulatory genes is largely similar among orthologs in all species (figure 4.11).
Figure 4.11: Min-max normalized $\Psi$ values of dynamic events in putatively regulatory kinases, transcription factors and intermediaries. Events shown exhibit $|\Delta \Psi_{\text{inner}}| \geq 0.1$ in at least one human timepoint comparison.

At a less stringent threshold of $|\Delta \Psi_{\text{inner}}| > 0$ the overlap is event more substantial, with 77 of the 262 expression2Kinases genes exhibiting alternative event dynamics (29.4%, $P < 2.2 \times 10^{-16}$ by one-sided Fisher’s exact test for enrichment).
4.3.2 Dynamic event alternative regions occur primarily in coding regions of protein coding genes

Having established the large number of alternative events exhibiting dynamics throughout differentiation, we sought to characterize the nature and putative functional consequences of those dynamics. We first examined the extent to which dynamic events occurred in different types of genes. Overwhelmingly, we find that dynamic events occur in protein coding genes (as identified by GENCODE biotype), but that a handful do occur in putatively non-coding genes (figure 4.12).

![Differential AS events in putative coding vs non-coding transcripts](image)

Figure 4.12: Counts of dynamic alternative events in coding and non-coding genes in all species (\(|\Delta \Psi| \geq 0.1\) and minimum 15 junction reads in the comparison).

Among the dynamic events in coding genes, we sought to establish the region of the transcripts undergoing alteration. We find that in the majority of cases, the alternative region of the event (i.e. the region not shared by both isoforms), overlaps the CDS (figure 4.13).
Having identified dynamic alternative events within each species, we sought to evaluate the extent to which orthologous events in different species behave similarly. In order
to perform this analysis, we first identified putative orthologous relationships of events in all species. To accomplish this, we generated pairwise whole genome alignments between each pair of species using minimap2 (Li et al [54]), and then used the minimap2 companion tool paftools liftover to lift the coordinates of alternative event exons from one species to another. We then re-assembled the events using the new exon coordinates, and established whether the exon-exon and exon-intron junctions matched any events already identified in the target species. We considered events with reciprocal relationships (i.e. event E in species A matches event F in species B after liftover from species A to B, and event F matches event E after liftover from species B to A) to be putative orthologs. While it is not possible to know the ground truth of orthologous relationships, we attempted to evaluate the success of this approach by examining the extent to which it agreed with event relationships implied by full-length transcript relationships of exon count-concordant transcripts in the human and macaque CAT-aligned GENCODE transcript annotations. Among events for which a relationship was proposed by both methods, 93% of the relationships were concordant. As the liftover method allows for identification of orthologous relationships amongst unannotated events we proceeded with its application to the full StringTie-augmented transcriptomes of all species. Of the 518842 events identified across all 4 species, putative orthologs in each species were identified for 310000 (59.7%). Among the 5611 events exhibiting $\Delta\Psi_{inner} \geq 0.1$ and $\geq 20$ junction reads in at least one timepoint comparison, putative orthologs were identified in all species for 4440 (79.1%).

We then examined the behavior of putatively orthologous events in all species. Broadly, it is the case that dynamic events exhibit similar behavior among all orthologs, and this relationship is particularly true among events exhibiting higher magnitude changes in inclusion (e.g.
In order to quantify conservation of alternative event dynamics, we determined the correlation of $\Delta\Psi_{mid}$ values in all timepoint-to-timepoint species comparisons. Among events with $|\Delta \Psi_{inner}| > 0$ for the timepoint comparison in at least one of the two species, and with $\geq 20$
junction reads associated to either event isoform all involved samples, positive correlation was observed (e.g. R = 0.69 for human-macaque shown in figure 4.15, and R = 0.66 for chimpanzee-orangutan shown in figure 4.16).

Figure 4.15: Orthologs generally exhibit the similar $\Delta \Psi$ values in the same timepoint comparison in human and macaque cells.
Figure 4.16: Orthologs generally exhibit the similar $\Delta \Psi$ values in the same timepoint comparison in chimpanzee and orangutan cells.

We further assessed the extent to which the $\Delta \Psi_{mid}$ sign is the same in analogous timepoint comparisons across orthologous events. We find that the sign is often the same in analogous comparisons across orthologous events in all species-to-species comparisons (e.g. figure 4.17 for human-macaque and figure 4.18 for chimpanzee-orangutan), and the extent to which that is true increases substantially with increasing $|\Delta \Psi|_{inner}$ threshold.
Figure 4.17: Orthologs generally exhibit the same sign $\Delta \Psi$ in the same timepoint comparison in human and macaque cells.
Figure 4.18: Orthologs generally exhibit the same sign $\Delta \Psi$ in the same timepoint comparison in orangutan and chimpanzee cells.

The extent to which the $\Delta \Psi$ signs agree increases as a function of $\Delta \Psi$ magnitude as well as with the temporal distance of the comparison. The greatest agreement in ortholog behavior is exhibited by the stem cell-to-week 5 comparison both in terms of $\Delta \Psi_{\text{mid}}$ correlation ($R = 0.81$ in human-macaque (figure ??) and 0.79 in chimpanzee-orangutan (figure 4.21)) and
in terms of sign agreement (figures 4.20 and 4.22).

Figure 4.19: Orthologs generally exhibit the similar $\Delta \Psi$ values in the stem cell to week 5 comparison in human and macaque cells.
Figure 4.20: Orthologs generally exhibit the same sign $\Delta \Psi$ in the SC-W5 comparison in human and macaque cells.
Figure 4.21: Orthologs generally exhibit the similar ΔΨ values in the stem cell to week 5 comparison in chimpanzee and orangutan cells.
Figure 4.22: Orthologs generally exhibit the same sign $\Delta \Psi$ in the SC-W5 comparison in orangutan and chimpanzee cells.

We further examined the $\Delta \Delta \Psi$ values of timepoint comparisons between pairs of species (for instance, the difference between the $\Delta \Psi_{SC-to-W5}$ values in human and macaque). We find that a large proportion of dynamic events exhibit $\max |\Delta \Delta \Psi_{inner}|$ values of zero (figures 4.23 and 4.25), and that those nonzero values are largely relatively small in magnitude (figures...
4.24 and 4.26).

Figure 4.23: Count of events with zero and non-zero $\max|\Delta \Psi_{inner}|$ values in human-macaque, given either ortholog exhibits $|\Delta \Psi_{inner}| \geq 0.1$ in at least one timepoint comparison and both have $\geq 20$ junction reads.

Figure 4.24: Density of $\max|\Delta \Psi_{inner}|$ values in human-macaque, given either ortholog exhibits $|\Delta \Psi_{inner}| \geq 0.1$ in at least one timepoint comparison and both have $\geq 20$ junction reads.
Figure 4.25: Count of events with zero and non-zero $\max|\Delta\Delta\Psi_{\text{inner}}|$ values in chimpanzee-orangutan, given either ortholog exhibits $|\Delta\Psi_{\text{inner}}|$ $\geq$ 0.1 in at least one timepoint comparison and both have $\geq$ 20 junction reads.

Figure 4.26: Density of $\max|\Delta\Delta\Psi_{\text{inner}}|$ values in chimpanzee-orangutan, given either ortholog exhibits $|\Delta\Psi_{\text{inner}}|$ $\geq$ 0.1 in at least one timepoint comparison and both have $\geq$ 20 junction reads.

Events with non-zero $\max|\Delta\Delta\Psi_{\text{inner}}|$ values also typically exhibit lower sequence con-
servation (assessed by both PhastCons [79] and phyloP [70] scores) in regions proximal to competitive splice sites. Alternative events generally exhibit the typical pattern of low intronic and high exonic conservation, with events having $max|\Delta \Psi_{inner}| = 0$ exhibiting higher conservation scores in most regions except flanking exons (e.g. figures 4.27 and 4.28 for PhastCons scores of SE and MS events in human-macaque and chimpanzee-orangutan respectively, figures 4.77 and 4.80 for phyloP scores). For all event types, only 50 exonic and 250 intronic bases proximal to competitive splice sites (i.e. splice sites in exon-exon junctions that are unique to one isoform, but that contain one splice site common to both isoforms).

![PhastCons in temporally conserved vs non-conserved human/macaque SE/MS events](image)

Figure 4.27: Human-macaque PhastCons and temporal conservation of SE and MS events.
Despite broad similarities in event behavior, there are examples of events exhibiting substantially species-specific behavior. For instance, a 5’-UTR cassette exon in POT1 undergoes a substantial increase in inclusion in the macaque time series, but does not appear utilized at all in the other species (figure 4.29).

Another example occurs in the gene TMEM165. A putatively PTC-containing and antisense
AluJb-overlapping exon undergoes a substantial increase in inclusion throughout the time series in human, modest increase in chimpanzee, and no increase in either orangutan or macaque.

Figure 4.30: Species-specific dynamics of TMEM165 SE event.

While the 3′-splice site sequence is identical in human, chimpanzee, and orangutan, the macaque sequence exhibits a GG rather than AG dinucleotide (figure 4.31). Moreover, while the upstream flanking-to-cassette exon-exon junction is supported by at least a handful of junction reads even in orangutan, no junction reads support the existence of the isoform in macaque. The inclusion of this exon in human neurons is also supported by our analysis of the Blair et al data [8], in which both 14 and 50-week neurons exhibit inclusion of this exon in the cytoplasmic fractions (though not, consistently, in any of the ribosomal fractions).
We highlight a handful of events exhibiting the greatest differences in temporal behavior in the human-macaque (table 4.1) and chimpanzee-orangutan (table 4.2) comparisons below.

Table 4.1: Top 15 most different events in human and macaque by $|\Delta \Delta \Psi_{mid}|$, given non-equivalent $\Delta \Delta \Psi_{inner}$ signs and a within-species $|\Delta \Delta \Psi_{inner}| \geq 0.3$ in at least one comparison.

```
<table>
<thead>
<tr>
<th>Gene</th>
<th>Event type</th>
<th>$\Delta \Psi_{mid}$</th>
<th>Human reads</th>
<th>Human $\Delta \Psi_{inner}$</th>
<th>Human $\Delta \Psi_{mid}$</th>
<th>Macaque reads</th>
<th>Macaque $\Delta \Psi_{inner}$</th>
<th>Macaque $\Delta \Psi_{mid}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLP1</td>
<td>AF</td>
<td>-0.77</td>
<td>21</td>
<td>0</td>
<td>29</td>
<td>-0.67</td>
<td>-0.77</td>
<td>-0.77</td>
</tr>
<tr>
<td>MEST</td>
<td>CF</td>
<td>-0.63</td>
<td>32</td>
<td>0</td>
<td>-0.09</td>
<td>20</td>
<td>-0.46</td>
<td>-0.70</td>
</tr>
<tr>
<td>MACF1</td>
<td>AF</td>
<td>0.62</td>
<td>30</td>
<td>0</td>
<td>-0.09</td>
<td>35</td>
<td>0.50</td>
<td>0.53</td>
</tr>
<tr>
<td>SRSF5</td>
<td>RI</td>
<td>-0.61</td>
<td>24</td>
<td>0.05</td>
<td>47</td>
<td>-0.34</td>
<td>-0.44</td>
<td>-0.44</td>
</tr>
<tr>
<td>RNF146</td>
<td>RI</td>
<td>-0.53</td>
<td>26</td>
<td>0.04</td>
<td>0.08</td>
<td>25</td>
<td>0.50</td>
<td>0.45</td>
</tr>
<tr>
<td>SETX</td>
<td>SE</td>
<td>-0.49</td>
<td>27</td>
<td>0</td>
<td>-0.15</td>
<td>29</td>
<td>-0.44</td>
<td>-0.65</td>
</tr>
<tr>
<td>CELF2</td>
<td>MF</td>
<td>0.42</td>
<td>22</td>
<td>0</td>
<td>0.05</td>
<td>20</td>
<td>0.37</td>
<td>0.47</td>
</tr>
<tr>
<td>TPM1</td>
<td>MF</td>
<td>-0.41</td>
<td>37</td>
<td>0</td>
<td>-0.06</td>
<td>22</td>
<td>-0.33</td>
<td>-0.48</td>
</tr>
<tr>
<td>MIA1</td>
<td>AL</td>
<td>0.40</td>
<td>108</td>
<td>-0.34</td>
<td>-0.40</td>
<td>126</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MEST</td>
<td>AF</td>
<td>0.39</td>
<td>40</td>
<td>-0.34</td>
<td>-0.39</td>
<td>285</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>TENM3</td>
<td>SE</td>
<td>0.38</td>
<td>48</td>
<td>0</td>
<td>0.06</td>
<td>32</td>
<td>0.38</td>
<td>0.44</td>
</tr>
<tr>
<td>POT1</td>
<td>SE</td>
<td>0.36</td>
<td>46</td>
<td>0</td>
<td>0.04</td>
<td>49</td>
<td>0.32</td>
<td>0.40</td>
</tr>
<tr>
<td>LRRC40</td>
<td>SE</td>
<td>0.36</td>
<td>37</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>0.31</td>
<td>0.36</td>
</tr>
<tr>
<td>KIF2A</td>
<td>SE</td>
<td>0.34</td>
<td>21</td>
<td>0</td>
<td>0.19</td>
<td>24</td>
<td>0.42</td>
<td>0.52</td>
</tr>
<tr>
<td>CHD9</td>
<td>AF</td>
<td>0.32</td>
<td>24</td>
<td>0</td>
<td>0.08</td>
<td>47</td>
<td>0.36</td>
<td>0.40</td>
</tr>
</tbody>
</table>
```

4.3.4 Alternative event orthologs involving CDS alterations behave more similar than those that do not

In an attempt to extract clues as to possible functional consequences for this vast array of alternative event dynamics, we sought to examine features of dynamic events in terms of their degree of dynamic conservation. We first grouped events by whether their alternative re-
Table 4.2: Top 15 most different events in chimpanzee and orangutan by $|\Delta\Delta\Psi_{mid}|$, given non-equivalent $\Delta\Delta\Psi_{inner}$ signs and a within-species $|\Delta\Delta\Psi_{inner}| \geq 0.3$ in at least one comparison.

| Gene   | Event type | $\Delta\Psi_{med}$ | Chimpanzee reads | Chimpanzee $\Delta\Psi_{inner}$ | Chimpanzee $|\Delta\Delta\Psi_{med}|$ | Orangutan reads | Orangutan $\Delta\Psi_{inner}$ | Orangutan $|\Delta\Delta\Psi_{med}|$ |
|--------|------------|--------------------|------------------|-----------------------|------------------|------------------|------------------|------------------|
| NDRG4  | MF         | 0.69              | 55               | 0.49                   | -0.54            | 25               | 0.11             | 0.15             |
| RGS3   | MF         | 0.64              | 46               | 0.00                   | 0.00             | 27               | 0.38             | 0.63             |
| RGS3   | MF         | 0.64              | 46               | 0.00                   | 0.08             | 21               | 0.49             | 0.69             |
| RGS3   | MF         | 0.61              | 30               | 0.54                   | 0.61             | 28               | 0.00             | 0.00             |
| HSH35D1| SE         | 0.59              | 46               | 0.00                   | 0.12             | 34               | 0.57             | 0.71             |
| EPB41  | MS         | 0.59              | 108              | -0.01                  | -0.10            | 118              | 0.38             | 0.49             |
| TRCN5  | AP         | 0.57              | 22               | 0.51                   | 0.57             | 41               | 0.00             | 0.00             |
| SPTAN1 | SE         | 0.57              | 26               | 0.36                   | 0.50             | 72               | 0.00             | -0.07            |
| MYO18A | SE         | 0.57              | 26               | -0.39                  | -0.49            | 52               | 0.00             | 0.08             |
| PFPICA | AF         | -0.56             | 59               | 0.65                   | 0.66             | 46               | 0.00             | 0.10             |
| TEME1K3B| SE         | -0.56             | 42               | -0.34                  | -0.13            | 41               | 0.36             | 0.39             |
| NDRG2  | SE         | 0.56              | 40               | -0.46                  | -0.56            | 26               | 0.00             | 0.00             |
| USP54  | AP         | 0.56              | 22               | -0.07                  | -0.10            | 28               | 0.45             | 0.46             |
| AORK1L2| MX         | 0.55              | 27               | 0.00                   | -0.01            | 28               | 0.36             | 0.54             |
| MID1   | AP         | 0.55              | 39               | -0.02                  | -0.12            | 34               | 0.39             | 0.42             |

Table 4.2: Top 15 most different events in chimpanzee and orangutan by $|\Delta\Delta\Psi_{mid}|$, given non-equivalent $\Delta\Delta\Psi_{inner}$ signs and a within-species $|\Delta\Delta\Psi_{inner}| \geq 0.3$ in at least one comparison.

gions overlap putative CDS regions. As described above, this is true for the majority of events.

We then determined the relative number of CDS-overlapping and non-overlapping events exhibiting non-zero $|\Delta\Delta\Psi_{inner}|$ values in cross-species comparisons. We find that a slightly higher proportion of CDS-overlapping events exhibit non-zero $|\Delta\Delta\Psi_{inner}|$ values in all species comparisons (e.g. figures 4.32 and 4.33 for human-macaque and chimpanzee-orangutan comparisons respectively).
Figure 4.32: Events with zero or non-zero $|\Delta \Psi_{inner}|$ by CDS-overlap status in human-macaque.
Figure 4.33: Events with zero or non-zero $|\Delta \Delta \Psi_{inner}|$ by CDS-overlap status in chimpanzee-orangutan.

We then examined the magnitude of the $|\Delta \Delta \Psi_{inner}|$ values directly. We find that CDS-overlapping events generally have lower $|\Delta \Delta \Psi_{inner}|$ values than non-overlapping events ($P = 3.41 \times 10^{-2}$ (Wilcoxon test) for human-macaque and $P = 3.09 \times 10^{-2}$ (Wilcoxon test) for the chimpanzee-orangutan comparison - figures 4.34 and 4.35 respectively).
Figure 4.34: $|\Delta \Delta \Psi_{inner}|$ by CDS-overlap status in human-macaque.
The above comparisons were conducted with events thresholded at $|\Delta \Psi_{inner}| > 0$ but remain largely the same at the higher threshold of $|\Delta \Psi_{inner}| \geq 0.1$ (figures 4.96 - 4.99), though the human-macaque comparison is no longer significant presumably due to the smaller sample size.
4.3.5 Conservation of alternative event dynamics increases as a function of PPI degree

Having examined the correlation of CDS-alteration with dynamic conservation, we sought to establish whether dynamic conservation is correlated with the number of putative protein-protein interactions in which an event’s gene’s protein product is predicted to engage. In order to perform this analysis, we used a human protein-protein interaction network generated by Alanis-Lobato et al [2] that has been disparity-filtered in order to reduce the impact of study bias on the structure of the network (described in detail by Schaefer et al [75]). We determined the degree (i.e. the number of putative interaction partners) of the genes in which we find dynamic alternative events, and examined the relationship between the degree and $|\Delta \Delta \Psi_{inner}|$ values. We first examined the distribution of degree values as a function of whether the event exhibited a zero or non-zero $\max |\Delta \Delta \Psi_{inner}|$ value. We find that events in which $\max |\Delta \Delta \Psi_{inner}| = 0$ generally exhibit slightly higher degree values in both the human-macaque ($P = \frac{1.38 \times 10^{-2}}{}$ - Wilcoxon test, figure 4.36) and chimpanzee-orangutan ($P = \frac{8.93 \times 10^{-4}}{}$ - Wilcoxon test, figure 4.37) comparisons.
Figure 4.36: PPI degree by zero or non-zero $\max|\Delta \Delta \Psi_{inner}|$ in human-macaque.
We then examined the relationship between $\max |\Delta \Delta \Psi_{\text{inner}}|$ magnitude and PPI degree. After binning observations into equally sized bins by PPI degree, we found that $\max |\Delta \Delta \Psi_{\text{inner}}|$ values are typically smaller in genes with higher PPI degree (figures 4.38 and 4.39 for human-macaque and chimpanzee-orangutan comparisons respectively).

Figure 4.37: PPI degree by zero or non-zero $\max |\Delta \Delta \Psi_{\text{inner}}|$ in chimpanzee-orangutan.
Human-macaque $|\Delta \Delta \Psi_{\text{inner}}|$ as a function of PPI degree bin

Given $|\Delta \Psi_{\text{inner}}| > 0$ and $\geq 20$ junction reads

Figure 4.38: $|\Delta \Delta \Psi_{\text{inner}}|$ by PPI degree bin in human-macaque.
Figure 4.39: $|\Delta \Delta \Psi_{inner}|$ by PPI degree bin in chimpanzee-orangutan.
4.3.6 NMD event dynamics are generally conserved and result in the relative increase of the NMD isoform inclusion

Nonsense-mediated decay has been reported in numerous studies (e.g. Lou et al [58], Alrahbeni et al [5]) to become less efficient as a function of neurodifferentiation, and also more broadly differentiation to ectoderm and mesoderm tissue types. To examine whether our data are consistent with these previous reports, we classified events based on whether one (and only one) of the alternative isoforms is exclusively associated with transcripts predicted to contain a PTC. We then examined the overall trend of the event dynamics, finding that events in which the included form is associated with putatively PTC-containing transcripts generally exhibit positive $\Delta \Psi$ values, while events in which the excluded form is associated with putatively PTC-containing transcripts exhibit negative $\Delta \Psi$ values. We highlight these biases in figure 4.40, which includes the stem cell-to-week 5 comparison.
We then sought to establish the dynamic conservation of these NMD events. We restricted the analysis to events which are consistently classified in both species in a given comparison. We then examined the correlation between $\Delta \Psi_{\text{mid}}$ values and the agreement of their signs between species. We observe as before that the putative PTC-associated isoform is typically increasing (i.e. exhibiting positive $\Delta \Psi$ when included and negative $\Delta \Psi$ when excluded),
and that the behavior, particularly at the level of sign agreement, is quite similar between species.

Figure 4.41: NMD event ESC-W5 $\Delta \Psi_{SC-W5}$ human-macaque scatter plot.

Figure 4.41: NMD event ESC-W5 $\Delta \Psi_{mid}$ human-macaque scatter plot.
Figure 4.42: NMD event ESC-W5 $\Delta \Psi_{SC-W5}$ chimpanzee-orangutan scatter plot.

Given $|\Delta \Psi_{\text{excl}}| > 0$ in either and $\geq 20$ junction reads in both
Figure 4.43: NMD event ESC-W5 $\Delta \Psi_{\text{mid}}$ human-macaque sign agreement. Regions outlined by a solid black line indicate events that are consistent with increasing stability of the NMD isoform.
Figure 4.44: NMD event ESC-W5 $\Delta\Psi_{\text{mid}}$ chimpanzee-orangutan sign agreement. Regions outlined by a solid black line indicate events that are consistent with increasing stability of the NMD isoform.
4.4 Conclusions and Discussion

4.4.1 In vitro neurodifferentiation is accompanied by an extensive alternative event dynamics

Our analysis highlights the large number of genes undergoing alternative processing dynamics through neurodifferentiation in models of human, chimpanzee, orangutan, and macaque in vitro cortical neurodifferentiation. We establish that alternative event dynamics occur both coincident with and independent of gene-expression level dynamics (through overall changes in transcriptional output or stability of gene products). Broadly these results suggest the importance of alternative isoforms to the process of neurodifferentiation, and is consistent with previous reports (e.g. [97]).

Our analysis reveals dynamics in many types of alternative events, though most substantially in SE, AF, and RI events (figure 4.3). Of the latter, it is important to note that interpretation is difficult in the case of these particular data. As these data are derived from whole-cell RNA-seq data, intronic reads could be derived from pre-mRNA in the nucleus. We nonetheless included them in the analysis, reasoning that observed changes in putative retention over time may still hint at interesting regulatory occurrences.

We also identify putative transcription factors and kinases underlying the gene expression changes using a method called expression2Kinases [18]. A significant number of these regulators or their putative intermediaries undergo alternative event dynamics. For instances, XPO1, which undergoes a large change in inclusion of a cassette exon, is a putative intermediary of a number of enriched regulators, such as the kinase MAPK14 and the transcription
factor E2F4. E2F4, which did not pass any thresholds for differential expression in our analysis, is subject to nuclear export via XPO1 [34]. Moreover, increased expression of E2F4 has been shown to promote neurodifferentiation in a mouse cell line [67]. Our analysis therefore suggests the hypothesis that XPO1-mediated nuclear export of E2F4 decreases on account of the shifting XPO1 isoform composition, enabling increased E2F4 activity. The inclusion of the XPO1 cassette exon is predicted to cause a frame shift resulting in a downstream PTC, and the modest decrease in XPO1 expression is consistent with the possibility of surveillance via NMD. However, the fact that the included isoform is observed in the heavy polysome fraction of neurons in Blair et al [9] suggests the possibility that the isoform is productive. This isoform is not found in the latest GENCODE annotation and to the best of our knowledge its functions and behaviors at the mRNA and/or protein level have not been established.

4.4.2 Orthologous events exhibit similar dynamics through neurodifferentiation

We further highlight the general similarity of event dynamics. From week-to-week, orthologous events at the very least typically exhibit changes of the same sign (e.g. 4.17), and to the extent that the magnitude of those changes is dissimilar, the difference is often not large (e.g. 4.24). The broadly similar behavior of these alternative events could be considered to be suggestive of the importance of the changes and the specific timing with which they occur.

Despite these broad similarities, we do note instances in which specific events exhibit a high degree of species-specificity in their dynamics. For instance, we noted events in POT1 (figure 4.29), and TMEM165 (figure 4.30), which exhibited macaque and human-specific dynamics respectively. The event in TMEM165 is of further interest because the cassette exon
lacks the canonical 3′-splice site dinucleotide in macaque. Moreover, the fact that this exon and its 3′-splice site overlap an antisense AluJb element suggest it as a possible instance of Alu exonization [80], [49]. Moreover, the fact that it is a putative Alu-derived that exhibits neurodifferentiation-associated dynamics and achieves high relative inclusion even in the cytoplasm despite containing a PTC (supported by the Blair et al data [8]), suggest that this event warrants further investigation.

### 4.4.3 Coding region alterations and protein-protein interactions are associated with more conserved event dynamics

Although observed dynamics are largely similar in orthologous events, we establish an apparent, if slight, bias towards greater temporal conservation of events overlapping (and thus altering) coding sequences, and events in genes with more putative protein interaction partners (e.g. figures 4.34, 4.38). The POT1 SE event previously highlighted (figure 4.29) is an example of a poorly conserved non-CDS overlapping event). The superior temporal conservation of CDS-overlapping events may be in part an indirect consequence of the fact that coding sequences are likely to be better conserved, and therefore the contribution of exonic splicing regulatory elements to the control of alternative splicing is likely to be more consistent among species. The superior conservation of events in genes with more putative protein interaction partners could similarly be an indirect consequence of increased sequence conservation of proteins with higher degree (or apparent degree - as we do not have knowledge of the ground-truth human protein interaction network we cannot guarantee the complete effectiveness of the disparity-filtration performed by Alanis-Lobato et al [2]). However, while the direction of the
causality is unavoidably inscrutable, these observations are consistent with previous hypotheses suggesting that alternative splicing may serve to rewire protein-protein interaction networks (e.g. Ellis et al [26], Irimia et al [43]).

Further in support of this idea is the fact that among alternative events whose alternative exons have some annotated protein feature (or at least one defined in the Exon Ontology databases [87]), the most abundant protein feature type is the intrinsically disordered region. While in aggregate the functional implications of inclusion or exclusion of most intrinsically disordered regions is unknown, a number of intriguing examples (reviewed by Buljan et al [14]) suggest, for instance, that these regions may alter a protein’s intermolecular interactions.

4.4.4 Increased relative abundance of putative PTC-containing transcripts with differentiation is conserved

We find that more often than not the relative abundance of PTC-containing transcripts increases as a function of neurodifferentiation. Moreover, this phenomenon is observed in all considered species, and is consistent with literature highlighting the apparent overall attenuation of NMD efficiency through neurodifferentiation (e.g. Lou et al [58], Alrahbeni et al [5]), and is broadly believed to be a consequence of a decrease in expression in one-or-more NMD factors. Despite this, to the best of our knowledge we are the first to report that a global transcriptomic analysis of AS-NMD is consistent with these previous reports, which emphasized reporter behavior.

These observations, combined with those made previously present something of a conundrum: if NMD efficiency is decreasing and relative abundance of NMD isoforms is in-
creasing, how are cells dealing with the putative increase in potentially toxic transcripts and/or their translated protein products? We examined the overall gene expression of genes in which an NMD isoform’s relative abundance is increasing, and found no evidence of a consistent compensatory transcriptional downregulation. Possibly the translation of NMD isoforms is in some way preferentially impeded, or possibly these isoforms are not (in the appropriate context) actually toxic and their up-regulation is a desired feature. Beyond its well established role in quality control, NMD is also (perhaps less well) known for its role in the programmed regulation of normal (i.e. non-toxic) coding transcripts (reviewed in Nasif et al [65]). Despite the overall bias towards up-regulation of NMD transcripts, it is clear that a number of AS-NMD events do exhibit a decrease in relative NMD isoform abundance. Possibly the down-regulated or unchanging NMD isoforms are indeed toxic and eliminated via orthogonal surveillance mechanisms, while the stabilized isoforms have some functional role to play. We note, for instance, the polysome association of a putative PTC-containing isoform of XPO1. This puzzle warrants considerable further investigation.

Broadly, we have established the identity of a conserved alternative splicing program accompanying in vitro neurodifferentiation in human, chimpanzee, orangutan, and macaque systems. We have provided support for the relevance of alternative splicing to the alteration of protein interaction networks, and support for the impact of global changes in NMD through differentiation.
Bibliography


[28] Ian T. Fiddes, Gerrald A. Lodewijk, Meghan Mooring, Colleen M. Bosworth, Adam D. Ewing, Gary L. Mantalas, Adam M. Novak, Anouk van den Bout, Alex Bishara, Jimi L.


[38] Xin Han, Yanling Wei, Hua Wang, Feilong Wang, Zhenyu Ju, and Tangliang Li. Nonsense-mediated mRNA decay: A ‘nonsense’ pathway makes sense in stem cell biology, Feb 2018.


[47] André Kahles, Cheng Soon Ong, Yi Zhong, and Gunnar Rätsch. SplAdder: identifica-


178


[61] Mohammad A. Mandegar, Nathaniel Huebsch, Ekaterina B. Frolov, Edward Shin, An-


[66] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep sur-
veying of alternative splicing complexity in the human transcriptome by high-throughput

[67] Stephan P. Persengiev, Ivanela I. Kondova, and Daniel L. Kilpatrick. E2F4 Actively
Promotes the Initiation and Maintenance of Nerve Growth Factor-Induced Cell Differenti-

[68] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T
Mendell, and Steven L Salzberg. StringTie enables improved reconstruction of a tran-

[69] Harold Pimentel, Nicolas L Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Dif-
fferential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*,

[70] Katherine S. Pollard, Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. De-
tection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*,

[71] Alejandro Reyes and Wolfgang Huber. Alternative start and termination sites of tran-
scription drive most transcript isoform differences across human tissues. *Nucleic Acids

[72] Andrea Riba, Noemi Di Nanni, Nitish Mittal, Erik Arhné, Alexander Schmidt, and Mi-
haela Zavolan. Protein synthesis rates and ribosome occupancies reveal determinants


[90] Juan L. Trincado, Juan C. Entizne, Gerald Hysenaj, Babita Singh, Miha Skalic, David J.


[95] Xi Wang, Jingyi Hou, Claudia Quedenau, and Wei Chen. Pervasive isoformspecific


[104] Xiaochang Zhang, Ming Hui Chen, Xuebing Wu, Andrew Kodani, Jean Fan, Ryan Doan, Manabu Ozawa, Jacqueline Ma, Nobuaki Yoshida, Jeremy F Reiter, Douglas L Black, Peter V Kharchenko, Phillip A Sharp, and Christopher A Walsh. Cell-Type-Specific Alter-

**Figure 4.45:** Gap statistic determination for clustering of differentially sedimenting h9 ESC events using the CLARA algorithm. As above differential status is assigned when an inner $|\Delta \Psi| \geq 0.1$ is observed between any two fractions, and a minimum number of 20 junction count reads supports the measurement in all samples.
Figure 4.46: Gap statistic determination for clustering of differentially sedimenting h9 NPC events using the CLARA algorithm. As above differential status is assigned when an inner $|\Delta \Psi| \geq 0.1$ is observed between any two fractions, and a minimum number of 20 junction count reads supports the measurement in all samples.
Figure 4.47: Non-NMD event CDS-length ratios of differentially sedimenting events by cluster in H9 ESC. CDS lengths for event isoforms with more than one consistent CDS were averaged. Given $|\Delta \Psi| > 0$ and $\geq 20$ junction reads.
Figure 4.48: Non-NMD event CDS-length ratios of differentially sedimenting events by cluster in H9 NPC. CDS lengths for event isoforms with more than one consistent CDS were averaged. Given $|\Delta \Psi| > 0$ and $\geq 20$ junction reads.
Figure 4.49: Density of Cyto vs Mono $\Delta \Delta \Psi$ between non-NMD proxies and NMD included form events given $|\Delta \Psi| \geq 0.1$ in at least one comparison.
Figure 4.50: Count of event pairs with given sign of Cyto vs Mono $\Delta\Delta\Psi$ between non-NMD proxies and NMD included form events given $|\Delta\Psi| \geq 0.1$ in at least one comparison.
Figure 4.51: Density of Cyto vs Mono $\Delta \Delta \Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta \Psi| \geq 0.1$ in at least one comparison.
Figure 4.52: Count of event pairs with given sign of Cyto vs Mono $\Delta\Delta\Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta\Psi| \geq 0.1$ in at least one comparison.
Figure 4.53: CDS length ratios of proxies and NMD included form events in h9 ESC ($|\Delta \Psi| \geq 0.1$ threshold). These are event pairs with given sign of Cyto vs Mono $\Delta \Delta \Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta \Psi| \geq 0.1$ in at least one comparison.
Figure 4.54: CDS length ratios of proxies and NMD excluded form events in h9 ESC (0.1 threshold). These are event pairs with given sign of Cyto vs Mono $\Delta\Delta\Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta\Psi| \geq 0.1$ in at least one comparison.
A greater number of comparisons are consistent with NMD-dependent effect in H9
P-value = 9.82e-10 (Exact binomial test with p = 0.5), |ΔΨ| ≥ 0.1

Figure 4.55: Number of comparisons between proxies and NMD events in h9 ESC with expected ΔΔΨ sign (0.1 threshold).
Figure 4.56: Density of Cyto vs Mono $\Delta \Delta \Psi$ between non-NMD proxies and NMD included form events given $|\Delta \Psi| > 0$ in at least one comparison.
Figure 4.57: Count of event pairs with given sign of Cyto vs Mono $\Delta \Delta \Psi$ between non-NMD proxies and NMD included form events given $|\Delta \Psi| > 0$ in at least one comparison.
Figure 4.58: Density of Cyto vs Mono $\Delta\Delta\Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta\Psi| > 0$ in at least one comparison.
Figure 4.59: Count of event pairs with given sign of Cyto vs Mono $\Delta \Delta \Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta \Psi| > 0$ in at least one comparison.
Figure 4.60: CDS length ratios of proxies and NMD included form events in h9 ESC (0 threshold). These are event pairs with given sign of Cyto vs Mono $\Delta \Delta \Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta \Psi| > 0$ in at least one comparison.
Figure 4.61: CDS length ratios of proxies and NMD excluded form events in h9 ESC (0 threshold). These are event pairs with given sign of Cyto vs Mono $\Delta\Delta\Psi$ between non-NMD proxies and NMD excluded form events given $|\Delta\Psi| > 0$ in at least one comparison.
A greater number of comparisons are consistent with NMD-dependent effect in H9
P-value = 4.52e-58 (Exact binomial test with $p = 0.5$), $|\Delta\Psi| > 0$

Figure 4.62: Number of comparisons between proxies and NMD events in h9 ESC with expected $\Delta\Delta\Psi$ sign (0 threshold).
Figure 4.63: Number of NMD switch events exhibiting the expected $\Delta \Psi$ sign.
Figure 4.64: Gene expression trends in chimpanzee.
Figure 4.65: Chimpanzee gene expression cluster enriched ontologies (GO biological processes).
Figure 4.66: Gene expression trends in orangutan.
Figure 4.67: Orangutan gene expression cluster enriched ontologies (GO biological processes).
Figure 4.68: Gene expression trends in macaque.
Table 4.69: Macaque gene expression cluster enriched ontologies (GO biological processes).

<table>
<thead>
<tr>
<th>Enrichment Terms</th>
<th>Log10(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood coagulation</td>
<td></td>
</tr>
<tr>
<td>Intercellular signaling pathway</td>
<td></td>
</tr>
<tr>
<td>Chromatin silencing at DNA</td>
<td></td>
</tr>
<tr>
<td>DNA replication-dependent nucleosome assembly</td>
<td></td>
</tr>
<tr>
<td>Nucleosome repositioning</td>
<td></td>
</tr>
<tr>
<td>Extracellular matrix organization</td>
<td></td>
</tr>
<tr>
<td>Regulation of gene silencing by mRNA</td>
<td></td>
</tr>
<tr>
<td>Protein turnover</td>
<td></td>
</tr>
<tr>
<td>Lipid transport</td>
<td></td>
</tr>
<tr>
<td>Cell adhesion</td>
<td></td>
</tr>
<tr>
<td>CMAH-dependent nucleosome assembly</td>
<td></td>
</tr>
<tr>
<td>DNA replication</td>
<td></td>
</tr>
<tr>
<td>DNA replication-dependent nucleosome assembly</td>
<td></td>
</tr>
<tr>
<td>G1/S transition of mitotic cell cycle</td>
<td></td>
</tr>
<tr>
<td>Cell proliferation</td>
<td></td>
</tr>
<tr>
<td>Cell death</td>
<td></td>
</tr>
<tr>
<td>Kinase-mediated organization</td>
<td></td>
</tr>
<tr>
<td>Regulation of translation involved in G1/S transition of mitotic cell cycle</td>
<td></td>
</tr>
<tr>
<td>Positive regulation of cell proliferation</td>
<td></td>
</tr>
<tr>
<td>Positive regulation of transcription by RNA polymerase II</td>
<td></td>
</tr>
<tr>
<td>Negative regulation of transcription by RNA polymerase II</td>
<td></td>
</tr>
<tr>
<td>Regulation of neurogenesis</td>
<td></td>
</tr>
<tr>
<td>Development of primary male sexual characteristics</td>
<td></td>
</tr>
<tr>
<td>Neural tube morphogenesis</td>
<td></td>
</tr>
<tr>
<td>Neural tube closure</td>
<td></td>
</tr>
<tr>
<td>Regulated axon guidance</td>
<td></td>
</tr>
<tr>
<td>Neuronal process</td>
<td></td>
</tr>
<tr>
<td>Negative regulation of neuron differentiation</td>
<td></td>
</tr>
<tr>
<td>Neuronal process</td>
<td></td>
</tr>
<tr>
<td>Regulation of post-synaptic membrane potential</td>
<td></td>
</tr>
<tr>
<td>Regulation of ion transport</td>
<td></td>
</tr>
<tr>
<td>Regulation of synaptic vesicle cycle</td>
<td></td>
</tr>
<tr>
<td>Dorsal spinal cord development</td>
<td></td>
</tr>
<tr>
<td>Neuropeptide signaling pathway</td>
<td></td>
</tr>
<tr>
<td>Neuropeptide signaling pathway</td>
<td></td>
</tr>
<tr>
<td>Glycan carbohydrate metabolic process</td>
<td></td>
</tr>
<tr>
<td>Synaptic transmission, GABAergic synaptic asymmetry</td>
<td></td>
</tr>
<tr>
<td>Positive regulation of synapse assembly</td>
<td></td>
</tr>
<tr>
<td>Chemical synapse transmission</td>
<td></td>
</tr>
<tr>
<td>Regulation of ion transport</td>
<td></td>
</tr>
<tr>
<td>Endocannabinoid receptor signaling pathway</td>
<td></td>
</tr>
<tr>
<td>Regulation of post-synaptic membrane potential</td>
<td></td>
</tr>
<tr>
<td>Excitatory synaptic transmission</td>
<td></td>
</tr>
<tr>
<td>Neuronal process</td>
<td></td>
</tr>
<tr>
<td>Neuron maturation</td>
<td></td>
</tr>
<tr>
<td>Neuronal process</td>
<td></td>
</tr>
<tr>
<td>Neuron maturation</td>
<td></td>
</tr>
<tr>
<td>Neuronal process</td>
<td></td>
</tr>
<tr>
<td>Neuronal process</td>
<td></td>
</tr>
<tr>
<td>Neuronal process</td>
<td></td>
</tr>
<tr>
<td>Excitatory synapse formation</td>
<td></td>
</tr>
<tr>
<td>Neuron differentiation</td>
<td></td>
</tr>
<tr>
<td>Membrane-bending movement</td>
<td></td>
</tr>
<tr>
<td>Positive regulation of ERK1 and ERK2 cascade</td>
<td></td>
</tr>
<tr>
<td>Positive regulation of cell cycle progression</td>
<td></td>
</tr>
<tr>
<td>Negative regulation of cell cycle progression</td>
<td></td>
</tr>
<tr>
<td>Negative regulation of peptidyl-tyrosine phosphorylation</td>
<td></td>
</tr>
<tr>
<td>Positive regulation of peptidyl-tyrosine phosphorylation</td>
<td></td>
</tr>
<tr>
<td>Negative regulation of protein kinase B signaling</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.70: Chimpanzee alternative event cluster trends and enriched ontologies (GO biological processes).
Figure 4.71: Orangutan alternative event cluster trends and enriched ontologies (GO biological processes).
Figure 4.72: Macaque alternative event cluster trends and enriched ontologies (GO biological processes).

Figure 4.73: Human-macaque PhastCons and temporal conservation of A3 events.
Figure 4.74: Human-macaque PhastCons and temporal conservation of A5 events.

Figure 4.75: Chimpanzee-orangutan PhastCons and temporal conservation of A3 events.
Figure 4.76: Chimpanzee-orangutan PhastCons and temporal conservation of A5 events.

Figure 4.77: Human-macaque phyloP and temporal conservation of SE/MS events.
Figure 4.78: Human-macaque phyloP and temporal conservation of A3 events.

Figure 4.79: Human-macaque phyloP and temporal conservation of A5 events.
Figure 4.80: Chimpanzee-orangutan phyloP and temporal conservation of SE/MS events.

Figure 4.81: Chimpanzee-orangutan phyloP and temporal conservation of A3 events.
Figure 4.82: Chimpanzee-orangutan phyloP and temporal conservation of A5 events.

Figure 4.83: Human-macaque PhastCons and temporal conservation of SE/MS events.
Figure 4.84: Human-macaque PhastCons and temporal conservation of A3 events.

Figure 4.85: Human-macaque PhastCons and temporal conservation of A5 events.
Figure 4.86: Human-macaque PhastCons and temporal conservation of A5 events.

Figure 4.87: Chimpanzee-orangutan PhastCons and temporal conservation of SE/MS events.
Figure 4.88: Chimpanzee-orangutan PhastCons and temporal conservation of A3 events.

Figure 4.89: Chimpanzee-orangutan PhastCons and temporal conservation of A5 events.
Figure 4.90: Human-macaque phyloP and temporal conservation of SE/MS events.

Figure 4.91: Human-macaque phyloP and temporal conservation of A3 events.
Figure 4.92: Human-macaque phyloP and temporal conservation of A5 events.

Figure 4.93: Chimpanzee-orangutan phyloP and temporal conservation of SE/MS events.
Figure 4.94: Chimpanzee-orangutan phyloP and temporal conservation of A3 events.

Figure 4.95: Chimpanzee-orangutan phyloP and temporal conservation of A5 events.
Figure 4.96: Events with zero or non-zero $|\Delta \Psi_{\text{inner}}|$ by CDS-overlap status in human-macaque.
Figure 4.97: Events with zero or non-zero $|\Delta \Psi_{inner}|$ by CDS-overlap status in chimpanzee-orangutan.
Figure 4.98: $|\Delta \Delta \Psi_{\text{inner}}|$ by CDS-overlap status in human-macaque.
Figure 4.99: $|\Delta \Delta \Psi_{inner}|$ by CDS-overlap status in chimpanzee-orangutan.