Development and Application of Next-Generation Sequencing Methods to Profile Cellular Translational Dynamics

by

Sang Young Chun

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2018

Doctoral Committee:

Assistant Professor Ryan E. Mills, Chair
Assistant Professor Jeffrey M. Kidd
Professor Mats Ljungman
Professor Alexey I. Nesvizhskii
Associate Professor Peter K. Todd

Sang Young Chun
stonyc@umich.edu
ORCID id: 0000-0002-4954-9446

For my family.

ACKNOWLEDGEMENTS

Foremost, I would like to thank my advisor Ryan Mills whose guidance, support and mentorship were instrumental to my professional development. Moreover, I am deeply appreciative of his enthusiasm for video games and other esotery that provided needed respite from the stresses of research. I am especially grateful for the advice and insight provided by my committee members, Jeff Kidd, Mats Ljungman, Alexey Nesvizhskii, and Peter Todd. In addition, I would like to express my deepest gratitude to my former graduate advisors, Arul Chinnaiyan and John Kim, whose guidance and advice were critical to my early development as a graduate student. Finally, I am thankful for my past mentors, Hye Sook Kim, Susan Lyons, Duyen Dang, and Long Dang, their advice and support helped spark my interest in bioinformatics.

I am indebted to the many colleagues and collaborators that I have had the privilege to work with over my graduate research career. From the Todd lab, I am grateful for the insight and expertise provided by Caitlin Rodriguez, her collaboration has informed much of my thesis. From the Chinnaiyan lab, I am especially thankful for the guidance provided by Catie Grasso, she helped me understand that efficient programming and lazy programming are often the same. I am also thankful to Mallory Freeberg from the Kim lab for her peer mentorship and support early in my graduate student career; in many ways, she served as a model for what I had aspired to achieve as a bioinformatician and student. I am also grateful for the guidance and support of Kim lab members Amelia Alessi, Allison Billi, Amanda Day, Vishal Khivansara, Arun Manoharan, Natasha Weiser, and Danny Yang. From the Mills lab, I had the distinct privilege to work alongside great

researchers, compatriots, and friends. I would like to thank Gargi Dayama and Arthur Zhou for making clear the importance of learning to plot outside of Excel. I am thankful to Xuefang Zhao for leading the way as the first student to graduate from the lab, and her ever-present humor. I am grateful to Yifan Wang for being an inferior D.va main, and for establishing the tail end of the distribution for Mills lab time-to-doctorate. I would also like to thank Marcus Sherman for his guidance on all things Python, Alex Weber for her good-natured patience at my jokes, and Catherine Barnier for absolutely not being a Fire Noodles challenge cheater. Finally, I am thankful to have worked alongside Akima George, Nan Lin, Chen Sun, Fan Zhang, and Zhenning Zhang.

I am extremely grateful for Brian Athey and his enthusiastic support over the years, as well as Margit Burmeister and Dan Burns; as a longtime student, I was able to experience firsthand how their devotion to the success of their students, like myself, led to the growth of the department to what it is today. In addition, I would like to thank Jeff de Wet for guiding my first steps as a programmer, and Julia Eussen for her tireless advocacy and enthusiasm. I would also like to take a moment to acknowledge the many friends that I made through my time at Michigan. I am especially grateful for the support of my friends Craig Biwer, Mallory Freeberg, Kathryn Iverson, Sunit Jain, Marianne Juarez, Andy Kong, Lisa LaPointe, Datta Mellacheruvu, Bryan Moyers, Arji Mufti, Lee Sam, Conner Sandefur, Avinash Shanmugam, Kraig Stevenson, Brendan Veeneman, Artur Veloso, Amanda Wilkinson, John Wilkinson, and Casey Wright.

Finally, I would like to thank my friends and family, without whose love and support none of this journey would have been possible. My partners in crime, Aash Bhatt, Matthew Jonovich, Jeff Keeler, Rod Rahimi, and Andrew Woodrow. My parents John and Keum Chun, my sister Jamie Chun, and my brother Danny Chun and his wife Brandie. Most importantly, my family Erin Chun, Elijah Chun, and Eleanor Chun: thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

ACSL3	Acyl-CoA synthetase long chain family member 3
ACTB	Beta-actin
ALK	Anaplastic lymphoma kinase
ALS	Amyloid lateral sclerosis
ATP5I	Adenosine triphosphate synthase, H+ transporting, mitochondrial F_0 complex subunit E
ARF4	Adenosine diphosphate ribosylation factor 4
ATF4	Activating transcription factor 4
AUC	Area under the curve
BAM	Binary alignment/map format
bp	Base pair(s)
BWA	Burrows-Wheeler aligner
CANT1	Calcium-activated nucleotidase 1
CC	Cellular component
CDF	Cumulative distribution function (also, empirical CDF)
CDKN1B	Cyclin dependent kinase inhibitor 1B
cDNA	Complementary deoxyribonucleic acid
CDS	Coding deoxyribonucleic acid sequence
CHX	Cycloheximide
COSMIC	Catalog of Somatic Mutations in Cancer
CTCF	Corrected total cell fluorescence
DE	Differential expression
DNA	Deoxyribonucleic acid
eIF	Eukaryotic initiation factor
ENCODE	Encyclopedia of DNA Elements
ERG	ETS-related gene
ETS	Erythroblast transformation-specific
ETV	Erythroblast transformation-specific variant
FBS	Fetal bovine serum
FDR	False discovery rate
FLOSS	Fragment length optimization similarity score
FMR1	Fragile X mental retardation 1
FMRP	Fragile X mental retardation protein
FPKM	Fragments per kilobase per million mapped reads
FXTAS	Fragile X tremor/ataxia syndrome

GADD34	Growth arrest and DNA damage-inducible 34
GEO	Gene Expression Omnibus
GFP	Green fluorescent protein
GTP	Guanosine triphosphate
HAND2	Heart and neural crest derivatives expressed 2
HGVS	Human Genome Variation Society
HNRNPA2B1	Heterogeneous nuclear riboprotein A2/B1
HRT	Harringtonine
IQR	Interquartile range
IRES	Internal ribosome entry site
kb	Kilobase(s)
LAMB1	Laminin subunit beta 1
LEPR	Leptin receptor
MAPQ	Mapping quality
mESC	Mouse embryonic stem cells
Met	Methionine
MF	Molecular function
MIEF1	Mitochondrial elongation factor 1
mRNA	Messenger ribonucleic acid
mRNA-Seq	Messenger ribonucleic acid sequencing
MS	Mass spectrometry
NanoLuc	Nano luciferase
nLuc	Nano luciferase
nt	Nucleotide(s)
OMIM	Online Mendelian Inheritance in Man
oORF	Overlapping opening reading frame
ORF	Open reading frame
PCBD1	Pterin-4 alpha-carbinolamine dehydratase 1
PCR	Polymerase chain reaction
PIC	Pre-initiation complex
PLXNA2	Plexin A2
PSA	Prostate-specific antigen
PTM	Post-translational modification
QDPR	Quinoid dihydropteridine reductase
RAM	Random access memory
RA	Retinoic acid
RCC1-201	Regulator of chromosome condensation 201
RNA	Ribonucleic acid
RNA-Seq	Ribonucleic acid sequencing
ROC	Receiver operating characteristic

RPKM	Reads per kilobase per million mapped reads
rRNA	Ribosomal ribonucleic acid
SA	Suffix arrays
SAM	Sequence alignment/map format
SRA	Sequence Read Archive
STAR	Spliced Transcripts Alignment to Reference
SV	Structural variants
TIS	Translation initiation site
TMPRSS2	Transmembrane protease, serine 2
TPM	Transcripts per million
TPM4	Tropomyosin 4
tRNA	Transfer ribonucleic acid
TSC1	Tuberous sclerosis 1
TXNRD1	Thiodoxin reductase 1
TSS	Translation start site
UCSC	University of California at Santa Cruz
uORF	Upstream open reading frame
UTP20	Component of the U3 small nucleolar RNA protein complex
UTR	Untranslated region

ABSTRACT

The transmission of genetic information from the transcription of DNA to RNA and the subsequent translation of RNA into protein is often abstracted into a linear process. However, as methods and technologies to measure the genomic, transcriptomic, and proteomic content of cells have advanced, so too has our understanding that the transmission of genetic information does not always flow in a lossless manner. For instance, changes observed in messenger RNA (mRNA) abundance are not always retained at the proteomic level. Indeed, a diverse array of mechanisms have been identified that exert regulatory control over this transmission of information. Next-generation short read sequencing has driven many of these insights and provided increasingly nuanced understanding of these regulatory mechanisms. However, the continued development and application of sequencing methodologies and analytics are required to properly contextualize many of these insights on a more global scale. Ribosome profiling is one such recent advancement which enriches for ribosome-protected fragments of mRNA; sequencing and analysis of these ribosome-protected mRNA fragments enables profiling of the translational content of a sample. The aim of this dissertation is to address the need for the development and application of statistical and analytical algorithms to profile the regulatory factors that contribute to the translational dynamics in cells.

In the first chapter, I survey the development and application of next-generation sequencing methods for the profiling and computational analysis of translation and translational dynamics. In

the second chapter of this thesis, I present SPECtre, a software package that identifies regions of active translation through measurement of the translational engagement of ribosomes over a transcript. SPECtre achieves high sensitivity and specificity in its classification of regions undergoing translation by leveraging the codon-dependent elongation of peptides; this tri-nucleotide periodicity is evident in the alignment of ribosome profiling sequence reads to a reference transcriptome. SPECtre classifies actively translated transcripts according to their coherence in read coverage over a region to an optimal tri-nucleotide signal.

In the third chapter, I describe the application of SPECtre to identify the translation of upstream-initiated open-reading frames that may regulate differentiation in a neuron-like cell model. uORFs are transcripts that result from the initiation of translation from AUG, and under certain biological constraints, from non-AUG sequences localized in the 5' untranslated regions of annotated protein-coding genes. Subsets of these uORFs have been implicated in the regulation of their downstream protein-coding genes in yeast, mice and humans. In this chapter, I provide further evidence for this regulation as well as the spatial context for the functional consequences of uORF translation on downstream protein-coding genes in a neuron-like cell line model of differentiation.

Finally, in the fourth chapter, I outline a strategy using our coherence-based translational scoring algorithm to profile ribosomal engagement over chimeric gene fusion breakpoints in prostate cancer. Here, known breakpoints from current annotation databases are integrated with novel junctions nominated by existing whole genome and transcriptomic gene fusion detection algorithms, and the translational profile over these chimeric junctions using SPECtre is measured. This provides an additional layer of translational evidence to known and novel gene fusion

breakpoints in prostate cancer. Ongoing development of a database and visualization platform based on these results will enable integrative insights into the transcriptional and translational topology of these breakpoints.

CHAPTER 1

INTRODUCTION

1.1 Translation

The transmission of genetic information may be conceptualized as a linear process where double-stranded DNA is transcribed into a single-stranded mRNA, and the translation of this mRNA into functional protein (Figure 1.1).[1] However, this transmission of encoded information in the genome from DNA into RNA to protein is subject to multiple points of coordinated governance; these include, but are not limited to, changes in DNA conformation, epigenetic regulation, transcription and post-transcriptional control, and translational and post-translational regulation. These regulatory mechanisms and checkpoints can exert profound, often non-linear effects on the ultimate abundance of protein in a cell.[2-10]

Translation is the mechanism by which the information encoded by an mRNA is converted into a protein through the systematic addition of tri-nucleotide sequences, alternatively referenced as codons or amino acids, by ribosomes. Thus, ribosomes and their associated complexes represent the translational machinery of a cell. Translation can be roughly divided into four parts: initiation, elongation, termination, and ribosomal dissociation. In eukaryotes, translation initiation is

typically mediated by a pre-initiation complex comprised of eIF2 bound to methionyl-initiator tRNA and a 40S ribosomal subunit. eIF4 facilitates the binding of the PIC to the 5' end of a 7-methylguanosine (m^7G) capped nascent mRNA and scans to the 3' end until an AUG tri-nucleotide translation initiation sequence is recognized.[11,12] In certain instances, translation may initiate from non-AUG near-cognate codons, and under other biological contexts translation may alternatively initiate from non-AUG non-cognate sites.[11,13] Upon successful initiation, a 60S ribosomal subunit is engaged by the PIC to form a complete 80S ribosomal complex, which mediates the elongation of the emergent peptide through the systematic recruitment and addition of subsequent tri-nucleotide codons. Peptide elongation proceeds until one of three in-frame termination signal moieties are recognized, at which point the 80S ribosomal complex is dissociated and the peptide completes its structural formation into protein.

1.2 Regulation of translation initiation

Although regulatory mechanisms may positively or negatively affect translation during elongation, termination or ribosomal dissociation, translational regulation at the point of initiation has been more comprehensively studied.[14-16] Briefly, regulation of translation initiation has multiple points of control including sequence context, mRNA secondary structure, or translation of alternative upstream open-reading frames.[17]

As mentioned previously, translation is typically initiated at AUG tri-nucleotide start sites. However, favorable flanking sequence context may enrich for initiation at specific AUG sites over

others. In bacteria, the Shine-Dalgarno (SD) sequence is a sequence motif that enhances the recruitment of the PIC to AUG initiation sites and promotes protein synthesis.[18] In eukaryotes, context for the site-specific preference of certain AUG sites over others is provided by the flanking Kozak consensus sequence.[12] Due to the 5' to 3' directionality of the scanning PIC, the AUG closest to the 5' end of the m^7G-capped mRNA typically dictates translation initiation. However, presumptive AUG initiation sites may be skipped by the PIC, termed leaky scanning, if their flanking sequence context is unfavorable (e.g. lacking a Kozak consensus sequence). Alternatively, near-cognate non-AUG tri-nucleotides, that differ from the canonical AUG initiation sequence by a one base substitution, may be selected by the PIC to initiate translation. Translation initiation from near-cognate sites is typically mediated by the abundance and activity of additional eIFs, such as eIF1 and eIF5.[19,20] eIF1 and eIF5 have cross-regulatory effects on the initiation of translation from near-cognate sites; eIF1 promotes the utilization of non-AUG start codons, whereas eIF5 antagonizes eIF1-mediated non-AUG near-cognate site translation initiation.

Shorter 5'UTRs, generally less than 20 nucleotides in length, may be prone to leaky scanning and have a detrimental effect on the translational efficiency of the encoded mRNA.[21,22] Further, the secondary structure of the 5' capped mRNA transcript may also modulate the efficiency of translation initiation. Longer 5'UTRs with a stable stem-loop structural conformation may stall the scanning PIC and promote the usage of sub-optimal AUG sites, or in some cases, bias translation initiation from near-cognate non-AUG start codons.[23] In addition, the normal scanning mechanism of the 5' capped mRNA by the PIC may be circumvented by internal ribosome entry sites; IRES are sequence elements that promote the entry of the PIC into the 5'UTR in a m^7G-

independent manner.[24] Taken together, specific sequence motifs in the 5'UTR or secondary structural conformation may promote the binding and recruitment of proteins and complexes that regulate the initiation of translation, sometimes from sites upstream of the canonically encoded protein-coding sequence translation initiation start site. Translation initiation and elongation of open-reading frames from these upstream start sites can modulate the translational efficiency of the canonical protein.[25-30] These ORFs may terminate upstream of the annotated protein-coding translation initiation site, or at an in-frame or out-of-frame termination site that overlaps the canonical CDS. The translation of upstream-initiated ORFs, inclusive of those terminated proximally upstream and within the canonical CDS, may hinder the translation of the annotated protein in a resource-dependent or sterically competitive manner.[31]

For instance, translation of one of two experimentally validated ORFs initiated upstream of the transcriptional regulator ATF4 are dependent on the phosphorylation of eIF2 in response to cellular stress.[32] ATF4 mediates the expression of genes that mitigate damage caused by cellular stress.[33,34] The translation of ATF4 is governed by two ORFs: uORF1, which terminates upstream of the ATF4 CDS, and uORF2, which terminates within the annotated CDS of ATF4. The 5' proximal uORF1 is a positive-acting *cis*-regulatory element that modulates ribosomal scanning and re-initiation of the downstream ATF4 coding sequence.[35] When eIF2-GTP is abundant under normal, non-stressed conditions, scanning ribosomes downstream of uORF1 re-initiate at uORF2, which inhibits translation of the ATF4 protein. In stressed conditions, eIF2 is phosphorylated and results in a reduction of free eIF2-GTP; reduced levels of eIF2-GTP increases the time required for the scanning ribosomes to re-initiate translation at uORF2. Thus, ribosomes downstream of uORF1 scan through and do not re-initiate translation of uORF2, instead translation

is re-initiated at the ATF4 CDS, which in turn promotes the transcription of downstream target genes in response to cellular stress.[35-37]

1.3 Upstream open-reading frames and disease

Translation of upstream-initiated ORFs contributes to the regulatory circuitry of a cell; indeed, their dysregulation has been implicated in the development and progression of various diseases. Sequence variants that introduce, ablate, or otherwise disrupt upstream-initiated ORFs may alter the translational efficiency of the downstream protein-coding sequence and affect phenotypic outcomes.[31] A 4 nucleotide deletion in the 5'UTR of the cyclin-dependent kinase inhibitor CDKN1B results in a frame-shifted termination signal of an upstream-initiated ORF leading to a reduction in the intercistronic space between the normally encoded uORF and the downstream CDS.[36] As a consequence, the efficiency of translation re-initiation at the CDKN1B CDS is decreased and results in down-regulation of the cell cycle inhibitor p27; patients with under-expression of p27 often develop multiple endocrine neoplastic syndrome.[37] Furthermore, translation may alternatively initiate from upstream non-AUG sequences; translation initiation from near-cognate and non-cognate AUG start sites have also been implicated in the etiology of different diseases. Expansion of a CGG tri-nucleotide repeat in the 5'UTR of the fragile X gene FMR1 induces a conformational change in the secondary structure of the nascent mRNA which promotes the recruitment of the 40S PIC to initiate translation in the absence of an AUG or near-AUG start site sequence.[38] Therefore, the translation of upstream-initiated ORFs comprise an

additional layer of complexity in the regulatory topology of both normal cellular function and disease-associated phenotypes.

1.4 Other regulatory factors that affect protein abundance

The abundance of mRNA in a cell may by modulated by transcriptional and post-transcriptional regulatory mechanisms, and the ultimate readout as proteomic abundance may be obscured biologically by other post-translational means. Exhaustive and ongoing research efforts have characterized many of these mechanisms, therefore only selected regulatory pathways that moderate mRNA and protein abundance are detailed here.

Similar to translation initiation, sequence-specific moieties and changes in structural conformation can promote, or alternatively, repress the transcription of DNA into RNA. In general, DNA is structurally condensed and organized around histones; and various proteins and transcription factors may act to mediate the accessibility of that DNA to the transcriptional machinery of the cell.[39] Sequence motifs in the 5'UTR, and in particular the 3'UTR may specifically bind small non-coding microRNAs; microRNAs bound to the 3'UTR of protein-coding transcripts may selectively target them for degradation.[3,40,41] In this way, post-transcriptional regulation of protein-coding genes by microRNAs may alter their ultimate abundance in the proteome.[42] Alternatively, termination signals in-frame to the canonical translation termination site in the CDS of protein-coding genes may trigger the nonsense-mediated decay pathway and repress the abundance of certain proteins.[43] Finally, post-translational modifications may alter the stability,

activity and, in some cases, the localization of proteins; altogether, these mechanisms contribute to determine the proteomic fate of various mRNA transcripts, or may constrain the accessibility of some proteins to orthogonal detection methods.[44]

1.5 High-throughput sequencing and RNA-Seq

The sequencing and assembly of the human genome ushered in an era of unprecedented discovery, and was made possible only by the inestimable efforts of thousands of researchers, as well as critical methodological and technological innovations.[45,46] Sanger sequencing, whole genome shotgun sequencing, and computational assembly algorithms empowered the completion of the initial draft of the human genome.[47-49] One unexpected result of the sequencing and assembly of the human genome was the paucity of protein-coding genes, especially for an organism considerably higher in complexity relative to mice, flies, or even single-celled organisms like yeast. Where it was once assumed that the human genome would consist of hundreds of thousands of protein-coding genes, scientists and lay people alike were astounded when finalized estimates placed this number around roughly 20,000 genes.[50] Thus, efforts were concentrated to develop experimental and computational methods to interrogate alternative models of gene expression regulation that might provide further insight into the complexity of the human transcriptome and proteome.

Highly reproducible, increasingly sensitive and massively parallel methods were developed to isolate and amplify selected sequences, and then computationally assess their composition at single

nucleotide resolution through the marriage of biochemical readouts and imaging technologies. In this way, total RNA-Seq and mRNA-Seq, and high-throughput protein arrays and mass spectrometric methods were developed to interrogate the transcriptomic and proteomic content of a cell, respectively.[51-56] In a typical mRNA-Seq experiment, total RNA is extracted from a single cell, collection of cells, or organism, and then reverse transcribed into cDNA. Following reverse transcription, the cDNA is fragmented into smaller sequences, size-selected, and then ligated to sequencing adapters. The adapter-ligated cDNA sequences are then subjected to several, or up to dozens of cycles of PCR amplification and then sequenced on a high-throughput sequencing machine. Depending on the sequencing technology, an mRNA-Seq experiment can produce anywhere from several million to hundreds of millions of single-ended or paired-end reads. Paired-end reads leverage the size selection step of library preparation to ligate both ends of a cDNA fragment with sequencing adapters; using the expected size distribution between the ends of these paired reads, additional information may be extracted upon sequence alignment that can aid in the identification of splice junctions or structural variants.[57] Single-end and paired-end sequences aligned to a reference transcriptome have been used to infer the relative abundance and structure of various RNA species, including non-coding transcripts, protein-coding genes, and alternatively spliced isoforms.[58]

1.6 Sequence alignment

After reads are collected from a high-throughput sequencing machine, bioinformatics analysis begins with an assessment of sequence content and quality. Artifacts of PCR amplification may be

identified and accounted for, if required, in downstream analyses. Prior to alignment, adapter and contaminant removal, followed by quality trimming of the sequence reads may be required; since many alignment algorithms are sensitive to the quality of reads at single nucleotide resolutions, inclusion of low quality nucleotides may affect not only where a read is aligned, but affect the overall confidence of its alignment.[59] Due to the biochemical properties of prolonged polymerase activity during sequencing-by-synthesis methods, the 3' ends of reads are often of lower quality than those towards the 5' end of the read.[60] Thus, quality trimming software may be used to survey the length of a sequenced read, and then selectively trim its ends if succeeding sequences of nucleotides fall below a pre-determined threshold for minimum quality.

Once read libraries have been assessed for sequence content and subjected to quality control measures, they are then ready for alignment to a reference genome or transcriptome. Since one-to-one string comparison of potentially hundreds of millions reads against millions or billions of possible reference target locations would be computationally expensive, reference sequences are often condensed and indexed using speedy compression algorithms, such as a Burrows-Wheeler transformation.[61] Compression and indexing of reference sequence target locations increases the speed and efficiency of string alignment search by many orders of magnitude. Representative alignment pipelines based on Burrows-Wheeler transformation include BWA, Bowtie, its successor Bowtie 2, and the splice-aware aligner TopHat.[62-65] Alternatively, the STAR algorithm leverages uncompressed suffix arrays to enable increasingly efficient alignment of sequence reads to a reference.[66] Other alignment algorithms, like Sailfish, utilize a k-mer assembly approach to estimate the relative abundance of transcripts.[67] Reads aligned to a reference genome or transcriptome are assessed for quality based on a number of factors, including

but not limited to, individual base pair quality, tolerance for insertions or deletions in the alignment, and parsimony of alignment, where sequence reads mapped to multiple loci are often scored lower than those that are aligned to unique locations on the reference. Since all subsequent analyses, including transcript abundance estimation and structure are based on the overall quality of sequence alignment, careful consideration must be given to aligner selection and parameterization.

1.7 Estimation of relative transcript abundance from RNA sequencing data

The relative abundance of transcripts and isoform structure may be inferred through assessment of those sequence reads that align across, or span, exon-exon splice junctions constructed from a reference transcriptome or transcript annotation database. In addition, paired-end sequence alignment and careful observation of deviations from the expected distribution of the distance between these read pairs can be used to guide increasingly sensitive and specific transcript structure assembly, or even identify *de novo* splice junctions.[58,64] Since the sampling of cDNA fragments from which the sequence reads are derived is typically assumed to be random, the relative abundance of reads aligned to a transcript may be used to infer its relative abundance. Since longer transcripts are more likely to be sequenced, relative transcript abundance is typically summarized as a library and transcript length-normalized estimate. RPKM, FPKM or TPM may be estimated directly from the number of reads annotated to each transcript, however more sophisticated algorithms like Cufflinks and RSEM build statistical models based on alignment parsimony to more accurately assemble transcripts and report their relative abundance.[58,68]

With multiple replicates, the differential expression of genes or transcript isoform abundance may be assessed using statistical metrics, or by using software packages like EdgeR or DESeq.[69-71] In this way, it is possible to observe and report global patterns in gene expression or relative isoform abundance across conditions or sample cohorts.

1.8 Mass spectrometry and estimation of protein abundance

Protein identification and their relative abundance may be estimated using immunoassay-based methods including Western blotting, ELISA, or through reporter-linked immunoluminescent readouts like GFP or firefly luciferase.[72-75] Immunoassay-based methods for protein identification and abundance estimation may be limited by the availability of high-quality, specific antibodies, and the concurrent measurement of multiple protein species in a single experiment can be technically challenger, or time and labor intensive. More recently, protein microarrays were developed for the high-throughput assessment of protein abundance in a sample; they are relatively cheap to produce, and may be customized to experimental requirements. Like immunoassay-based methods, protein microarrays may be limited by the availability of suitable bait antigens, and protein-protein interactions may not be accurately assessed due to immobilization of secondary or tertiary structure. In addition, the overall search space of a protein microarray experiment may be artificially constrained by the density available on the chip.[76]

Technological advances in automation and detection sensitivity have enabled increasingly high-throughput investigation of the proteomic content of a sample using mass spectrometry-based

approaches. In a typical high-throughput MS experiment, proteins are selected according to size, or by affinity purification. Isolated proteins are then fragmented, their constituent peptides are ionized, and the mass-to-charge time of flight of these peptide-ion species is measured. Peptide sequences may then be assembled *de novo*, or more generally, compared against a reference protein and peptide sequence database using search algorithms like MASCOT, X!Tandem, PeptideProphet, or MSFragger.[77-81] Most search algorithms take an input MS/MS spectrum and compare it against a theoretical peptide sequence fragmentation database. Sensitivity and specificity of the peptide search against the fragmentation database may be customized, including parameterization for mass tolerances allowed, proteolytic enzyme constraints, and for the types of post-translational modification that may be considered. The output from the database search is a list of peptide sequence matches ranked according to their likelihood of being a positive match. Relative protein abundance may be estimated from the number of spectra assigned, or more accurately, normalized according to identification parsimony.[82]

1.9 Confounding variables in the comparison of transcriptome and protein abundance

Comparison of transcript abundance estimates based on mRNA-Seq, and those based on MS experiments and database search, are positively but only moderately correlated.[83] There are several factors which may complicate the comparison of abundance estimates across -omic space, some of which have been detailed above. Briefly, post-transcriptional regulation by microRNAs may target certain mRNAs for degradative clearance, and may result in anti-correlated transcript abundance with protein expression. The translation of some mRNAs into protein may be

negatively regulated through competitive or steric inhibition by uORFs. Moreover, PTMs like ubiquitination may result in the selective degradation of protein species and subsequent down-regulation of their abundance when measure by mass spectrometry. In addition, other PTMs may alter the structure, activity, or sub-cellular localization of a protein and complicate their isolation, fragmentation, or detection using MS-based assays.

Global comparison of transcript expression with protein abundance may be further confounded by bioinformatics-based biases inherent to transcriptome or protein searches. For example, liberal parameterization of sequence alignment or during transcript assembly may artificially enrich for false positive transcript structures, or splice junctions with poor evidentiary support. Peptide databases derived from less stringently assembled transcripts may significantly enlarge the potential search space, and result in spurious peptide matches. Furthermore, depending on how the peptide database is constructed, or its search parameterized, certain transcript variants may be alternatively identified as a novel peptide, instead of a post-translational modification of an existing peptide. Thus, considerable thought must be given to how comparative studies of transcript expression and protein abundance are conducted, and interpreted.[84]

1.10 Translational profiling through sequencing of ribosome-protected mRNA

Ribosome profiling, or Ribo-Seq, is a next-generation high-throughput sequencing methodology that was developed to directly investigate the content and dynamics of actively translating ribosomes.[85] Ribosome profiling involves the isolation and massively parallel sequencing of

ribosome-protected fragments of mRNA; since these mRNA fragments, or footprints, are protected from enzymatic digestion during library preparation by their complexed ribosomes, they are representative of regions in a transcript under active translation. A typical ribosome profiling experiment starts with cellular lysis to isolate mRNA bound by ribosomal complexes. Treatment with biochemical translation elongation inhibitors, like cycloheximide, or through combinatorial cold-shocking of the samples, immobilizes ribosomal complexes over an mRNA.[86,87] Exposed regions of mRNA not protected by a ribosome are enzymatically digested using a ribonuclease, and the mRNA:ribosome complexes are isolated using sucrose gradient density centrifugation. Ribosomes are proteolytically disaggregated to free the bound mRNA, and the de-coupled mRNA is then size selected in order to enrich for a range of 28-30 nt fragments typically protected by a ribosome. Sequencing adapters are ligated to the 3' ends of the isolated fragments of mRNA, biotinylated, and then purified to deplete the sample of ribosomal RNA contaminants. The adapter-ligated mRNA fragments are reversed transcribed into cDNA, subjected to several cycles of PCR amplification, and then sequenced on a high-throughput next-generation sequencing machine.[88]

1.11 Bioinformatic considerations for the analysis of ribosome profiling data

Once ribosome profiling library sequence reads are collected from a next-generation sequencing machine, the RPFs are removed of sequencing adapters and then trimmed by minimum base quality thresholds. Ribosome profiling reads may then be aligned to a reference genome or transcriptome using established methods suitable for RNA- or mRNA-Seq experiments. However, even after immunoaffinity depletion during library preparation, rRNA contamination can persist

in the sequenced ribosome profiling reads due to the presence of several kilobases of rRNA in a ribosomal complex.[89] rRNA contaminants can comprise anywhere from 10% to up to 80% of sequenced reads in a given ribosome profiling experiment.[90] Thus, prior to alignment, ribosome profiling reads are typically aligned to a contaminant sequence database to further deplete them of rRNA sequences. Following this bioinformatic depletion of rRNA contaminant species, the remaining unmapped reads are then aligned to a reference genome or transcriptome. Alignment parameters must be carefully selected, and post-alignment quality control must be judiciously applied; because of their shorter length (28-30 nt) compared to the typical length of fragments isolated in many contemporary mRNA-Seq experiments (50-150 nt), ribosome profiling reads are more sensitive to non-parsimonious alignment.[91] In addition, treatment with different biochemical translation elongation inhibitors may enrich for varying lengths of mRNA due to changes in the conformation of immobilized ribosomes. Cycloheximide treatment enriches for mRNA fragments with an average length of 28-30 nt, whereas immobilization through cold-shock alone, enriches for a secondary population of mRNA fragments which range in size from 18 to 22 nucleotides.[92]

Following alignment, positional analysis of RPFs over a transcript must account for the physical localization of an isolated mRNA fragment inside the ribosome.[93] Internal mRNA fragment localization is also inhibitor-dependent; cycloheximide is a translation elongation inhibitor that induces a conformational change in the E-site of the ribosome.[94] Therefore, the aligned position of ribosome profiling sequence reads derived from cycloheximide immobilization must be centered to the A-site of the ribosome. In contrast, harringtonine is a translation elongation inhibitor that prevents the peptidyl transfer of the amino-acid:tRNA to the nascent peptide. Thus,

the aligned position of ribosome profiling reads in harringtonine-derived sequence libraries must be adjusted to the P-site of the ribosome. To determine the A- or P-site offset to adjust the aligned position of ribosome profiling reads derived from cycloheximide or harringtonine immobilization, respectively, reads of the same length that overlap annotated canonical translation initiation sites are extracted. Reads of the same length are collected, and the most common distance from the 5' ends of these reads is used to adjust the aligned position of all reads of identical length (Figure 1.7).[93] For example, if the most common distance from the 5' end of 28 nt length reads overlapping canonical AUG initiation sites is 13 nt, then the 5' aligned position of all reads 28 nucleotides in length are adjusted by 13 nucleotides from their 5' end. In this way, the A- or P-site adjusted aligned position of reads is reflective of the structural-dependent position of the ribosome over the bound mRNA.

1.12 Identification of regions under active translation

Given some of the limitations of comparative analysis between transcript expression and protein abundance, ribosome profiling is strategically positioned to provide an intermediate view of translation and translational dynamics. However, because of the generally shorter lengths of RPFs, alignment of ribosome profiling sequence reads can be especially sensitive to non-parsimonious alignment to multiple locations on a reference genome or transcriptome. Therefore, rudimentary coverage metrics may not be sufficiently robust enough on their own to reliably identify regions in the transcriptome under active translation. To that end, several algorithms have been developed

that leverage critical features of translation, ribosome profiling library preparation, and fragment alignment to more accurately detect actively translated regions of the transcriptome.

Size selection during ribosome profiling library preparation typically enriches for fragments of 28-30 nt in length.[85] Thus, an early translational scoring rubric was developed that was based on deviation from this expected enrichment of specific fragment lengths; this fragment length optimization similarity score was calculated by comparing the distribution of fragments aligned to a transcript against an aggregated distribution of reads aligned to all annotated protein-coding genes.[95] Transcripts with fragment length distributions that closely matched that of the aggregated reference distribution were classified as actively translated. However, because the scoring metric is dependent on efficient size selection during library preparation, FLOSS classification of translation may be sensitive to technical variations in ribosome profiling sequence library generation.[96]

Subsequent translational classification algorithms based on analysis of ribosome profiling sequence data have taken advantage of additional characteristics inherent to translation and the alignment of RPFs to a reference transcriptome. Peptide elongation occurs through the systematic formation of peptidyl:tRNA bonds as amino acids are sequentially added to a nascent protein. Since amino acids, or codons, consist of triplet nucleotides, when ribosome profiling reads are aligned to a reference transcript, the A- or P-site adjust aligned position of these fragments tracks the tri-nucleotide codon-dependent elongation of a peptide (Figure 1.8).[95,97,98] ORFscore is a translational classification metric that scores the enrichment of RPFs over the canonical reading

frame in a transcript.[99] However, ORFscore employs a conservative scoring heuristic that penalizes severely for RPF enrichment outside of the canonical reading frame. Therefore, classification of the translational activity of transcripts based on ORFscore is less robust than other more recently developed algorithms.[96]

More recently, translational profiling algorithms including ORF-RATER, RiboTaper, and SPECtre were developed; all three model the tri-nucleotide periodicity of aligned A- or P-site adjusted ribosome profiling reads to measure translational activity.[96,100,101] ORF-RATER utilizes a regression-based model to score the tri-nucleotide periodicity of aligned RPFs against a re-sampled transcript read coverage profile. Alternatively, RiboTaper and SPECtre directly model aligned and adjusted RPF coverage as a signal process, and then scores its coherence, or power relationship, to a reference tri-nucleotide periodic signal.[102,103] RiboTaper uses mRNA-Seq to build a null distribution and applies Slepian functions to normalize and score the coherence of RPF coverage over a transcript. In contrast, SPECtre scores the windowed (Welch's) coherence of RPF coverage over a transcript against an idealized tri-nucleotide periodic signal. RiboTaper and SPECtre are robust to technical variations in ribosome profiling library preparation, and demonstrate comparable levels of sensitivity and specificity for the identification of regions in the transcriptome under active translation.[96] Notably, SPECtre achieves this accuracy in translational classification in the absence of mRNA-Seq data.

1.13 Perspectives on ribosome profiling as a platform for translational discovery

Entrenched between transcriptional and proteomic survey methodologies, the analysis of ribosome-protected sequence fragments enables the surveillance of translation, and translational dynamics. In the proceeding chapters, we detail work that begins with the development of SPECtre, which measures the translational activity of regions in the transcriptome as a function of its tri-nucleotide periodicity. This work forms the basis for the robust detection of uORFs involved in the regulation of retinoic acid induced neuronal differentiation of SH-SY5Y cells. Finally, we extend SPECtre profiling to investigate the translational landscape of chimeric gene fusion breakpoints in prostate cancer. Breakpoints are identified through integrative curation of known fusion transcripts with *de novo* predictions from mRNA sequencing studies; profiling of these breakpoints enables the integrative characterization of chimeric gene fusion events from transcription to translation.

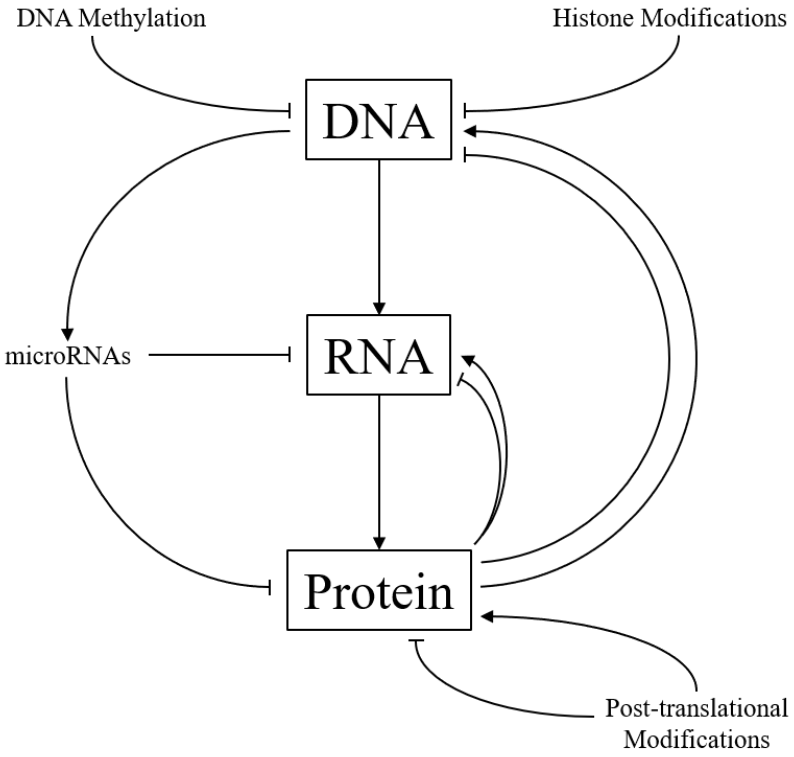

Figure 1.1 Gene expression and protein synthesis regulation. Genomic information encoded as DNA is first transcribed into RNA, then translated into protein through the multi-factorial and omni-directional coordination of various regulatory mechanisms. In concert, these regulatory mechanisms may positively or negatively regulate the abundance, function, or localization of protein in a cell.

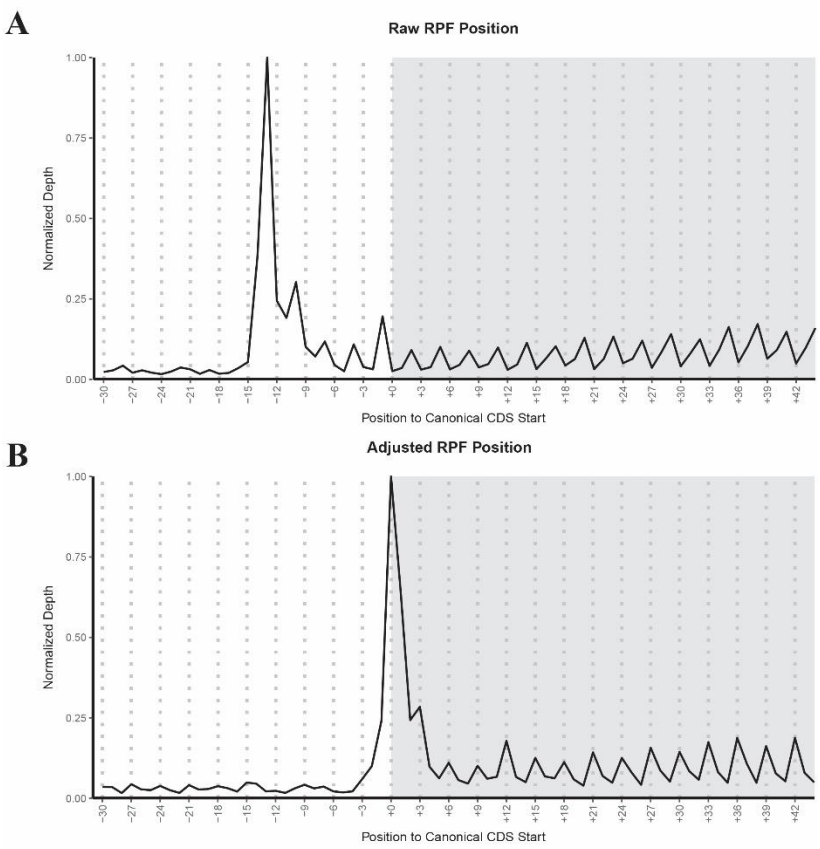

Figure 1.2 Adjustment of RPF alignment position. Metagene profile of raw, and P-site adjusted ribosome profiling reads proximally aligned to canonical AUG translation initiation start sites. The shaded region in the plots represents the canonical coding sequence regions of the metagene profile. A) The raw aligned position of RPFs, often reported as the 5' end of the aligned read to the reference, does not account for its physical position inside of the ribosome it was once protected by. Therefore, raw read coverage of RPFs is enriched upstream of the annotated meta- translation initiation start site of protein-coding genes. B) Following P-site adjustment of the reported RPF alignment positions, the meta-coverage of ribosome profiling reads on protein-coding genes is enriched over the annotated position of translation initiation (denoted as +0).

CHAPTER 2

Spectral profiling of uORF translation in non-differentiated and differentiated neuroblastoma cells

Modified from: Chun, S.Y.*, Rodriguez, C.M.*, Todd, P.K. and Mills, R.E. (2016) SPECtre: a spectral coherence-based classifier of actively translated transcripts from ribosome profiling sequence data. *BMC Bioinformatics* 17(482). DOI: 10.1186/s12859-016-1355-4

The work presented in this chapter of the dissertation was published in BMC Bioinformatics. Prior to its publication, several other algorithms were published, including those that leveraged the tri-nucleotide periodicity of ribosome profiling alignments. Thus, multiple groups were interested in using similar ideas to identify regions of active translation using ribosome profiling sequence data, demonstrating the utility in measuring this phenomenon. Since its publication, this work has been cited a total of five times in the past year, further demonstrating the utility of this approach. Although this work is now published, much of its development was done concurrent to the work presented in the following chapter. Dr. Todd and Ms. Rodriguez provided experimental support, and Ms. Rodriguez generated preliminary ribosome profiling sequence libraries. Dr. Mills guided the computational approach, and I developed and implemented the coherence algorithm.

2.1 Introduction

Ribosome profiling is a next-generation sequencing strategy that enriches for ribosome-protected mRNA footprints indicative of active protein translation [85]. Fragments of mRNA bound by ribosomal complexes are selected for by enzymatic digestion, isolated using a sucrose cushion or gradient, released from their occupying ribosome, size selected by gel electrophoresis, and then sequenced. Thus, sequencing and analysis of ribosome protected fragments of mRNA enables profiling of the translational content of a sample on a transcriptome wide level.

Various algorithms have been developed to differentiate protein coding and non-coding transcripts in ribosome profiling sequence data using fragment length distribution differences and read frame enrichment of aligned reads,[95,99] However, classification based on extreme outlier analysis of fragment length organization similarity score differences are agnostic to the ribosome protected fragment abundance over a transcript. Furthermore, classification based on read frame alignment enrichment (ORFscore) is optimized for canonical open reading frame usage only. In addition, neither of the algorithms described above are available as standalone packages and must be implemented by the user. Published more recently, RiboTaper utilizes a coherence based approach to detect actively translated transcripts from the alignment of ribosome protected fragments; however, the RiboTaper algorithm requires matched ribosome profiling and mRNA sequence libraries and can take multiple days to analyze a single sample.[100]

Here we introduce SPECtre, a classification algorithm based on spectral coherence to identify regions of active translation with high sensitivity and specificity using aligned ribosome profiling sequence reads without the requirement of a matched mRNA sequence library (Figure 2.1 a). SPECtre leverages a key feature of ribosome profiling where sequence reads aligned to a reference transcriptome will track the tri-nucleotide periodicity characteristic of transcripts as they are translated by ribosomes, and reports both significant signals of translation as well as windowed periodicity scores for visualizing results within a genomic context. Options to change the size of windows analyzed, the step size between adjacent windows, false discovery rate, abundance cutoffs to define actively translated versus untranslated score distributions, and parameters to optimize runtime performance are provided to the user to customize. Implementations of FLOSS and ORFscore are included with SPECtre for comparative purposes.

2.2 Implementation

In contrast to non-coding transcripts, ribosome profiling fragments aligned to protein coding transcripts are characterized by a tri-nucleotide periodic signal as ribosome bound mRNAs are translated into protein in a codon dependent manner (Figure 2.1 b). Thus, coding transcripts may be differentiated from non-coding transcripts by the presence or absence of a strong tri-nucleotide periodic signal. To measure the strength of this tri-nucleotide signal, we calculate the spectral coherence over sliding N nucleotide length windows across a transcript (see also Appendix B for Supplemental Materials and Methods).[103] Spectral coherence is a measurement of the power relationship between two signals over the frequency domain, such that two signals with shared frequencies will have high coherence, whereas wo unrelated signals will be of low coherence. The SPECtre score, based on a modified Welch's spectral density estimate of overlapping windows, is calculated for each transcript from a user provided transcript annotation database.[104]

For a given transcript with coordinates defined by the set *C*, the A- or P-site adjusted read positions overlapping those coordinates are extracted from a BAM alignment file. The coverage over each coordinate in the set is summed, then normalized to the position with the highest coverage, such that all coordinate positions defined by the set *C* range from zero (no coverage) to one (highest coverage). The default SPECtre score is calculated as the average (Welch's) coherence over N nucleotide sliding windows across a normalized coverage region against an idealized trinucleotide control signal of the same length. Therefore, the SPECtre score across a normalized coverage

region *R*, with coordinates *C*, against an idealized trinucleotide periodic signal *S* with frequency *j*, over adjacent *N* nucleotide windows is given by:

(1)

$$Spec_{RS,j} = \frac{1}{M}\sum_{m=1}^{M} Coh_{R_{m,m+N}S_N,j} \ for\ all\ m+N \in C$$

Alternatively, the number of sliding windows (W_n) over the coordinate set *C*, may be modified based on the step size between each window. Therefore, given a coordinate set *C*, and step size of *L*:

(2)

$$W_n = C_{Ln}, for\ n \geq 1\ and\ L \geq 1$$

Therefore, the default SPECtre score of a normalized coverage region *R*, at frequency *j* of an idealized trinucleotide signal *S*, over *N* nucleotide sliding windows with a step size of *L*, is given by the equation:

(3)

$$Spec_{RS,j} = \frac{1}{M}\sum_{m=1}^{M} Coh_{R_{m,m+N}S_N,j} \ for\ all\ m \in W_n\ and\ all\ m+N \in C$$

Distributions of these scores are generated using a user-defined fragments per kilobase per million reads cutoff to differentiate transcripts under active translation from those that are not; these distributions are then used to derive a minimum SPECtre score threshold for active translation given a pre-determined false discovery rate, as well as the posterior probability that a given transcript or region is actively translated.[105]

Ribosome profiling libraries treated with cycloheximide typically isolate RPFs of 28 to 30 nucleotides in length; these fragments align with high fidelity to protein-coding regions. However, in the absence of cycloheximide, conformational changes in the ribosomal complex enrich for a shorter range of RPFs that also map with high fidelity to protein-coding regions.[92] Enrichment of these shorter-range fragments may obscure the trinucleotide signal profiled by coherence-based classifiers, like SPECtre, and may under-estimate the number of actively translated ORFs. We simulated increasing variance of RPF lengths outside of the expected enrichment of 28-30 nt length fragments through a biased sampling of reads aligned to the housekeeping gene ACTB. With increased bias, the RPF length distribution is no longer enriched in fragments of 28-30 nt in length, but instead progressively resembles a uniform distribution (Supplemental Figure A.1). Biased re-sampling of 10,000 out of over 500,000 P-site adjusted reads aligned to ACTB was performed over 10,000 trials, and in each trial the sampled reads were converted into normalized coverage, then scored by SPECtre. Using an extreme outlier cutoff, this biased re-sampling analysis suggests that SPECtre scoring remains robust under increased variance in sequence library fragmentation (Supplemental Figures A.1 and A.2).

2.3 Results

We assessed the sensitivity and specificity of each classification algorithm using recently published ribosome profiling and mRNA-Seq data derived from HEK293 cells.[100] For the comparative analysis of each classification algorithm in the HEK293 ribosome profiling library,

RiboTaper (version 1.3) was run against published read alignments using the included GENCODE (v19) transcript annotation database.[106] The highest scoring RiboTaper ORFs were extracted from the *orfs_found* results file using the transcript identifiers and scoring method from the *ORFs_max* output. These ORFs were then scored by SPECtre (using default parameters), FLOSS and ORFscore, and the relative performance of each algorithm was assessed by receiver operating characteristic analysis. Previous work has benchmarked classifier performance using a series of transcript FPKM cutoffs or other coverage based metrics [95,99,100]. Therefore, ROC analyses were performed using a series of ORF abundance cutoffs based on FPKM to differentiate those under active translation from those that are not. In this manner, we are able to assess the ability of each approach to identify ORFs with signatures of active translation in the interrogated cell type. We performed ROC analyses and calculated the AUC over pre-defined RPF abundance cutoffs (0.5, 1.0, 3.0, 5.0 and 10.0 FPKM) to assess the relative performance of each classification algorithm to accurately define regions of active translation. In HEK293 cells, SPECtre conforms with high fidelity to RiboTaper classification and outperforms both FLOSS and ORFscore to identify actively translated ORFs (Figure 2.2 a and b).

We also used previously published ribosome profiling data derived from mouse embryonic stem cells and zebrafish embryos to assess the performance of SPECtre, FLOSS and ORFscore in the absence of mRNA-Seq data (Supplemental Table A.1); RiboTaper was excluded from these analyses due to its requirement of matched mRNA-Seq data. Ribosome profiling sequence reads from each set were aligned to the mouse or zebrafish reference genome and transcriptome, respectively. Antisense, overlapping and neighboring protein coding and non-coding transcripts were removed from the analysis using methods described previously [95]. The FLOSS, ORFscore

and SPECtre metrics were calculated for each remaining transcript and ROC analyses were carried out as described above. SPECtre remains robust in its classification of actively translated transcripts in the standalone mESC ribosome profiling library (Figure 2.2 c and Supplemental Table A.2), and exhibits a marked improvement in accuracy in a meta-analysis of ribosome profiling libraries derived from zebrafish embryos (Figure 2.2 d).

A unique feature of SPECtre is its ability to report and visualize signals of periodicity in the context of surrounding genomic features. Graphical output from SPECtre analysis is shown for two representative transcripts (Figure 2.3 a and b). A condensed transcript profile of RCC1-201 (ENST0000398598) is show in Figure 2.3(a) with the 5'UTR and 3'UTR depicted by the narrow black lines, and the CDS region depicted with the thicker black line. In gray is the normalized P-site adjusted read coverage over the transcript, with the posterior probability calculated by SPECtre denoted by the black line. The dashed horizontal line represents the translational threshold calculated by SPECtre at a false discovery rate of 0.05. In addition to the transcript structure depicted in Figure 2.3(b) are two upstream open reading frames detected by separately by RiboTaper (asterisked black bars) in the MIEF1 (ENST0000325301) transcript. Although the 5'UTRs of both RCC1-201 and MIEF1 are profiled by RPF coverage, SPECtre analysis identifies only the uORFs in the 5'UTR of MIEF1, also identified previously by RiboTaper, with a trinucleotide signal of sufficient strength to be indicative of translational potential.[100]

A further analysis of these and other ORFs assessed by both SPECtre and RiboTaper show a very high degree of score consistency between the two algorithms (Figure 2.3 c) in addition to their

comparable overall accuracy. However, SPECtre has been designed to be fast and efficient and exhibits a runtime almost one third of that required by RiboTaper (Figure 2.3 d) without requiring RNA-Seq data. This is achieved through SPECtre's ability to chunk experiments and parallelize analyses over multiple threads, depending on available computational resources, which enables this exceedingly fast runtime relative to existing methods and decreases the computational barrier between library alignment to application and validation. For these experiments, SPECtre analysis was split by chromosome and run using 8 processors, with 32 gigabytes of RAM allocated; RiboTaper was run with default parameters, using 8 processors and 64 gigabytes of RAM. Both SPECtre, and RiboTaper were run on a high-performance computing cluster running Red Hat Enterprise Linux version 6.4 (Santiago). For installation simplicity and application efficiency, SPECtre has been written in Python with minimal third-party dependencies; the only non-standard Python libraries required for SPECtre analysis are RPy2, NumPy, HTSeq, SAMTools, PyFASTA, PySAM, and the R package ROCR.

2.4 Conclusions

SPECtre is a flexible, lightweight, command line driven analytical package that identifies regions of active translation through modeling of the tri-nucleotide periodicity characteristic of translation by ribosomes, and does so with high fidelity to a recently published method that relies on a similar coherence based approach. SPECtre classification also outperforms prevailing algorithms based on fragment length distribution profiling and reading frame occupancy enrichment. SPECtre is robust across ribosome profiling libraries derived from multiple organisms and cell types, even in

the absence of matching mRNA-Seq data, and is capable of identifying active translation in regions previously thought to be non-coding. Furthermore, SPECtre is under continuous development to optimize compute run time and memory overhead in order to facilitate the efficient and accurate investigation of translational dynamics through ribosome profiling sequence analysis.

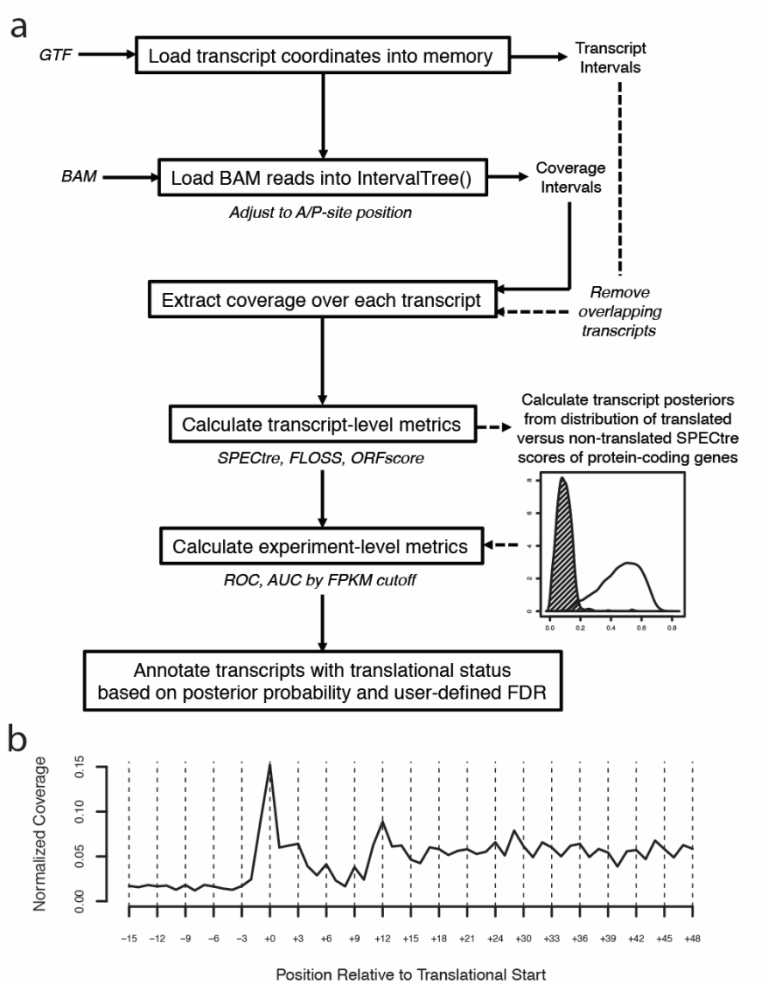

Figure 2.1 SPECtre pipeline and tri-nucleotide periodicity a) SPECtre analytical pipeline, input files, formats and outputs. b) Ribosome profiling read coverage averaged over annotated protein coding transcripts demonstrates a tri-nucleotide periodic signal characteristic of translation by ribosomes.

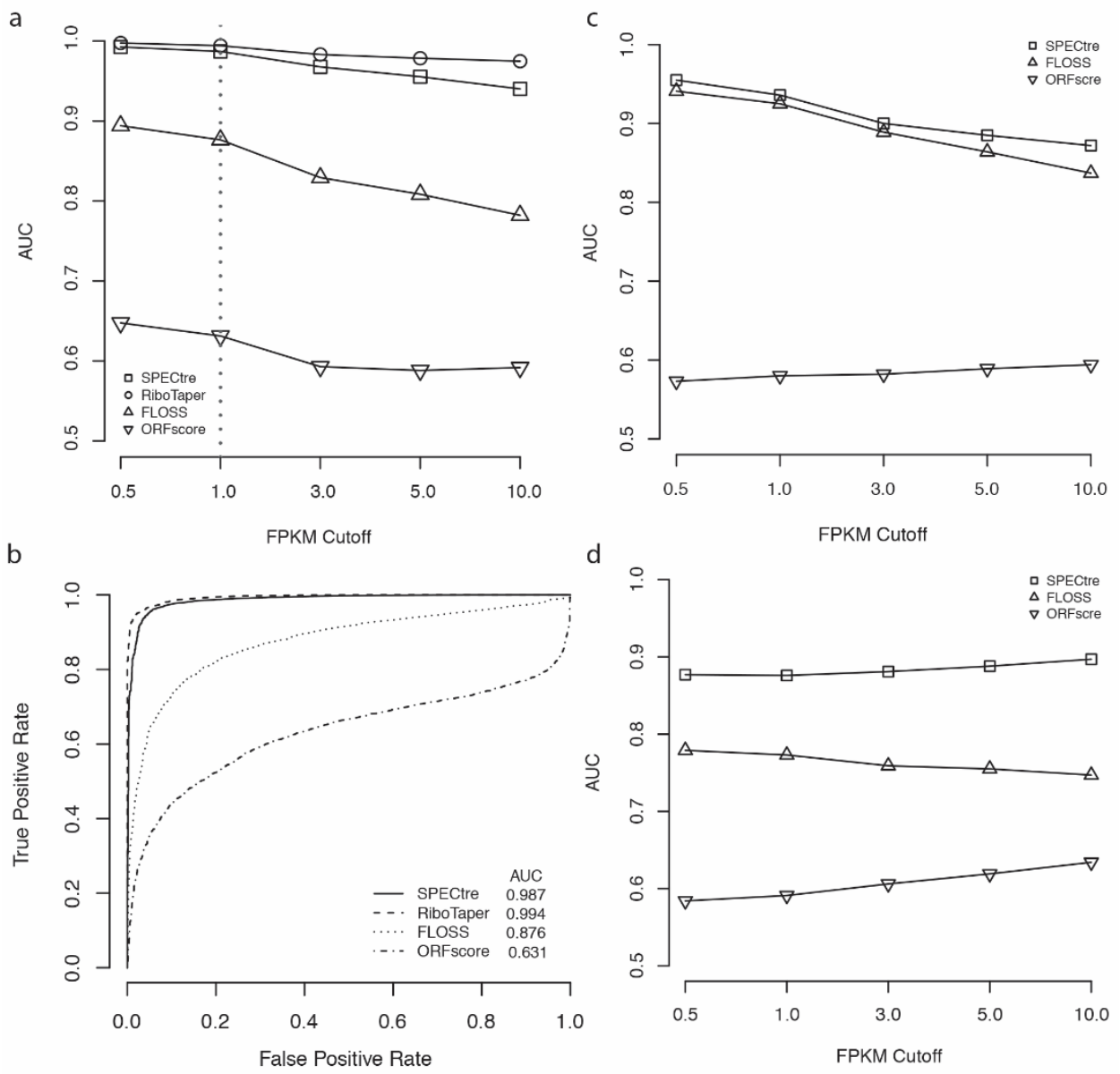

Figure 2.2 Comparative analysis of SPECtre against previously published methods. a) Performance of SPECtre, RiboTaper, FLOSS and ORFscore classification of ORF translation at various PRF abundance cutoffs as measured by AUC in ribosome profiling of HEK293 cells.[100] b) ROC curves of SPECtre, RiboTaper, FLOSS, and ORFscore at a cutoff of 1.0 FPKM. c) Performance of SPECtre, FLOSS, and ORFscore classification of ORF translation in ribosome profiling of mESC at various RPF abundance cutoffs as measured by AUC.[95] d) Performance of SPECtre, FLOSS, and ORFscore classification of ORF translation in a meta-analysis of ribosome profiling in zebrafish over various RPF abundance cutoffs as measure by AUC.[99] All

SPECtre analyses were based on 30 nt sliding windows, using a step size of three between each window.

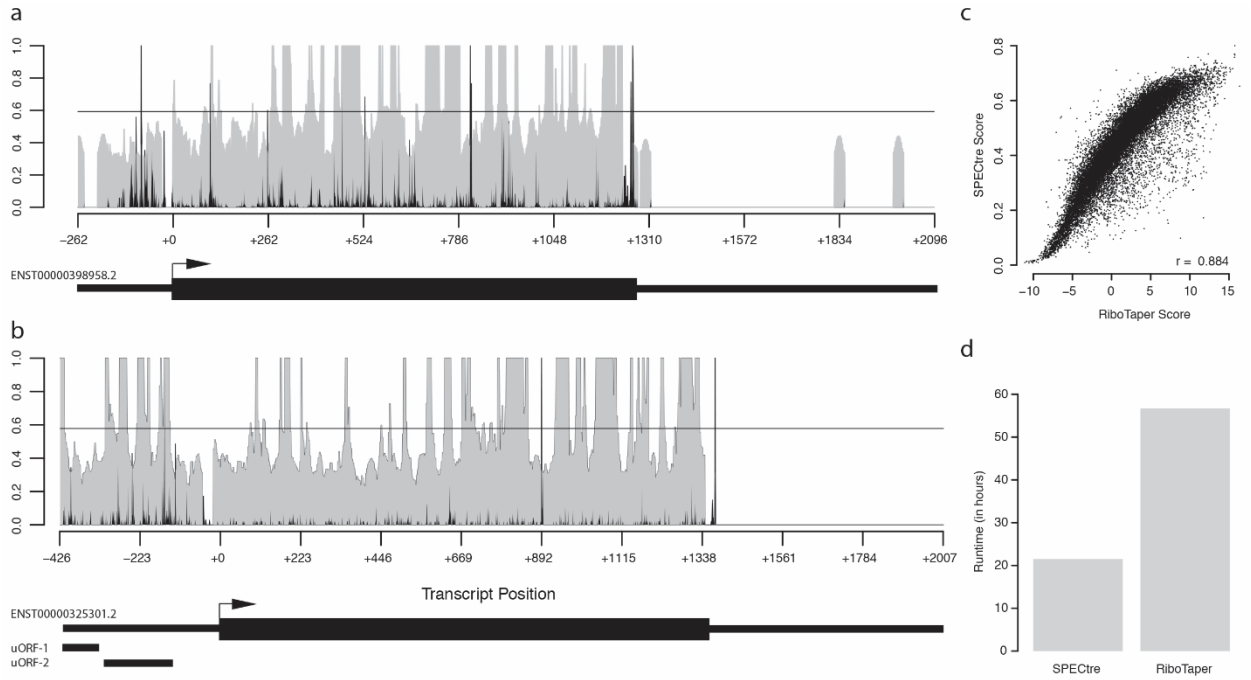

Figure 2.3 Examples of SPECtre results and runtime comparison to RiboTaper. a) SPECtre posterior probability profile (shaded gray) and normalized P-site adjusted read coverage (black bars) over the transcript structure of RCC1-201. Solid, horizontal black line represents the translational threshold as calculated by SPECtre at a false discovery rate of 0.05. Arrow indicates position of annotated translational start site. Thin black boxes (left to right) denote the 5'UTR and 3'UTR, respectively, with CDS (thick black box) in between. b) SPECtre posterior probability profile (as above) over the transcript structure of MIEF1. Thin, black boxes under transcript structure denote two uORFs previously identified by RiboTaper analysis. c) Scatter plot of SPECtre and log2(RiboTaper) scores over assessed ORFs. d) Comparison of SPECtre (left) and RiboTaper (right) total compute time, in hours.

CHAPTER 3

Translational profiling of uORFs in a cellular model of neuronal differentiation

Modified from manuscript in preparation: Rodriguez, C.M.*, Chun, S.Y.*, Mills, R.E. and Todd, P.K. (2017) Translational profiling of uORFs in a cellular model of neuronal differentiation.

The work presented in this chapter has been modified from a manuscript in preparation. The project context, design and scope were conceived by Dr. Todd, and Ms. Rodriguez. Ms. Rodriguez grew and differentiated the SH-SY5Y cells, prepared the sequencing libraries, and performed the biological validation experiments. Dr. Mills directed the computational approach and analyses, and I developed the uORF prediction pipeline, processed the data, and led the bioinformatics analyses.

3.2 Introduction

The development of massively parallel next-generation sequencing and application of robust analytical methods have empowered the quantitative characterization of the gene expression changes that underlie phenotypic diversity and cellular function. Corresponding advancements in mass spectrometry and peptide identification algorithms have made it possible to track gene expression from the transcriptome through the proteome.[84] However, critical differences in sampling, scale, and search methodologies makes linear comparison across the two platforms imperfect.[107] Furthermore, we have increasingly come to understand the complex regulatory network of interactions and changes that makes comprehensive profiling of the genome, transcriptome and proteome elusive. Although components of this underlying regulatory network may be inferred from evaluating changes in transcript abundance, translational profiling may

provide additional context for how these changes contribute to cellular function and diversity.[85,108-110]

Translational regulation of protein expression drives changes affecting the state, or programming, of a cell under a dynamic set of conditions.[26-30] As such, profiling ribosome occupancy across the transcriptome under various cell states has proven a powerful platform to assess the relationship between mRNA abundance and its translational output to the proteome.[85,108-110] Early ribosome profiling experiments demonstrated the broad capacity of the technique to provide sensitive and condition-specific measurement of changes in mRNA translational efficiency that impact cellular processes from meiosis to development.[85,108-119] Furthermore, these studies also revealed for the widespread presence of translating ribosomes in regions outside of annotated protein-coding sequences.[99,108,109,120-127] In this study, we used SH-SY5Y neuroblastoma cells treated with retinoic acid to investigate the translational landscape that underlies their terminal differentiation into a population of cells that resemble neurons. While previous studies have investigated the genetic changes that contribute to this process, many have focused on transcript-level changes that may not be reflected as protein-level differences.[128,129] This work provides further transcriptomic and translational context for the network of protein synthesis shifts associated with cellular differentiation.

Early ribosome profiling experiments presented evidence for the enrichment of RPFs in the 5'UTRs of protein-coding transcripts. [99,108,109,120-126] Based on these reports, several computational and analytical algorithms have been developed to assess translation of these 5' leader sequences, or uORFs.[95,96,99-101,130,131] These studies underscored the regulatory

significance of uORF translation on downstream protein-coding genes, and established their prevalence as translational regulators, but often lacked one of three features: 1) implementation of stringent parameters to identify translated uORFs, 2) inclusion of uORFs that utilize a non-canonical TIS, and 3) investigation of a biological process that define a novel role for uORFs. In this study, we seek to uncover new roles for translation in the 5' leader under various cell conditions apart from those already studied. Several studies have elucidated a role for uORFs in circadian clock regulation, zebrafish development, cell proliferation, and tumorigenesis.[99,132-135] Moreover, as studies emerge in distinct cell-types, more multi-cistronic transcripts are unveiled across the genome with potential to contribute to the regulation of the proteome, but may be constrained to a specific set of conditions.

The 5' leader sequence in mRNA is a well-characterized source of protein synthesis regulation.[17] Translation of regions in the 5' leader sequence may indirectly, or directly, regulate the synthesis of the canonical protein product; this can occur by altering mRNA and translation factor binding, influencing mRNA stability, or through the interaction of the newly synthesized 5' protein product with the ribosome.[31] The translation of one of two experimentally validated uORFs initiated upstream of the transcriptional regulator ATF4 is dependent on the phosphorylation of eIF2 in response to cellular stress (Figure 1.3).[32] ATF4 is a transcription factor that mediates the expression of genes that mitigate cellular damage caused by conditions of stress.[33,34] The translation of ATF4 is governed by two ORFs: uORF1 is terminated upstream of ATF4 and the second ORF, uORF2, which terminates within the annotated CDS of ATF4. The 5' proximal uORF1 is a positive-acting *cis*-regulatory element that modulates ribosomal scanning and re-initiation of the downstream ATF4 coding sequence.[35] When eIF2-GTP is abundant

under normal, non-stressed conditions, scanning ribosomes downstream of uORF1 re-initiate at uORF2, which inhibits translation of the ATF4 protein. In stressed conditions, eIF2 is phosphorylated and results in a reduction of free eIF2-GTP; reduced levels of eIF2-GTP increases the time required for the scanning ribosomes to re-initiate translation. Thus, ribosomes downstream of uORF1 scan through and do not re-initiate translation of uORF2, instead translation is re-initiated at the ATF4 CDS which mediates the activity of downstream target genes in response to cellular stress.[35-37]

Previous ribosome profiling studies have revealed an overall repressive role for uORFs on the translational efficiency of the downstream protein; however, these analyses limited their investigation to uORFs initiated from AUG start codons in the 5'UTR. Although translation initiation from non-AUG start sites remains less characterized, the prevalence and functional impact of non-AUG-initiated translation has been studied in certain disease-related contexts.[38,136,137] Furthermore, it is unknown how uORFs might contribute to protein synthesis regulation during neuronal differentiation. In this study, we identified 27.6% of mRNA transcripts contain one or more uORFs using the translational classification algorithm SPECtre and stringent heuristic filtering. 32.6% of the identified uORFs are predicted to use an AUG initiation site, with the near-cognate codon CUG being the next most abundantly used TIS. uORFs with AUG and near-cognate start codons can be experimentally validated, demonstrating a biologically founded algorithm allows for reliable annotation of these non-canonical events. We found that both overlapping and non-overlapping uORFs are conserved, and have increased GC content. Interestingly, uORF translation shifts between the non-differentiated and RA-differentiated cell types; this is important because our data show that uORFs exert a repressive

effect on downstream protein synthesis based on the proximity of their termination relative to the annotated CDS start site. Our work expands on established observations by building a set of principles which may be used to determine the translatability of a 5' leader, and its potential for affecting canonical protein translation, while detailing all of the translational changes associated with RA differentiation of SH-SY5Y cells.[129,138-140]

3.3 Materials and Methods

SH-SY5Y cell maintenance and differentiation

SH-SY5Y cells were grown in DMEM:F12 media (Invitrogen) supplemented with 10% FBS, 0.01 mg/mL Gentamicin and 0.25 µg/mL Amphotericin B. Cells were plated on 150 mm plates that were either coated with 0.1 mg/mL poly-D lysine (Millipore) for differentiation or uncoated. Cells were allowed to propagate to 80% confluency for 1-2 days prior to lysing for ribosome profiling or induction of differentiation. SH-SY5Y cells were differentiated for 6 days in 10 µM retinoic acid (all-trans, Sigma), with media changed every 24 hours prior to lysing.

Construction and next-generation sequencing of ribosome profiling libraries

Cells were washed with ice cold PBS with CHX at 100 µg/mL, flash frozen in liquid nitrogen, and lysed on ice (in the presence of CHX) to prevent ribosome loading and runoff. Additional lysates were processed in parallel for poly(A) mRNA purification and library preparation using the TruSeq Library Preparation Kit (Illumina). Polysomes were isolated from the ribosome foot printing lysates on a 1 M sucrose gradient with high speed centrifugation using a 70.1 Ti rotor (Beckman)

at 55,000 revolutions per minute for 4 hours at 4°C. RNA footprints were processed according to previously described methods.[89] rRNA was eliminated prior to linker ligation using Ribo-Zero Gold rRNA Removal Kit (Illumina). Ribosome profiling libraries were barcoded and multiplexed with 2-4 libraries per lane, and sequenced on a HiSeq 2000 (Illumina) using 50 cycles of single end reads. mRNA libraries were multiplexed on a single lane. All sequencing was conducted at the University of Michigan DNA Sequencing Core.

Plasmid construction

pcDNA 3.1 plasmid was modified to encode NanoLuc and GGG-NanoLuc as previously published [CITE]. gBlocks® (IDT) were ordered of the 5' leader sequence to the last codon before the in-frame stop of selected genes flanked by restriction sites. These were restriction cloned upstream of GGG-nLuc using PacI and XhoI (NEB) with 12 nucleotides between the start of the 5' leader and the T7 promoter sequence to reduce spurious initiation in sequences specific to the plasmid. Restriction digest and Sanger sequencing were used to confirm plasmid sequence. Additional reporters were cloned so that the NanoLuc tag was shifted out of frame with the predicted ORF and the CDS start site (if present in the reporter).

SH-SY5Y transfection and nanoluciferase assay

SH-SY5Y cells were seeded on 6-well culture plates at 3x105 cells per well. 24 hours post seeding, each well was transfected using 7.5 μL FUGENE HD (Promega) and 1.25 μg nanoluciferase reporter plasmid along with pGL4.13 (internal transfection control that encodes firefly luciferase) were added at the same concentration in 300 μL of OptiMEM (Invitrogen). Transfections of

differentiated cells were performed on day 5 in RA supplemented media. Cultures were allowed to grow for 24 hours after transfection. Cells were lysed in 250μL Glo Lysis Buffer (Promega) for 5 minutes at room temperature. 50 μL lysate was mixed with 50 μL prepared Nano-GLO or ONE-Glo (Promega) for 2 minutes, and bioluminescence was detected using a GloMax®96 Microplate Luminometer. Nanoluciferase signal was normalized to FFluc signal in each sample. Two pcDNA vectors encoding nanoLuc and nanoLuc with the AUG start codon mutated to a GGG (GGG-nanoLuc) were run in parallel with each experimental nanoluciferase plasmid and subjected to both conditions to serve as a control for normalization.

Immunocytochemistry and microscopy

Cells were fixed at 37°C with 4% PFA/4% sucrose in PBS with 1 mM $MgCl_2$ and 0.1 mM $CaCl_2$ (PBS-MC), permeabilized for 5 minutes in 0.1% Triton-X in PBS-MC, and blocked for 1 hour with 5% bovine serum albumin in PBS-MC. Cells were incubated in blocking buffer and primary antibodies against β-actin (Santa Cruz Biotechnology, 1:1000) and neurofilament (supplier, 1:1000) for 1 hour at room temperature. Following 3x10 minute washes in PBS-MC, cells were incubated in PBS-MC with Alexa Fluor 488 conjugated goat anti-rabbit IgG and Alexa Flour 635 conjugated goat anti-mouse IgG to achieve secondary detection (Thermo Fisher, 1:1000). Cells were washed again, and placed in ProLong TM Gold antifade reagent with DAPI (Invitrogen). Imaging was performed on an inverted Olympus FV1000 laser-scanning confocal microscope using a 40X objective with a 1X digital zoom. Acquisition parameters were identical for each condition and optimized to eliminate signal bleed-through between channels. Images were converted to maximal-intensity z-projections in ImageJ. Cytoplasmic β-actin was quantified by averaging the integrated density corrected for background signal of the cells in each condition. The

length of one main neurofilament labeled primary neurite per cell was determined in ImageJ and converted from pixels to μm, and averaged for each condition.

Western blotting

Cells were maintained as described above. Cells were washed 2X in PBS, and RIPA buffer was added to a single well of a 12-well dish either at 80% confluency or after 6 days of retinoic acid differentiation. Cells were agitated for 40 minutes at 4°C to ensure complete lysis. Lysates were clarified by centrifugation at 15x103 RPM for 20 minutes at 4°C. The supernatant was mixed with reducing SDS sample buffer and boiled for 5 minutes at 90°C. Equal amounts of lysate were loaded on an 8% SDS-PAGE gel and subsequent western blotting was performed with primary antibodies against FMRP (mouse, 1:1000, 6B8, BioLegend) and GAPDH (mouse, 1:1000, Santa Cruz Biotechnology)—both in 5% (wt/vol) non-fat dry milk in TBS-T (NFDM). An HRP conjugated goat antibody to mouse IgG was used for secondary detection (1:5000, Jackson ImmunoResearch Laboratories) in 5% NFDM.

A 12% SDS-PAGE gel was used to resolve eIF2α and ATF4. Antibody for phosphorylated-eIF2α (rabbit, Thermo Fisher) was used at 1:500, after secondary detection blots were stripped in a low pH glycine buffer and re-probed with antibody against total eIF2α (rabbit, Cell Signaling Technology, 1:1000) and E7 tubulin (mouse, DSHB,1:1000). Blots were stripped again and probed for ATF4 (rabbit, Cell Signaling Technology, 1:1000). Secondary detection was achieved using HRP conjugated goat antibodies to rabbit IgG or to mouse IgG in 5% NFDM (1:5000, Jackson ImmunoResearch Laboratories).

Pre-processing and alignment of mRNA and ribosome profiling sequence libraries

Ribosome profiling and mRNA-Seq reads were trimmed of adapters, and then by quality using *fastq-mcf* from the *ea-utils* package (https://github.com/ExpressionAnalysis/ea-utils). Ribosome profiling and mRNA-Seq reads in FASTQ format were trimmed based on quality if four consecutive nucleotides were observed with Phred scores of 10 or below. The minimum read length required after trimming was 25 nucleotides. Trimmed sequences were then aligned to a ribosomal RNA sequence index using Bowtie v1.1.2 to deplete them of contaminant sequences.[62] Alignment to the rRNA sequence contaminant index was performed using the following parameters: seed alignment length of 22 nucleotides, no mismatches in the seed alignment were allowed, with the unmapped reads written to a separate FASTQ file.

Calculation of translation efficiency

Ribosome profiling or mRNA-Seq reads were counted over each region (5'UTR, CDS, and 3'UTR), transcript, or upstream-initiated ORF and then normalized to length and library size as transcripts per million.[68] To calculate translational efficiency over a region, transcript or upstream-initiated ORF, ribosome profiling TPM in each biological replicate across each condition was quantile normalized and then divided by the quantile normalized TPM in mRNA-Seq.[141] Read and RPF counts from mRNA-Seq and ribosome profiling libraries does not include those that overlap the 5'UTR and 3'UTR. Furthermore, to limit the boundary effects due to translation initiation and termination, RPF and read counts do not include those whose A- or P-site adjusted

position for harringtonine and cycloheximide libraries, respectively, overlap the first or last 15 nucleotides of an annotated CDS.

Differential expression analysis and gene set enrichment testing in mRNA-Seq

As described previously, the read abundance over annotated protein-coding transcripts was calculated as TPM, then quantile normalized across conditions using the preprocessCore package in R, and then ranked.[142] The change in rank for each gene was calculated across the non-differentiated and RA-differentiated conditions, and the significance of the up- or down-regulation of these rang-changes across conditions was classified using an extreme outlier cutoff.[143] Functional characterization of these significantly rank-changed genes across the non-differentiated and RA-differentiated conditions was analyzed using the goseq package in R, and corrected for multiple testing using Benjamini-Hochberg adjusted p-values.[144,145]

Differential translation analysis and gene set enrichment testing in ribosome profiling

Ribosome profiling read fragments were A- or P-site adjusted, and then counted over annotated protein-coding CDS regions in each biological replicate using the metagene profiles generated by Plastid.[93] As described previously, ribosome-protected fragments with A- or P-site adjusted positions that overlapped the first or last 15 nucleotides of the boundaries defined by the annotated CDS region were masked from the analysis. DESeq2 was used to identify those genes with differential translation across the two states of cellular differentiation.[71] Genes were annotated as significantly up- or down-regulated using a Benjamini-Hochberg adjusted p-value cutoff of 0.1, and fold-change in counts greater than 1, or less than 1, respectively. Functional characterization

of these significantly up- and down-regulated genes was analyzed by goseq using parameters specified previously.

Differential translation efficiency and gene set enrichment testing in ribosome profiling

For each biological replicate, ribosome profiling read fragments were A- or P-site adjusted, and then counted over annotated protein-coding CDS regions using the metagene profiles generated by Plastid. As above, read counts over the first and last 15 nucleotides of protein-coding CDS regions were masked for subsequent analyses. In addition, mRNA-Seq read counts were extracted from each condition, with the proximal and terminal 15 nucleotide ends of the CDS masked for consistency with the RPF counts. The DESeq2 wrapper for differential translational efficiency analysis, was used to identify those genes with significant changes in translational efficiency.[146] Genes were annotated as significantly up- or down-regulated using a Benjamini-Hochberg adjusted p-value cutoff of 0.1, and absolute fold-change of 1. Functional characterization of the sets of genes enriched in each condition by translational efficiency was analyzed by goseq using parameters described previously.

Percent change in transcript abundance

In addition to significance cutoffs for up- and down-regulated regions or transcripts based on mRNA-Seq and ribosome profiling sequence alignments, global changes were assessed by percent change in abundance. TPM based on mRNA-Seq and RPF alignments were calculated, quantile normalized, and then compared by percent change in abundance across the non-differentiated and

RA-differentiated libraries. The number of genes with a change in TPM across the two conditions was evaluated at pre-defined percent-change cutoffs of 10%, 20%, 30%, 40% and 50%.

Conservation and GC nucleotide content analysis

To assess the conservation of the various regions, transcripts and ORFs, the PhyloCSF over each target region was extracted.[147] For upstream-initiated ORFs, the PhyloCSF score was extracted according to its predicted phase. In order to de-convolute the contribution of regional conservation due to overlap with annotated CDS regions, predicted ORFs that did not initiate and terminate wholly upstream of a CDS were also scored according to the subset of their coordinates defined by the 5'UTR alone. The mean PhyloCSF over each of these regions and ORFs was calculated, and then mean-shifted to canonical (+0) reading frame of the annotated CDS for comparison. The G/C nucleotide content of each region was calculated as the number of G and C nucleotides in a given region divided by its length. The G/C nucleotide content was calculated for annotated 5'UTRs, CDS, 3'UTRs and over each predicted ORF; the G/C content for the 5'UTR portion of ORFs predicted to terminate in the CDS was calculated separately.

Cluster analysis of differential uORF translation by SPECtre score

In order to identify subsets of upstream-initiated ORFs with differential translation in one state of cell differentiation compared to the other, the SPECtre score for each predicted ORF was calculated (see Appendix B). The SPECtre score of each predicted ORF was classified by k-means clustering in R to define sets of uORFs with differential translation in one of the conditions, and those with no difference in translational potential between the two conditions.

Kernel density estimation of differential uORF regulation on CDS translation efficiency

To further differentiate those uORFs with differential translation and identify those that contribute to the regulation of downstream CDS, the log-change in predicted ORF TPM was compared against the log-change in downstream CDS TPM across the conditions. The differential translational identity of each predicted ORF was retained from the SPECtre clustering analysis, and kernel density estimation was performed using R.

Multiple regression analysis of uORF regulation on CDS translation efficiency

In order to assess the global contribution of uORFs and oORFs on the change in translational efficiency of the downstream CDS, a multiple regression model was built using: the change in abundance (in TPM) over each predicted ORF across conditions (*delta_tpm*), the GC content over each predicted ORF restricted to the portion of each ORF overlapping the 5'UTR (*gc*), the PhyloCSF conservation score over the predicted ORF restricted to the part of each ORF overlapping the 5'UTR (*cons*), the proximity of the termination site defined by each ORF as its absolute distance, in nucleotides, to the annotated CDS translation initiation site (*dist*), and the binary classification of the predicted translation initiation site, as AUG-initiated or non-AUG-initiated, for each ORF (*tis*). Each of the above parameters was input as a prediction variable, with the change in translational efficiency of the CDS across the RA-differentiated and non-differentiated states as the outcome variable, using the linear model function in R. Pairwise interactions between the prediction variables were also tested.

Gene set enrichment testing of multiple regression negative residuals

Residuals for genes tested in the multiple regression model were extracted and ranked according to their magnitude; genes were extracted based on their rank in the top 5th, the top 10th, and the top 25th percentile of negative residuals. These residual sets were then tested for gene set enrichment using the goseq package in R, and the p-values were corrected for multiple testing (Benjamini-Hochberg). The union of all gene sets across the three percentile residual sets tested was compiled, and the adjusted p-values in each test was tabulated for visualization.

Regression analysis of uORF proximity on CDS translation efficiency

To investigate the relationship between the proximity of upstream-initiated ORFs and the translational efficiency of the downstream CDS, a linear regression model was built using the SPECtre score of the predicted ORF as the prediction variable and the translational efficiency of the CDS as the dependent variable. To assess the relative contribution of proximity, predicted ORFs were binned according to their termination position relative to the annotated CDS translation initiation site. Predicted ORFs were binned every 30 nucleotides based on their termination position relative to the annotated CDS start site, with maximum bins limited to 300 nucleotides upstream, and 600 nucleotides downstream of the annotated CDS translation initiation sequence position. Based on the predicted ORFs terminated in each bin, the regression coefficient in each biological replicate over each state of differentiation was calculated.

3.4 Results

Molecular and bioinformatic validation of retinoic acid induced differentiation of SH-SY5Y cells

SH-SY5Y neuroblastoma cells were induced to undergo neuron-like differentiation by treatment with retinoic acid (Figure 3.1A); efficacy of RA-induced differentiation was assessed by immunocytochemistry staining using neuron-specific markers. Neurofilament staining is more prominent in RA-differentiated SH-SY5Y cells (Figure 3.1B, left) than in non-differentiated cells. In addition, each main neurite was selected and analyzed for neurofilament staining and length; neurites in RA-differentiated cells are longer than in non-differentiated cells (Figure 3.1D). Furthermore, beta-actin staining is more diffuse in non-differentiated cells compared to the punctate staining observed in RA-differentiated cells (Figure 3.1A, left middle). Individual cell fluorescence measurements confirmed the higher staining of non-differentiated cells by beta-actin (Figure 3.1D). Finally, the post-synaptic marker FMRP was detected by Western blot, and enriched in RA-differentiated cells compared to non-differentiated cells (Figure 3.1E and F). In sum, the RA differentiated cells displayed characteristics phenotypically consistent with post-mitotic neuron-like cells.

Differential expression analysis and gene set enrichment testing of sequenced mRNA transcript abundance was used to further confirm the neuron-like differentiation of SH-SY5Y cells by retinoic acid. Based on the rank-change in abundance, calculated as TPM, across the two cell conditions (Figure 3.1G), significantly up- and down-regulated gene sets were identified in the RA-differentiated mRNA-Seq libraries. Significantly up-regulated gene sets in the RA-differentiated condition included those terms related to cell communication, signaling and stimulus

response (Figure 3.1H). In contrast, significantly down-regulated gene sets in the RA-differentiated state included those related to mitosis, cellular division and regulation of the cell cycle (Figure 3.1I). Enriched terms related to molecular function and cellular components based on rank-change analysis of mRNA abundance are shown in Supplemental Figures B.1 and B.2, respectively. Taken together, experimental validation and bioinformatic analysis confirms the retinoic acid induction of SH-SY5Y cells into a differentiated state phenotypically consistent with neuron-like cells.

Integrative analysis of mRNA and ribosome profiling sequence data was used to further characterize the RA-differentiated and non-differentiated SH-SY5Y cells. Ribosome-protected fragments over each transcript were counted and differential abundance was assessed by DESeq2 (Figure 3.2A). Significantly up- and down-regulated genes were identified, and gene set enrichment testing was performed (Supplemental Figures B.3 and B.4). In order to account for the influence of transcriptional regulation on translational abundance, we performed an additional set of differential analyses using translational efficiency (Figure 3.2B). Significantly up-regulated gene sets identified by differential translational efficiency analysis included those related to protein targeting and localization (Figure 3.2C). Enriched terms related to molecular function and cellular components based on differential translational efficiency analysis are shown in Supplemental Figures B.5 and B6, respectively. A representative transcript with differential translational efficiency across the two conditions is the axon guidance gene PLXNA2 (Figure 3.2D); PLXNA2 is abundantly transcribed in both conditions (top and bottom, gray bars), but limited in the abundance of RPFs in the non-differentiated condition (top) compared to the RA differentiated state (bottom). Similar to the differential mRNA analysis, significantly down-regulated gene sets

identified by differential translational efficiency analysis (Figure 3.2E) include those related to mitosis, cell division, and cell cycle control. The gene TFPI2 is less efficiently translated in the RA-differentiated state (bottom) compared to the non-differentiated cell condition (top). Globally, we observe shifts in transcription and translation across the two cell conditions in a majority of transcripts profiled by both mRNA-Seq and ribosome profiling (Figure 3.2G). Based on these observations, we find global changes in transcription and translational efficiency consistent with RA-induced differentiation of SH-SY5Y cells in genes related to signaling, regulation of the cell cycle, including subsets of genes related to neuronal modeling.

uORF prediction and filtering statistics

uORFs were computationally predicted from annotated 5'UTRs of protein-coding transcripts; translation initiation was permitted to utilize both AUG and non-AUG near-cognate sites, where near-cognate sites differ from the canonical AUG initiation site sequence by one nucleotide. RPF coverage over the predicted uORFs was normalized, then scored by SPECtre for translational potential, and then filtered according to a series of heuristics (Figure 3.3). These included minimum RPF coverage in the 5' leader region of the predicted uORF, removal of in-frame N-terminal extensions, and minimum mRNA-Seq coverage in the annotated CDS region. Minimum mRNA-Seq coverage in the CDS was required for subsequent translational efficiency comparisons across the two cell conditions. In addition, redundant isoforms were converged, and overlapping ORFs were prioritized based on optimal RPF coverage and SPECtre score.

Based on the filtered set of uORFs, we identified roughly 3,100 genes with at least one predicted uORF (Figure 3.4A); of these, approximately 360 of these genes harbored both an overlapping ORF and an ORF terminated upstream in the 5'UTR (Figure 3.4B). The filtered set of predicted uORFs initiate proximally to the annotated protein-coding translational start site (Figure 3.4B), and tend to be short in length (Figure 3.4C). We next examined the sequence characteristics of the predicted uORFs through measurement of their G/C nucleotide content. For comparison, we found that annotated 5'UTRs were significantly enriched in G/C content over both CDS regions, and 3'UTRs (Figure 3.4D). Furthermore, annotated 5'UTRs with a predicted uORF were significantly higher in G/C content than 5'UTRs without a predicted uORF; this is suggestive of a sequence-based context in the 5'UTR for the efficient translation of uORFs (Figure 3.4E). However, upstream-terminated ORFs and the full sequence of oORFs had G/C nucleotide content that resembled 5'UTRs that were not predicted to have uORF (Figure 3.4F). Interestingly, when the CDS region of these overlapping ORFs were partitioned and the portion of these uORFs contained in the 5'UTR were measured, their G/C content resembled that of the 5'UTRs predicted to have a uORF. Although uORFs tend to resemble the 5'UTR sequences from which they are derived, these results are not immediately suggestive of a sequence-level difference that distinguishes them from their flanking 5'UTR sequence context.

Translation initiation site annotation

Given the preference for translation initiation using AUG sites, the translation initiation sites of predicted ORFs were re-annotated based on the proximity or presence of an in-frame AUG site. Based on this re-annotation, AUG and putative AUG initiation sites were the most commonly used start codons in the predicted ORFs (Figure 3.5A). Approximately 32.6% of predicted uORFs

utilized an AUG start codon, with CUG the next most common site used (16.2%). These results support previous reports of the preferential initiation of translation from AUG codons. Moreover, CUG is identified as the next most common start site, which is consistent with previous reports, and biochemically-validated models that suggest translation initiation from CUG sites are second in efficiency only to AUG start codons.[109,148]

Validation of uORF translation

A subset of transcripts with predicted uORFs were selected for experimental validation, including HAND2 and ARF4 (Figure 3.5B and C). Initial assessment of the translatability and regulatory impact of predicted ORFs was done using a nano-Luciferase reporter construct (Figure 3.5D). Predicted uORF sequences were inserted in-frame to an nLuc reporter with the AUG translation initiation site converted to a GGG; translation of the nLuc reporter would be driven by the sequence context provided by the 5' leader. In addition to HAND2 and ARF4, the set of genes selected for validation included LAMB1, TSC1, and the leptin receptor gene, LEPR (Figure 3.5E). Two 5' leader sequences from genes not thought to have highly translated uORFs were included for analysis as well (Figure 3.5F, black). Compared to the 5' leader-less control construct, the 5' leader sequence of all candidates tested with highly translated predicted uORFs (Figure 3.5E) drove expression of the nLuc reporter (Figure 3.5F, green). In addition, Western blotting confirmed the presence of highly translated products from the predicted uORF in LEPR, PCBD, QDPR and ATP5I (Figure 3.5G). Finally, we assessed the frame specificity of the predicted uORFs by shifting the nLuc reporter out-of-frame of the 5' leader (Figure 3.5H). Taken together, these results validate the translation of predicted uORF leader sequences in an *in vivo* system.

Global landscape of uORF translation on protein synthesis regulation

To investigate the potential regulatory impact of uORF translation on a global scale within the context of neuronal differentiation, we identified subsets of genes with differential translation in the non-differentiated and RA-differentiated conditions (Figure 3.6A). SPECtre scores over the CDS of protein-coding transcripts were clustered into groups that were enriched in non-differentiated cells (gold), enriched in the RA-differentiated cells (cyan), or similarly translated in both states (gray). Next, we coupled these translated CDS to their predicted uORFs (Figure 3.6B) and observed no defined relationship between uORF translation and CDS regulation. As denoted by the density of transcripts both above and below the line, uORF translation alone was not predictive of downstream CDS translation. Annotation of a predicted uORF in the 5' leader of an mRNA was also not predictive of CDS regulation in both non-differentiated and RA-differentiated specific transcripts (Figure 3.6E, F and G). However, we do observe localized examples where uORF translation is sufficient to negatively impact the translation of the downstream protein (Figure 3.6H and I). Although the mere presence of a uORF is not predictive of translational repression, we identified sets of translationally repressed genes with predicted uORFs in both non-differentiated and RA-differentiated cells. Furthermore, we identified and experimentally validated subsets of these uORFs that conditionally repress synthesis of the downstream protein-coding gene in a state-dependent manner, including the leptin receptor gene LEPR. The putative uORF-mediated translational regulation of the leptin receptor gene in neuronal cells invites further scrutiny. Leptin is a hormone related to signaling of appetite satiety; the leptin receptor is involved in the regulation of glucose homeostasis through neuronal inhibition in the parabrachial nucleus.[149-151]

Sequence and proximity as contexts for uORF regulation

Since prediction of a uORF and its translation were not predictive of protein synthesis regulation, we explored other sequence and physical components of the predicted uORFs that might be predictive of their regulatory impact. To this end, we built a multiple regression model based on the GC content of the 5' leader region of predicted ORFs, their conservation, change in translational efficiency over the two conditions, TIS sequence, and the proximity of the termination site of the uORF to the annotated CDS (Figure 3.7A). We observed that proximity of termination to the annotated CDS, GC content, change in uORF translation, and conservation were all significant predictors of downstream CDS translation whereas the identity of the predicted TIS had little predictive power (Figure 3.7A and D). Furthermore, we extracted the set of genes that most negatively influence the regression model and found that they were enriched for biological processes related to cell cycle regulation, cell division and chromosome structure (Figure 3.7B). The full cohort of genes, with positive and negative residuals, is listed in Supplemental Table B.2. We decided to examine the impact of termination proximity further and built proximity-dependent regression models using the SPECtre score of the predicted uORFs. Based on this model, we found that termination proximity, in particular those uORFs that terminated inside the annotated CDS region, had profound effect on the downstream translation of the canonically-encoded protein (Figure 3.7C). Taken together, these results provide additional support for a steric hindrance model for translational regulation of protein synthesis by upstream open-reading frames.

3.5 Discussion

Comprehensive characterization of the diverse network of regulatory control that mediates protein synthesis remains elusive; however, advances in the integrative analysis of multi-scalar next-generation sequencing data are driving increasingly sensitive insight and dissection of these regulatory factors that control the efficiency of protein synthesis. Translation of upstream-initiated open-reading frames has been well studied in various systems and organisms, and much effort has been made to catalog and characterize the prevalence and impact of these ORFs.[152] In this study, we investigated the landscape of uORF translation in a cellular model of neuronal differentiation and examined the significance of their regulatory potential on the synthesis of proteins relevant to this change in cell state. We found that upon the retinoic-acid induced conversion of SH-SY5Y cells into a terminally differentiated state, pathways relevant to signaling and neuronal function were up-regulated, and gene networks related to cell cycle regulation and division were down-regulated. This was confirmed experimentally through immunocytochemistry, and across multiple scales of next-generation sequencing data; changes in gene set enrichment related to the terminal differentiation of SH-SY5Y cells were consistent through transcriptomic and translational profiling.

We established a rigorous framework for the computational prediction of uORFs, and characterized their potential significance in the regulation of cell differentiation. Like previous studies, we found that predicted uORFs were concomitantly less conserved than protein-coding CDS regions, but more highly conserved than the background regional context of 5’ untranslated regions. A limited subset of protein-coding genes with translational evidence harbored a predicted ORF; approximately 31% of annotated protein-coding transcripts were predicted to have a

potentially translated uORF. Similar to earlier studies, most of these uORFs were predicted to initiate translation from canonical AUG start sites, with CUG translation initiation sites being the next most commonly utilized sequence moiety. To further disentangle the potential role of these uORFs in the regulation of canonical protein synthesis, we characterized global changes in the translational activity of protein-coding genes; we found subsets of genes that were differentially regulated corresponding to cellular state using spectral classification and transcript abundance metrics.

Initial efforts to classify the regulatory significance of predicted ORFs based on their translational activity alone were inconclusive; although we validated the translational activity of specific 5' leader sequences, only a limited subset of these were experimentally observed to regulate their downstream protein-coding CDS as computationally predicted. Based on these results, we examined additional factors that might more accurately predict the regulatory potential of these uORFs. We built a multiple regression model that accounted for the translational activity of the predicted uORFs, as well as their conservation status and sequence context. Furthermore, we accounted for their predicted initiation site identity, and their spatial proximity to the annotated protein-coding start site. Based on this multiple regression model, we found that all of these factors except initiation site identity were significant predictors of changes in the translational efficiency of the annotated protein-coding CDS. Among the genes that were most translationally suppressed upon terminal differentiation of SH-SY5Y cells were those related to regulation of the cell cycle, cell division, and chromosome organization. Indeed, we found that the spatial proximity of the predicted uORFs contributed significantly to their translational regulation of protein-coding CDS.

Taken together, we find that a subset of uORFs negatively regulate the translation of their downstream protein-coding CDS; furthermore, the conservation of these ORFs, their GC nucleotide content, translational activity, and their spatial proximity to the annotated CDS contribute significantly to their regulatory potential. Translation of many of these uORFs results in the suppression of genes related to the cell cycle and cellular division; pathways that are down-regulated upon retinoic acid differentiation of SH-SY5Y cells. These results contribute additional evidence for the importance of uORF translation in the control of protein synthesis, and suggest additional factors that may mediate their regulatory potential within the context of a cellular model of neuronal differentiation.

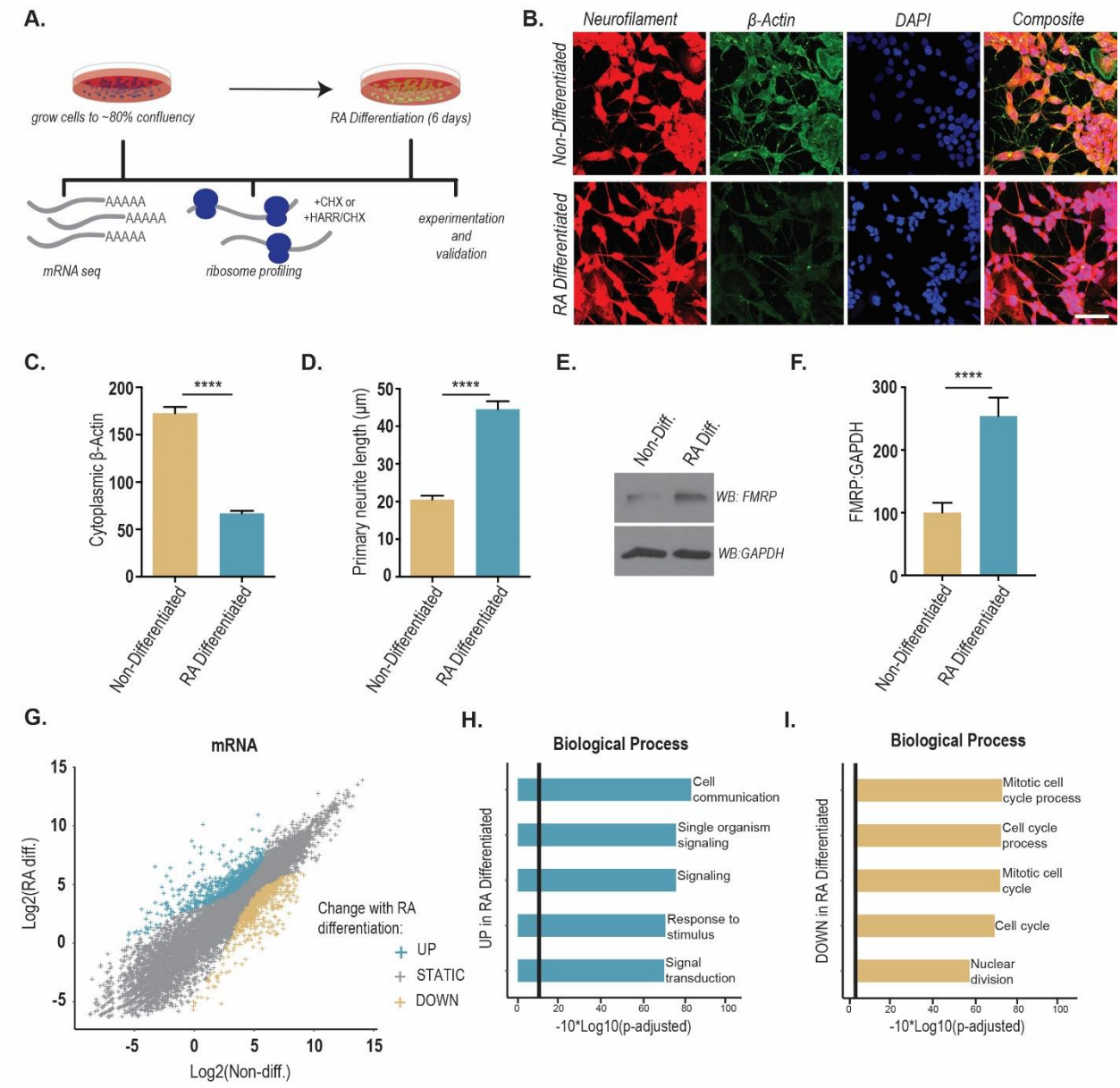

Figure 3.1 Retinoic acid treatment induces neuronal differentiation of SH-SY5Y cells. A) Schematic of the experimental design and the basic workflow for data acquisition. B) Immunocytochemistry performed on non-differentiated and RA-differentiated SH-SY5Y cells confirmed the shift to a neuronal phenotype with RA treatment. Cells of both conditions were fixed and stained with antibodies against neurofilament (red), and beta-actin (green), and nuclei were DAPI stained (blue). C) Individual cell fluorescence was quantified and represented as corrected total cell fluorescence for beta-actin; n=119 for non-differentiated and n=118 for RA-differentiated. D) Each main neurite was selected and analyzed for neurofilament staining and

length; n=109 for non-differentiated and n=100 for RA-differentiated. E) Western blotting of both cell conditions showed an increase in the post-synaptic marker FMRP in the RA-differentiated cells, and quantified in F); n=4 for both conditions. For panels C), D), and F) Student's T-test p-value<=0.0001. Graphs represent mean ± S.E.M. G) Differential mRNA abundance based on non-differentiated versus RA-differentiated transcripts per million. Transcripts were defined as significantly up-regulated (cyan) or down-regulated (gold) in the RA-differentiated cell condition based on rank-change in TPM compared to the non-differentiated condition. H) Significantly enriched up-regulated gene sets in RA-differentiated mRNAs based on Benjamini-Hochberg multiple testing corrected p-values. Black vertical line in panels H) and I) denotes a corrected p-value cutoff of 0.05. I) Significantly enriched down-regulated gene sets in RA-differentiated mRNAs based on Benjamini-Hochberg multiple testing corrected p-values.

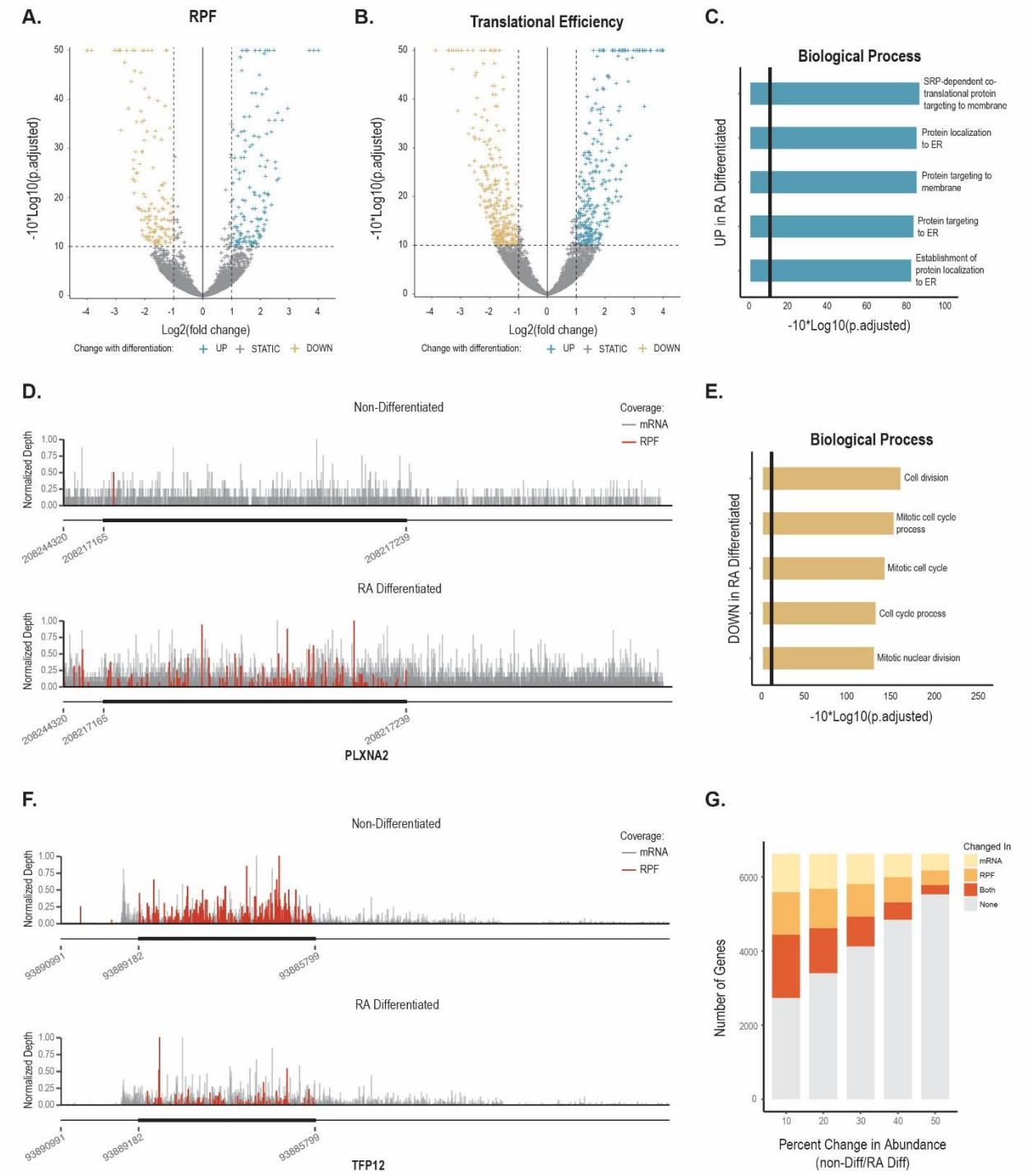

Figure 3.2 Differential translation and translational efficiency in SH-SY5Y cells. A) Volcano plot of transcripts with differential translation by ribosome-protected fragment counts in non-differentiated and RA-differentiated cells. Significantly up-regulated (cyan) and down-regulated (gold) genes in RA-differentiated cells are defined by a log2-normalized fold-change cutoff of ±1

(vertical lines), and a multiple testing corrected p-value cutoff of 0.1 (horizontal line). B) Volcano plot of transcripts with differential translational efficiency by Riborex analysis in non-differentiated and RA-differentiated cells. Significantly up-regulated (cyan) and down-regulated (gold) genes in RA-differentiated cells are defined by a log2-normalized fold-change cutoff of ±1 (vertical lines), and a multiple testing corrected p-value cutoff of 0.1 (horizontal line). C) Gene sets with significantly up-regulated translational efficiency in RA-differentiated cells. The top five biological processes with significant enrichment using a multiple testing p-value cutoff of 0.05 (vertical line) include those terms related to cell division, and the regulation of the cell cycle. D) Normalized mRNA (gray) and RPF coverage (red) over the 5'UTR (thin line, left), CDS (thick line, middle), and 3'UTR (thin line, right) of the axon guidance gene, PLXNA2, is representative of a transcript with higher translational efficiency in the RA-differentiated cell condition. E) Gene sets with significantly down-regulated translational efficiency in RA-differentiated cells. The top five biological processes, using a multiple testing p-value cutoff of 0.05 (vertical line) include terms related to protein target, and localization. F) Transcript coverage plot of TFP12 demonstrates higher coverage by RPFs (red) relative to mRNA (gray) in the non-differentiated cell condition. G) Genes that change in RPF abundance as measured by mRNA-Seq and ribosome profiling. Pre-defined cutoffs in percent change of TPM across the two cell conditions was evaluated by mRNA and RPF abundance, and the number of genes that changed as measured by one, both, or neither was counted.

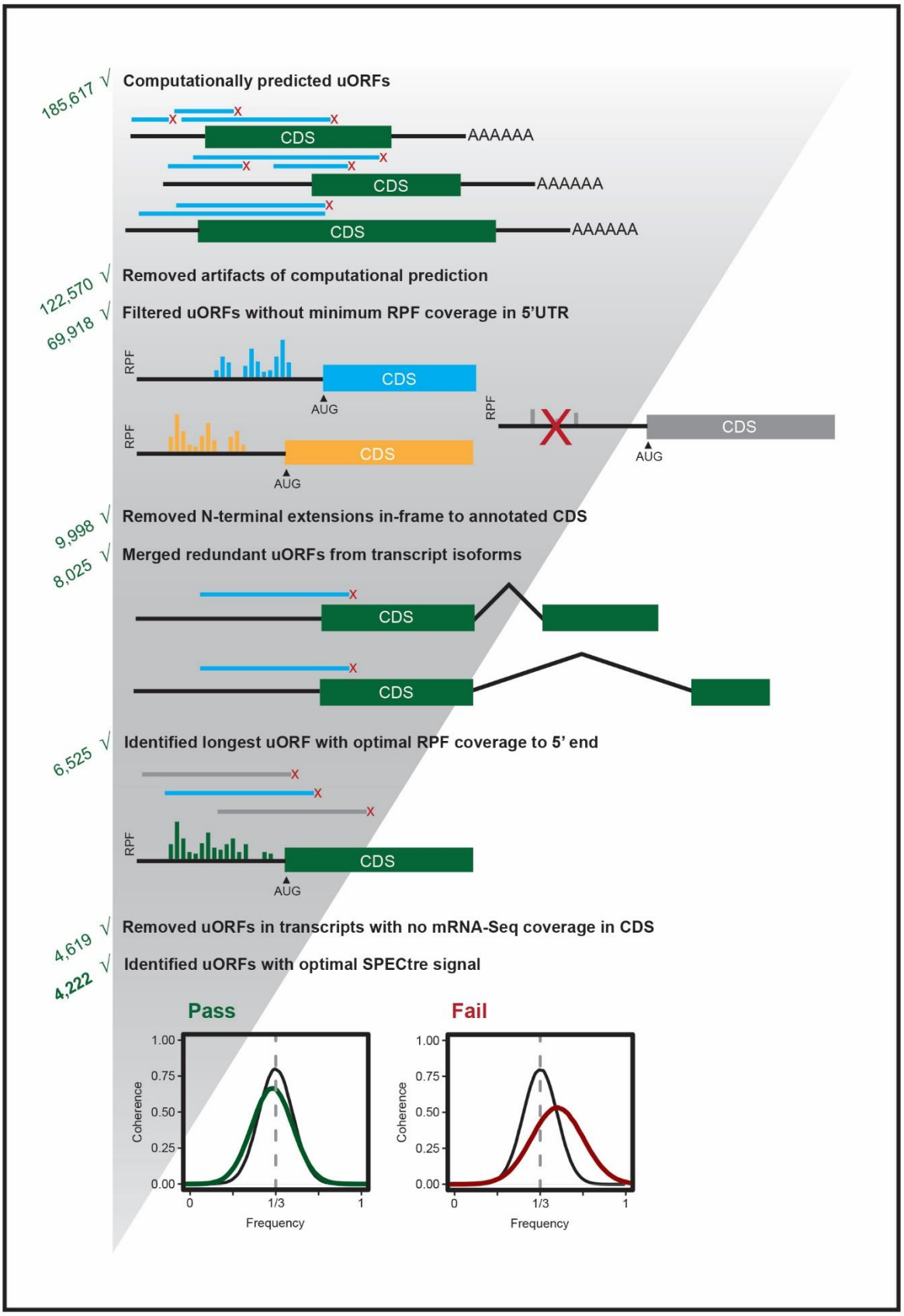
185,617 √
Computationally predicted uORFs
CDS
AAAAAA
CDS
AAAAAA
CDS
AAAAAA
122,570 √
Removed artifacts of computational prediction
69,918 √
Filtered uORFs without minimum RPF coverage in 5'UTR
RPF
CDS
AUG
RPF
CDS
AUG
RPF
CDS
AUG
9,998 √
Removed N-terminal extensions in-frame to annotated CDS
8,025 √
Merged redundant uORFs from transcript isoforms
CDS
CDS
6,525 √
Identified longest uORF with optimal RPF coverage to 5' end
RPF
CDS
AUG
4,619 √
Removed uORFs in transcripts with no mRNA-Seq coverage in CDS
4,222 √
Identified uORFs with optimal SPECtre signal
Pass
Fail
Coherence
1.00
0.75
0.50
0.25
0.00
0
1/3
1
Frequency
Coherence
1.00
0.75
0.50
0.25
0.00
0
1/3
1
Frequency

Figure 3.3 Computational prediction and filtering of ORFs. ORFs were predicted from AUG and non-AUG near-cognate translation initiation sites in the 5'UTR of annotated protein-coding genes, and computationally extended to the first in-frame termination site encountered in the 5'UTR (upstream-terminated ORFs) or CDS (overlapping ORFs). Predicted ORFs were then filtered according to a series of heuristic filters including: minimum RPF coverage in the 5'UTR, 2) in-frame N-terminal extensions, 3) redundant isoforms, 4) minimum length, 5) sufficient SPECtre signal, and 6) those CDS with no mRNA-Seq coverage to be tested for translational efficiency.

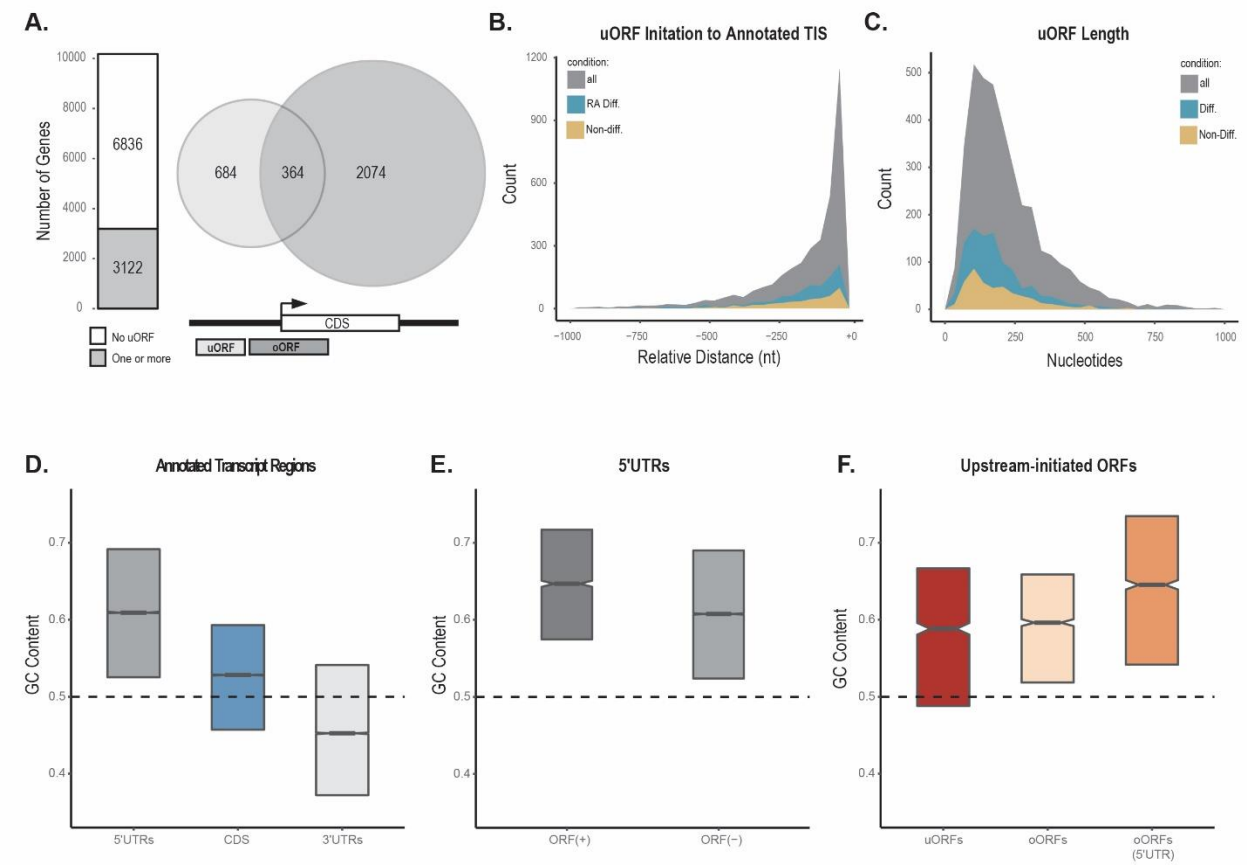

Figure 3.4 Characterization of predicted ORFs. A) The number of genes with at least one predicted ORF (bar plot) in the 5'UTR of evaluated protein-coding genes. Breakdown of the number of genes with a predicted ORF terminated upstream in the 5'UTR only (left circle), terminated in the CDS only (right circle), or with both a predicted upstream-terminated and CDS-overlapping ORF (overlap). B) Distribution of predicted ORF translation initiation relative to the annotated protein-coding CDS start site in non-differentiated specific genes (gold), RA-differentiated specific genes (cyan), and in aggregate (gray). Distribution of predicted ORF lengths in non-differentiated specific genes (gold), RA-differentiated specific genes (cyan), and in aggregate (gray). D) GC nucleotide content in annotated transcript 5'UTRs (dark gray), CDS (blue), and 3'UTRs (light gray). E) GC nucleotide in annotated 5'UTRs with a predicted ORF (dark gray), and those 5'UTRs without a predicted ORF (light gray). F) GC nucleotide content of predicted upstream-terminated ORFs (red), overlapping CDS-terminated ORFs (light orange), and the 5'UTR specific portion of predicted CDS-terminated ORFs (dark orange).

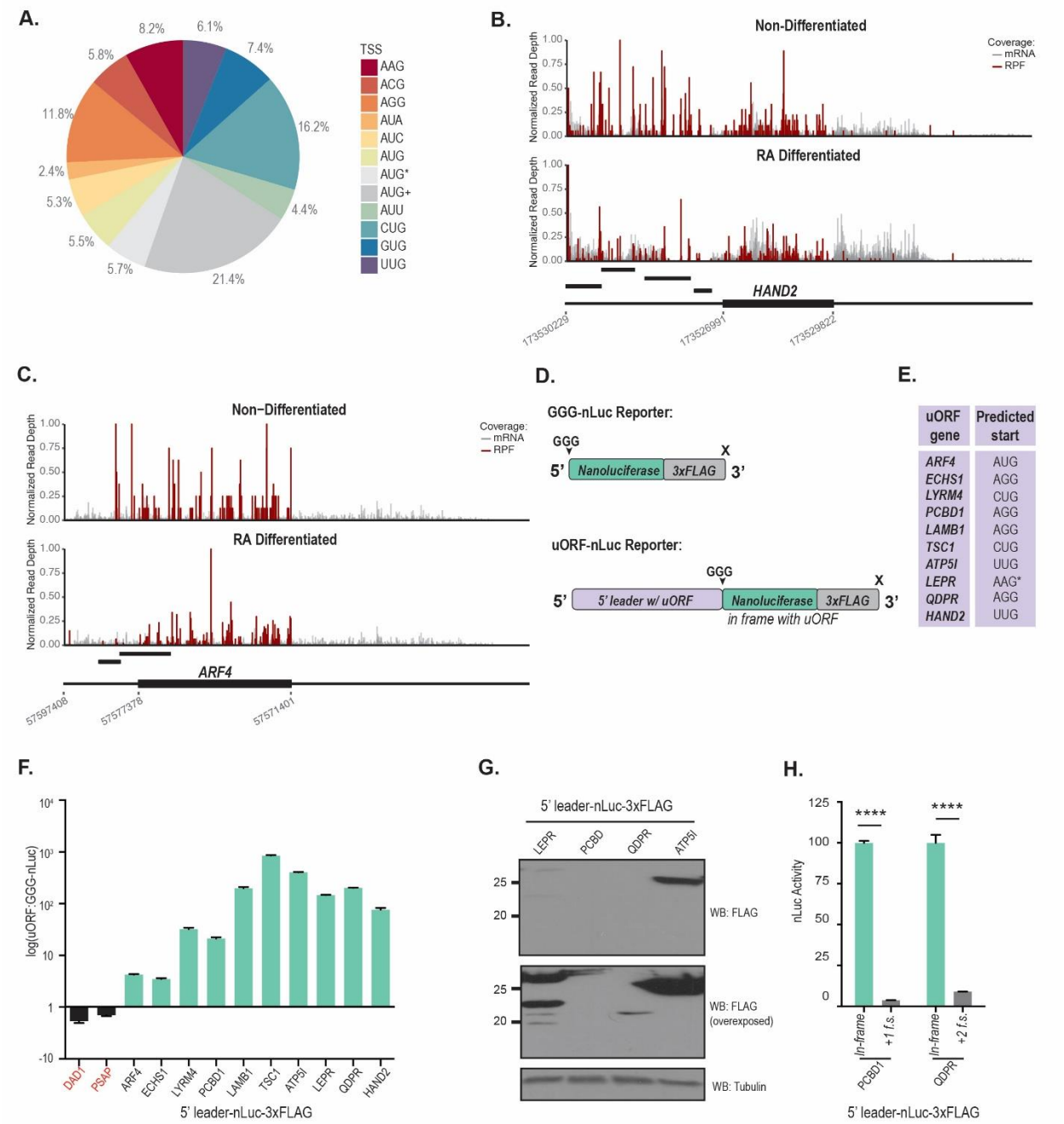

Figure 3.5 Validation of SPECtre scored uORFs. A) Distribution of predicted ORF translation start site sequences. Near-cognate start codons are utilized in the majority of predicted ORFs. AUG is the single most common initiation codon; this was either directly identified by SPECtre (AUG), present within 30 nucleotides upstream of the start of the SPECtre signal without an intervening stop site (AUG*), or located greater than 30 nucleotides upstream of the start of the SPECtre signal without an intervening stop codon (AUG+). Due to the high translatability of an ORF with an

AUG start codon, these were annotated as putatively AUG-initiated events. B) and C) Plots show mRNA read coverage (gray), and RPFs (red). B) HAND2 is a transcript with 4 predicted upstream-initiated ORFs. C) ARF4 is a transcript with one predicted upstream-terminated ORF, and one CDS-overlapping ORF with condition-dependent translation. D) Schematic of the uORF nanoluciferase (nLuc) reporters used in this study. GGG-nLuc serves as a negative control, as its AUG initiation start codon is mutated to a GGG codon. This reporter supports very little translation. E) A table of the predicted start sites for each uORF reporter from the high confidence dataset. F) nLuc assays performed in SH-SY5Y cells confirmed expression of high confidence uORFs (purple). 5' leaders not included in the high-confidence dataset (black) are below the GGG-nLuc reporter activity and considered not translated. All values are normalized to the GGG-nLuc control performed in parallel during experimentation, data for individual reporters was collected in triplicate in multiple experiments. G) Western blotting confirmed protein production of the highly translated ORFs. H) Reporters were cloned so that the nLuc tag was shifted out of frame (f.s.) with the predicted ORF and the CDS start site (if present in the reporter). This lead to a significant decrease in reporter signal and confirmed frame specificity of our detection algorithm. n=3 for all experiments. **** denotes a Student's T-test $p \leq 0.0001$. All green ORFs in panel F) have a p-value ≤ 0.0001, and all black ORFs have a p-value < 0.05. Graphs represent mean ± S.E.M.

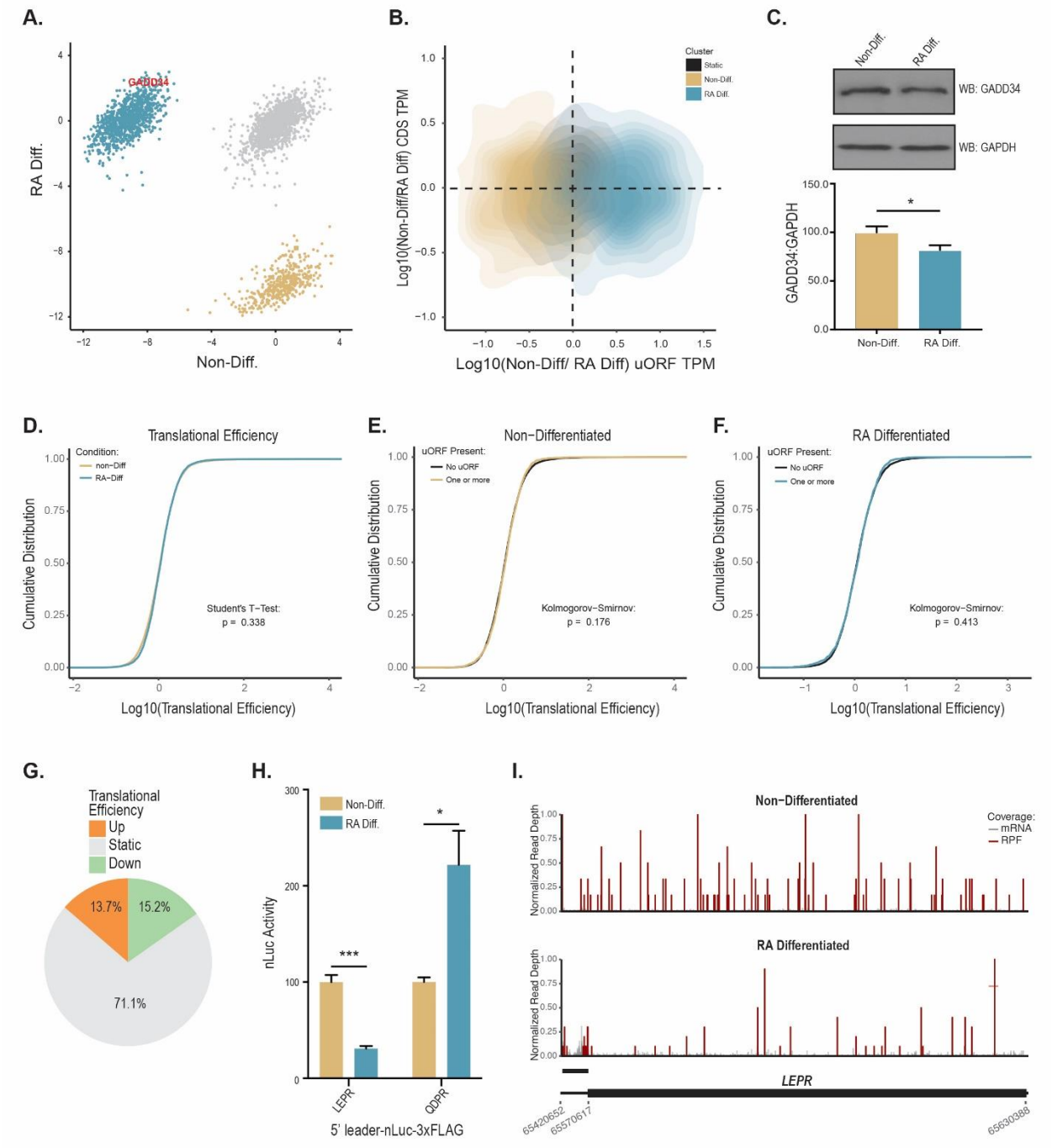

Figure 3.6 Translational efficiency of CDS with predicted uORFs. A) K-means clustering analysis of protein-coding CDS translational potential as score by SPECtre with predicted upstream-initiated ORFs in non-differentiated and RA-differentiated cells. Three clusters of CDS regulation emerge: those CDS that are up-regulated in RA-differentiated cells (cyan), up-regulated in non-differentiated cells (gold), and CDS with no change in translational potential (gray). B) Kernel

density estimation analysis of changes in TPM over annotated protein-coding CDS as a function of changes in TPM over predicted upstream-initiated ORFs. Cluster identity of predicted ORF changed in translational potential as scored by SPECtre is identical to panel A): predicted ORFs enriched for translation in RA-differentiated cells (cyan), predicted ORFs with enriched translation in non-differentiated cells (gold), and those with static translation across the two conditions (black, for visibility) are annotated to protein-coding CDS with higher RPF abundance in non-differentiated cells (above horizontal line), and those with higher RPF abundance in RA-differentiated cells (below horizontal line). C) Western blotting for GADD34 levels in non-differentiated and RA-differentiated cells showed a decrease with differentiation. This is opposite GADD34 ORF translation, which falls into the cluster of higher translation in the RA-differentiated condition A). D) Empirical cumulative distribution of translational efficiency in all protein-coding CDS across RA-differentiated (cyan) and non-differentiated cells (gold). E) Empirical cumulative distribution of protein-coding CDS with a predicted upstream-initiated ORF (gold) and those CDS without a predicted ORF in non-differentiated cells. F) Empirical cumulative distribution of protein-coding CDS with a predicted upstream-initiated ORF (cyan) and those CDS without a predicted ORF in RA-differentiated cells. G) Percent of genes with a predicted upstream-initiated ORF with higher translational efficiency in RA-differentiated cells (orange), lower translational efficiency in RA-differentiated cells (green), or no change in translational efficiency between RA-differentiated cells and non-differentiated cells (gray). H) nLuc assays of the specified ORF reporters transfected in both non-differentiated and RA-differentiated cells confirmed a shift in translation as predicted by SPECtre. n=3 for each condition in panel H). * denotes a Student's T-test $p \leq 0.05$ and **** $p \leq 0.0001$. Graphs represent mean ± S.E.M.

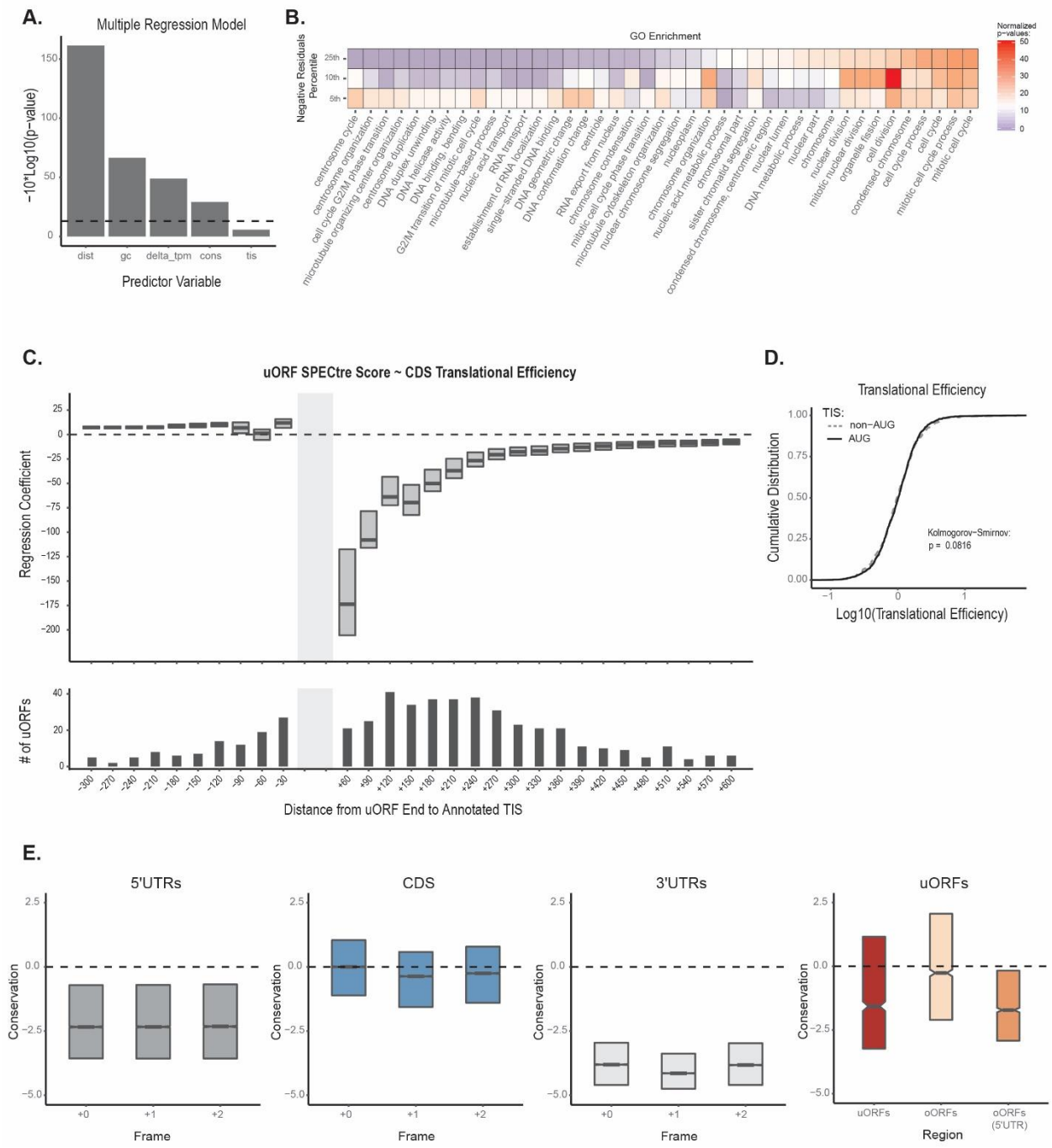

Figure 3.7 Characterization of predicted uORF regulation and downstream CDS. A) Significance of prediction variables in multiple regression model of protein-coding CDS translation regulation. Absolute distance of the predicted ORF position (*dist*), GC content of the 5' annotated region of predicted ORFs (*gc*), change in TPM over predicted ORFs (*delta_tpm*), and conservation of the 5' annotated region of ORFs (*cons*) are significant predictors of changes in CDS translation

(horizontal line denotes a significance value cutoff of 0.05), whereas translation initiation site identity of the predicted ORF (*tis*) is not a significant predictor. B) Gene set enrichment testing of the 5th, 10th and 25th percentile of genes with negative residuals in the multiple regression model. Gene sets are considered as significantly enriched using a multiple testing corrected p-value cutoff of 0.05 (white), with highly significant enriched terms in red. C) Linear regression model of protein-coding CDS translational efficiency as a function of predicted ORF SPECtre score and proximity of ORF termination to annotated CDS start site. Boxes are comprised of the linear regression coefficient in both biological replicates in non-differentiated and RA-differentiated cells, with the mean regression coefficient of all four replicates denoted by a black bar. Dashed horizontal line denotes a linear regression coefficient of zero. D) Empirical cumulative distribution of protein-coding CDS with upstream-initiated ORFs with a predicted AUG translation initiation start site (solid line), and those predicted to initiate at a non-AUG start codon (dashed line). E) Conservation analysis of annotated 5'UTRs in all three reading frames (far left, dark gray), annotated CDS regions over all three frames (middle left, blue), annotated 3'UTRs in all three reading frames (middle right, light gray), and predicted ORFs. Predicted ORFs are scored according to their termination in the 5'UTR (red), in the annotated CDS (light orange), or the portion of CDS-terminated ORFs that overlap the 5'UTR (dark orange).

CHAPTER 4

Integrative profiling of chimeric junctions with ribosome associated translation in prostate cancer

Modified from work in progress: Chun, S.Y. and Mills, R.E. Integrative profiling of chimeric junctions with ribosome-associate translation in prostate cancer.

The material presented in this chapter is derived from preliminary work in progress and is not being pursued as a manuscript at this time. Dr. Mills and I devised the context and scope of the analysis. Dr. Mills provided guidance for the experimental approach, and I led the development and implementation of the computational and analytical pipeline.

4.1 Introduction

Cancer is a significant threat to public health, and is the leading cause of death in the United States. In men, prostate cancer is the leading type of malignancy diagnosed, with the majority of men developing the disease by the age of 80. In 2017, prostate cancer is projected to be the most commonly diagnosed cancer type in men with over 160,000 new cases in the United States alone, and is predicted to be the third-leading cause of cancer related death regardless of gender.[153] As such, early detection and further insight into the mechanism of its development are critical to long-term prostate cancer treatment and patient outcomes.

Existing methods for early detection in the clinic are reliant on screening for prostate-specific antigen serum levels. However, tests for PSA may not accurately differentiate aggressive prostate

cancer from other milder forms, or distinguish them from benign dysplasias. Therefore, the need is exigent for biomarkers and methodologies that accurately and sensitively prioritize aggressive malignancies of the prostate.[6,154-157]

The development and application of high-throughput sequencing approaches have enabled deeper insight into the genomic, transcriptomic and proteomic changes that underlie prostate cancer. The landscape of prostate cancer specific single nucleotide variants, insertions, deletions, copy number variants, and structural variations has emerged with the continued advancement of profiling methodologies and analytical algorithms.[158-162] Perhaps among the most well-characterized are the highly recurrent gene fusions that involve the ETS family of transcription factors, like TMPRSS2-ERG.[163-167] Aberrations like TMPRSS2-ERG confer a selective advantage for tumor cell growth and proliferation; these mutations 'drive' the progression of cancer. In contrast, the landscape of mutations in cancer are also comprised of 'passenger' events, which are those that do not provide a direct or indirect advantage to tumor cells.[168] Thus, clinicians and researchers recognize the need to not only identify the total catalog of aberrations in cancer, but to distinguish mutations that directly influence the development or progression of the disease from those that do not.

Chimeric gene fusions offer significant diagnostic and therapeutic potential; their expression often dysregulates critical pathways related to cell cycle regulation, proliferation and differentiation, and since they are specific to malignant cells, may present opportunities for targeted therapy. Approximately 50% of tumor samples collected from patients positively screen for serum PSA harbor a TMPRSS2-ERG gene fusion.[167] As such, methods for the detection and classification

of known gene fusion events, as well as the identification of novel variants is under continued and rigorous development. [169-171] Gene fusions often drive the aberrant overexpression of the downstream partner. Alternatively, gene fusion events result in the creation of a protein product with novel or altered function from the two partners. Other gene fusions may produce a truncated protein with altered stability, structure, or function.[172,173] Although transcriptome sequencing has proven useful in the efforts to identify and characterize the contribution of chimeric gene fusion events to cancer, comprehensive profiling at the proteome has been elusive. Due to the lack of highly-specific antibodies, immuno-based detection methods are prone to high rate of false positive identification.[174] Furthermore, the short length and altered stability of certain gene fusions makes mass spectrometry-based detection less than ideal.

Building on advances in next-generation library preparation and sequencing, a method was recently developed for the high-throughput profiling of mRNA actively bound by translating ribosomes.[85] Ribosome profiling is a next-generation sequence methodology designed to survey the landscape and dynamics of mRNA translation into protein. In tandem with mRNA-Seq, analysis of ribosome-protected fragments of mRNA offers significant potential to profile the translational efficiency of protein synthesis, and contextualize those changes against the topology of chimeric gene fusion events. Although transcriptional profiling has driven many insights in understanding the impact of gene fusion events in cancer, placing those insights within the context of translational dynamics may offer further clues regarding the mechanism of their expression, as well as their functional importance. To this end, we propose the development of an integrative analytical and data visualization platform for the translational profiling of junctions with ribosome-associated translation (juncRAT).

In this study, we aim to establish the computational and analytical framework to survey the translational dynamics of chimeric gene fusions in prostate cancer; more specifically, we seek to characterize the translational context of these gene fusion and deconvolute the regulatory contribution of chimeric gene fusions to the development and progression of prostate cancer. In particular, we aim to spatially differentiate chimeric gene fusions by their transcriptional, and their translational activity. To that end, we aim to profile the translational dynamics of known, and newly detected gene fusion events through the integrative analysis of mRNA and ribosome profiling sequence data.

4.3 Materials and Methods

Prostate cancer cell lines and data access

In total, we evaluated five publicly available sets of next-generation sequencing data derived from the prostate cancer cell line PC-3: three (3) sets of paired-end mRNA-Seq data, one (1) set of singled-end mRNA-Seq data, with a matched set of ribosome profiling sequence libraries (Table 4.1). The paired-end mRNA-Seq libraries are 100 to 102 nucleotide fragments, derived from three separate experiments on control PC-3 cells. The single-ended mRNA-Seq library consists of 40 nucleotide length reads from vehicle-treated PC-3 cells. The matched set of ribosome profiling reads is derived from the same set of vehicle-treated PC-3 cells. Biological replicate sequence libraries were aligned separately, and merged for subsequent analyses. Ribosome profiling and mRNA sequencing libraries are listed in Supplemental Table C.1.

Integration of known and novel gene fusion breakpoints

Known prostate cancer specific breakpoints were downloaded from the COSMIC database (v82) in HGVS format.[175,176] COSMIC gene fusion records are supported by various levels of computational and experimental evidence through manual curation of the available research literature. HGVS variant records were converted to transcript coordinates, and 150 nucleotides of sequence upstream and downstream flanking the breakpoint were extracted from the hosted transcript sequence database. FASTA records for each flanked breakpoint sequence were generated for alignment index generation.

Novel and known chimeric gene fusion breakpoints were identified using three previously published detection algorithms: TopHat-Fusion, STAR-FUSION, MACHETE.[169-171] TopHat-Fusion uses a read segmentation algorithm to identify novel and known gene fusion breakpoints, re-maps single- or paired-end reads across those breakpoints, and then identifies candidates based on minimum read coverage heuristics over the breakpoint. STAR-FUSION uses a two-pass alignment algorithm and stringent filtering criteria to identify high confidence fusion gene breakpoints. MACHETE uses a two-stage approach to identify candidate gene fusions: potential breakpoints are identified using spanning and split read alignments, and the alignment score, mapping quality, and the amount of overlap of junction-spanning reads are used to build a statistical model and score each nominated breakpoint.

For higher sensitivity and specificity of fusion breakpoint detection, only the paired-end and not single-ended mRNA-Seq libraries were aligned to the hg38 genome and transcriptome. Default parameters were specified for each detection algorithm, and all nominated breakpoints from the

three detection algorithms were aggregated. Specific junctions identified by a single algorithm were retained for subsequent re-alignment using the single-ended mRNA-Seq and ribosome profiling libraries. Flanking 150 nucleotide sequences upstream and downstream of the candidate breakpoint were extracted from hg38 transcript records and output to a FASTA file. Novel and known breakpoints were aggregated, and combined with annotated CDS sequences (hg38) to build an integrative transcriptome and breakpoint compressed query index using Bowtie (v1.1.2).

Breakpoint sequence alignment and profiling by SPECtre

For consistency in alignment parameterization, single-ended mRNA and ribosome profiling sequence reads were aligned to the aggregated fusion breakpoint and annotated CDS reference using Bowtie (v1.1.2).[62] Alignment was based on an initial seed alignment length of twenty-four (24) nucleotides, allowing up to two (2) mismatches in the seed. Following alignment, breakpoints without at least one mRNA-Seq or ribosome profiling read spanning the junction were discarded. Putative breakpoints were further filtered based on the alignment quality of the reads spanning the junction, with no mismatches in the alignment permitted immediately upstream and downstream of the breakpoint. Thus, putatively transcribed breakpoints are defined as those junctions with minimum mRNA-Seq read coverage, and putatively translated breakpoints are defined as junctions with minimum coverage by ribosome profiling RPFs.

Aligned RPFs were adjusted to their P-site offset position, and the normalized depth at each position in mRNA-Seq and ribosome profiling was calculated as the per-position reads per million mapped reads. The RPM normalized depth at each position in the breakpoint was divided by the maximum depth across both sequencing profiles to determine relative positional depth. The

coherence of this depth-normalized RPF coverage was scored against an idealized tri-nucleotide signal over 30 nucleotide sliding windows across the breakpoint and flanking regions. An empirical model of translational potential was calculated from the distribution of SPECtre scores over annotated protein-coding CDS regions. In addition to the SPECtre signal profiled over each breakpoint and protein-coding CDS, the corresponding phase in each sliding window was extracted.

4.4 Results

Integration of known and novel chimeric gene fusion breakpoints

Technical issues precluded the prediction of breakpoints using TopHat-Fusion and MACHETE; TopHat-Fusion has a known memory allocation error that prevents proper execution, and parallel processing of MACHETE is only possible on a cluster running Sun Grid Engine. Therefore, only the results of COSMIC database and STAR-FUSION prediction are presented at this time.

Forty-two (42) prostate-specific HGVS gene fusion annotations were extracted from the COSMIC database, and converted to FASTA-formatted breakpoint sequences using a custom Python script (Supplemental Methods, Appendix C). Of these 42 COSMIC database breakpoints, 19 were identified in aggregate from STAR-FUSION prediction using the seven paired-end mRNA-Seq libraries (Figure 4.2A); since PC-3 is TMPRSS2-ERG gene fusion negative, it was not predicted by STAR-FUSION.[165] In addition to these previously annotated fusion events, STAR-FUSION predicted an additional 142 breakpoints that passed their default heuristic filtering criteria. These

previously annotated and *de novo* predicted breakpoints define the set of chimeric fusions to be re-aligned by matched single-end mRNA-Seq and ribosome profiling PC-3 libraries.

Breakpoint alignment using single-ended mRNA-Seq and Ribo-Seq

As previously described, nominated breakpoints derived from COSMIC and STAR-FUSION detection were used to construct an integrative breakpoint database of previously annotated and *de novo* predicted junctions. Following alignment with single-ended mRNA-Seq and ribosome profiling libraries against this integrative junction database, 12 COSMIC gene fusions and 32 STAR-FUSION predicted events had evidence of transcriptional or translational activity, as defined by minimum spanning read coverage over their junctions (Figure 4.2B). Of the 32 STAR-FUSION nominated events, the majority were identified by only one out of the seven original paired-end mRNA-Seq libraries. However, several of the events are supported by a minimum of two, and up to six of the individual paired-end mRNA-Seq libraries (Figure 4.3). In this study, chimeric gene fusion breakpoints are profiled evenly by both technologies, with breakpoints covered by similar numbers of junction spanning reads in both mRNA-Seq and ribosome (Figure 4.4A, p-value > 0.05). In addition, COSMIC integrated and STAR-FUSION nominated breakpoints are profiled to similar depths of coverage by junction spanning reads (Figure 4.4B and Figure 4.4C, respectively). Although only a subset of the integrated set of previously annotated and *de novo* predicted gene fusions are identified by the re-alignment of single-ended mRNA and RPF sequence libraries, this is not unexpected due to their shorter length (40 nt for single-ended mRNA and RPF libraries). Furthermore, paired-end alignments provide additional structural evidence for previously unannotated junctions in the absence of breakpoint spanning reads. Thus, alignment of single-ended mRNA-Seq and ribosome profiling reads positively identifies both

previously annotated and novel breakpoints generated by integrative database and *de novo* breakpoint discovery.

Coherence profiling of fusion breakpoints

In order to further define a set of highly transcribed or translated fusion breakpoints, additional heuristic coverage quality metrics were applied; based on a minimum junction spanning depth of 3 mRNA-Seq or ribosome profiling fragments, and upstream and downstream read abundance, 14 breakpoints were selected for additional spectral analysis for translational potential (Supplemental Table C.2). Among these 14 breakpoints, four are derived from COSMIC database integration, and include known ETS-family fusion events (ETV1-ACSL3, ETV1-HNRNPA2B1, and ETV4-CANT1), and the TPM4-ALK chimeric gene fusion. Of these fourteen events, 9 are putatively transcribed and translated according to the mRNA-Seq and ribosome profiling spanning read coverage, respectively.

Based on the mRNA-Seq and ribosome profiling read coverage over these junctions, several patterns of transcription and translation emerge, including breakpoint partners with varying levels of mRNA and RPF support before and after the junction (Figure 4.5), and fusion partners with evidence of transcriptional and translational activity both upstream and downstream of the breakpoint (Figure 4.6 and Figure 4.7). Transcriptional activity without corresponding translational evidence in the form of junction spanning RPFs may suggest alternative models of transcriptional or post-transcriptional regulation for other candidates (Figure 4.8). Given the limited sample of candidate fusions, it is still clear that in conjunction with conventional mRNA-Seq studies, coupling them to ribosome profiling experiments may offer potential utility as

translational validation for chimeric transcripts, as well as present opportunities to infer alternative regulatory mechanisms for putative fusion partners.

However, simple RPF coverage over chimeric junctions may not be sufficient to infer their active translation. Although the ETV1-HNRNPA2B1 junction is supported by mRNA-Seq coverage that spans the breakpoint, RPF coverage over the junction is nominal (Figure 4.5). Indeed, when the normalized RPF coverage over the breakpoint region is scored by SPECtre, a dramatic drop in tri-nucleotide coherence is observed well before the breakpoint; this break in tri-nucleotide signal is suggestive that this breakpoint is not actively translated, despite the nominal spanning RPF coverage evidence that might indicate otherwise (Figure 4.9A). Although the coherence over the TXNRD1-UTP20 breakpoint decreases relative to the immediate regions upstream and downstream of the junction, the tri-nucleotide periodic signal is maintained across the breakpoint (Figure 4.9B). Although the coherence over the breakpoint itself falls relative to the proximal regions surrounding it, this fusion transcript may warrant further scrutiny. Likewise, coherence to the tri-nucleotide signal indicative of active translation is maintained across another ETV1 family fusion transcript (Figure 4.9C). Taken together, these results suggest alternative models of transcriptional and translational regulation that might be further leveraged to predict, or classify, the pathogenicity of gene fusion events based on their translational efficiency.

4.5 Discussion

Initial analysis of the integrative transcriptional and translational profiling of chimeric gene fusion junctions in cancer are limited both in scope and depth. Despite this, preliminary results suggest the potential utility of further dissecting the transcriptional activity over these gene fusion breakpoints with added translational context. Furthermore, the computational and analytical foundations are in place for the high-throughput integration and interpretation of multi-scalar transcriptional and translational profiling of curated and *de novo* gene fusion predictions. Based on this integrative profiling, deeper insight might be gained regarding the translational efficiency of specific gene fusion breakpoints, and provide additional granularity for the pathogenicity of 'driver' events from 'passenger' events.[168]

Integration of copy number alterations for coverage normalization

Initial transcriptional and translational profiling of chimeric transcripts like the ETV1-HNRNPA2B1 fusion event invite further mechanistic scrutiny, as well as highlight areas in which the proposed integrative profiling approach might be improved. The coverage in both mRNA and RPFs is drastically higher in the 5' partner relative to the 3' partner. Mechanistically, the presence of an alternative in-frame stop codon proximal to the loss of coverage is one plausible explanation. However, another regulatory factor that was not accounted for in this preliminary analysis is copy number variation. The overexpression of many genes, including gene fusion partners, may be driven by underlying alterations in copy number.[177] Thus, the enrichment in mRNA and RPF abundance on one side of the ETV1-HNRNPA2B1 breakpoint over the other may be an artifact of copy number variation. One avenue to assess the underlying copy number architecture of gene fusion partners might be to analyze whole genome sequence data in PC-3 cells; the mRNA and

RPF coverage over the breakpoint could be further normalized to account for CNV differences in the two gene fusion partners.[178]

Phase deconvolution of chimeric junction translation

In addition to the efficiency of translation, spectral coherence profiling may also be used to predict the phase of the signal measured. Taking advantage of the tri-nucleotide periodicity inherent to the codon-dependent elongation in peptide synthesis, the phase of predicted gene fusion produces could be de-convoluted from the annotated canonical protein product through re-sampling analysis ribosome profiling reads over a chimeric junction. In this way, the translational efficiency of the competitively synthesized protein products (e.g. canonical versus chimeric) could be compared and assessed for state-dependent changes in the magnitude of translational dysregulation, such as solid tumors versus metastatic malignancies.

Annotation of chimeric junctions with ribosome-associated translation

To promote integrative approaches to study the genomic, transcriptional and translational factors that contribute to the etiology of cancer, the computational and analytical framework used for this pilot study could be extended and scaled for larger-scale data integration and discovery. Additional sources of curated and semi-curated repositories may be integrated into the database of previously annotated breakpoints; these resources include the Mitelman database of gene fusions (http://cgap.nci.nih.gov/Chromosomes/Mitelman) and other hosted repositories like ChimerDB.[179] As described in Section 4.3 (Methods), novel fusion breakpoints identified by additional detection algorithms like MACHETE and TopHat-Fusion are planned for integration.

Since ribosome profiling serves as a level of translational validation, one advantage of the proposed pipeline is that rare, lowly expressed variants, or other fusion transcripts that are discarded by the heuristic filters of certain fusion discovery algorithms may be included for ribosome profiling analysis. For instance, one of the heuristic filters employed by the STAR-FUSION pipeline filters out promiscuous fusion partners; fusion transcripts that include promiscuous 5' or 3' partners may be retained for full translational investigation. To this end, we propose the creation of a data warehouse and visualization platform for the integrative analysis and annotation of chimeric junctions with ribosome-associated translation, or juncRAT.

Adaptation of randomized approaches to define breakpoints in the genome and transcriptome

Relatively few studies have evaluated the impact of genomic structural variants that result in the translocation of constitutively or aberrantly active promoter regions proximal to genes relevant to cancer development and progression. Furthermore, many of the current discovery platforms for fusion detection rely on conservative heuristic algorithms to filter out candidate breakpoints. For example, TopHat-Fusion and other callers like STAR-FUSION employ minimum junction overlap by spanning reads to identify a high confidence set of fusion transcripts.[169,171] In some respects, these minimum overlap cutoffs may be considered arbitrarily conservative and enrich for artifacts of amplification.[170] Randomized approaches to identify breakpoints using short read sequencing technology have been shown to sensitively and accurately identify SVs in the genome.[180] To fully capture the landscape of genomic and transcriptomic structural variants we propose to apply and adapt existing software to call pathogenic breakpoints from whole genome sequencing data, and identify previously annotated and novel chimeric junctions from paired-end transcriptomic data, respectively. Integration and translational validation of breakpoints identified

from this method established for calling genomic structural variants offers additional opportunities for collaborative and comparative research in the investigation of pathogenic gene fusion events in cancer.

4.6 Conclusion

The work detailed in this chapter of the dissertation lays the computational and analytical framework for the integrative profiling of the transcriptional and translational landscape of chimeric gene fusion transcripts in prostate cancer. Analysis of ribosome profiling coverage and translational activity over chimeric gene fusion breakpoints provides further evidence for their potential pathogenic contribution to cancer. In this way, the translational context provided by ribosome profiling may enable the increasingly granular characterization of the genomic and transcriptomic events that drive cancer development and its progression.

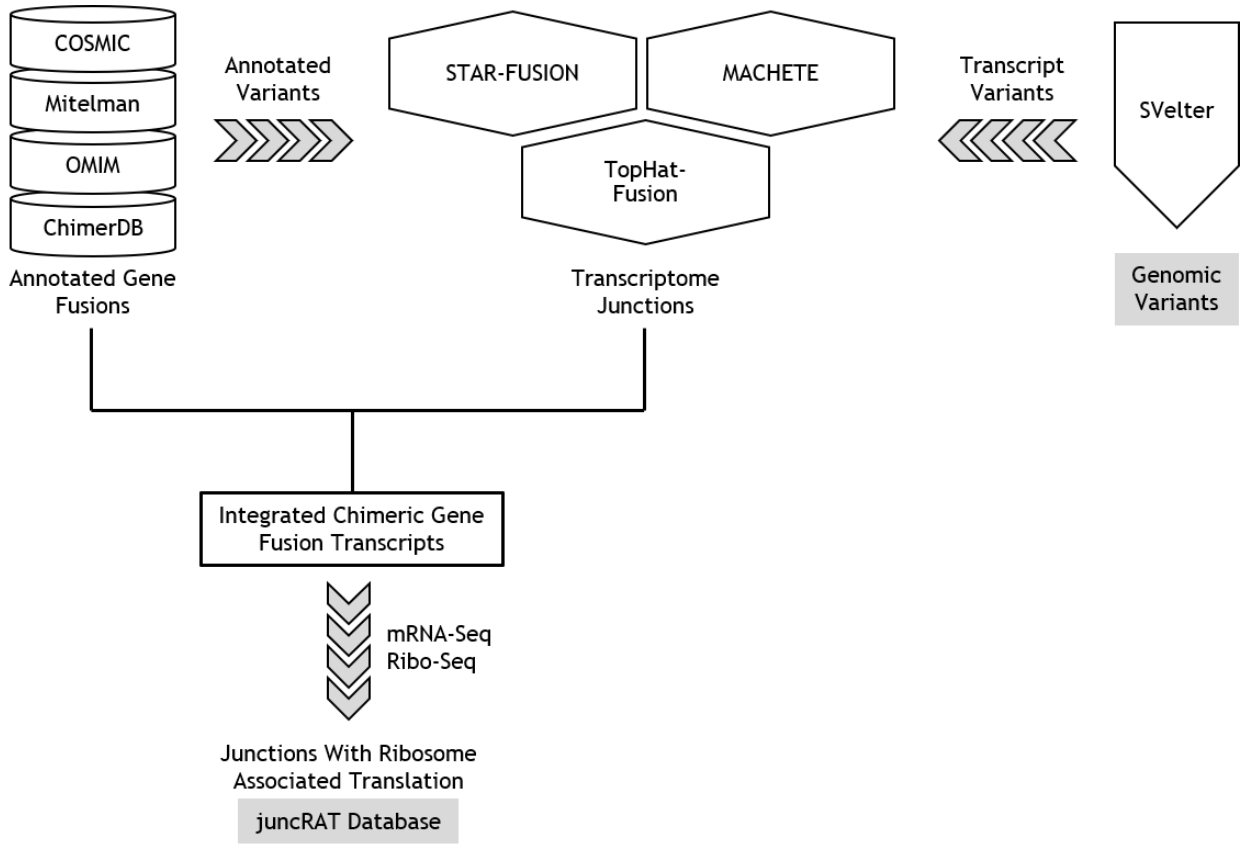

Figure 4.1 Schematic of the juncRAT alignment and analytical pipeline.

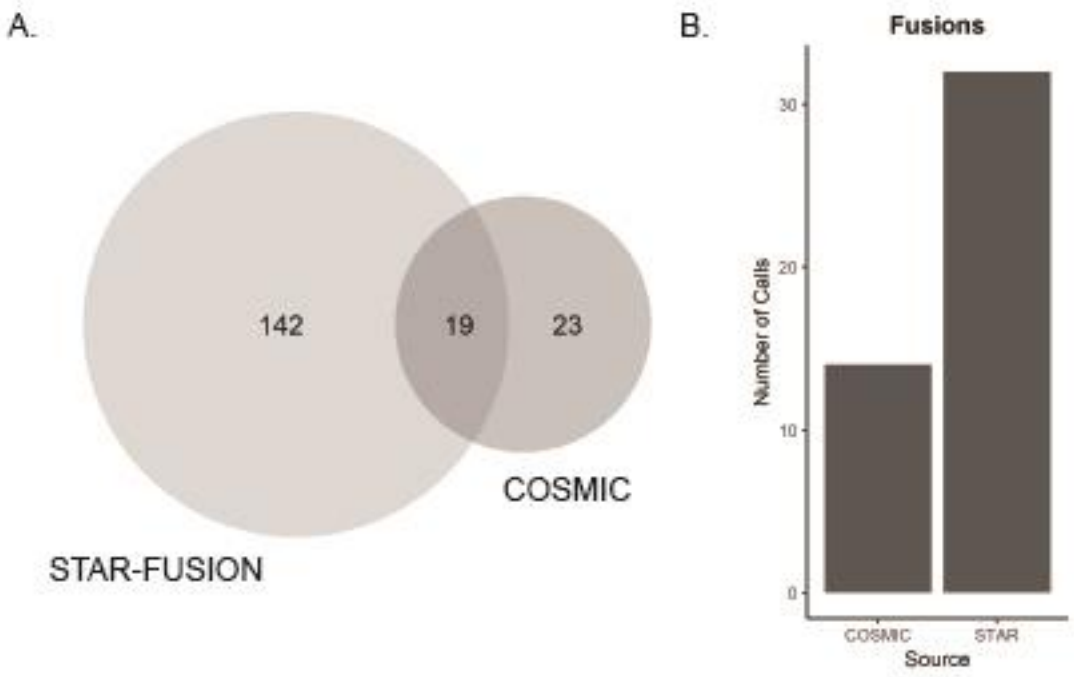

Figure 4.2 Integrative chimeric gene fusion breakpoint alignment. A) Number of novel fusion events (left), and the number of previously annotated fusion events identified by STAR-FUSION (middle). COSMIC events are the total number of prostate cancer specific events curated in the annotation database (v82). B) The number of COSMIC and novel events identified by re-alignment over those breakpoints with single-ended mRNA-Seq and ribosome profiling reads.

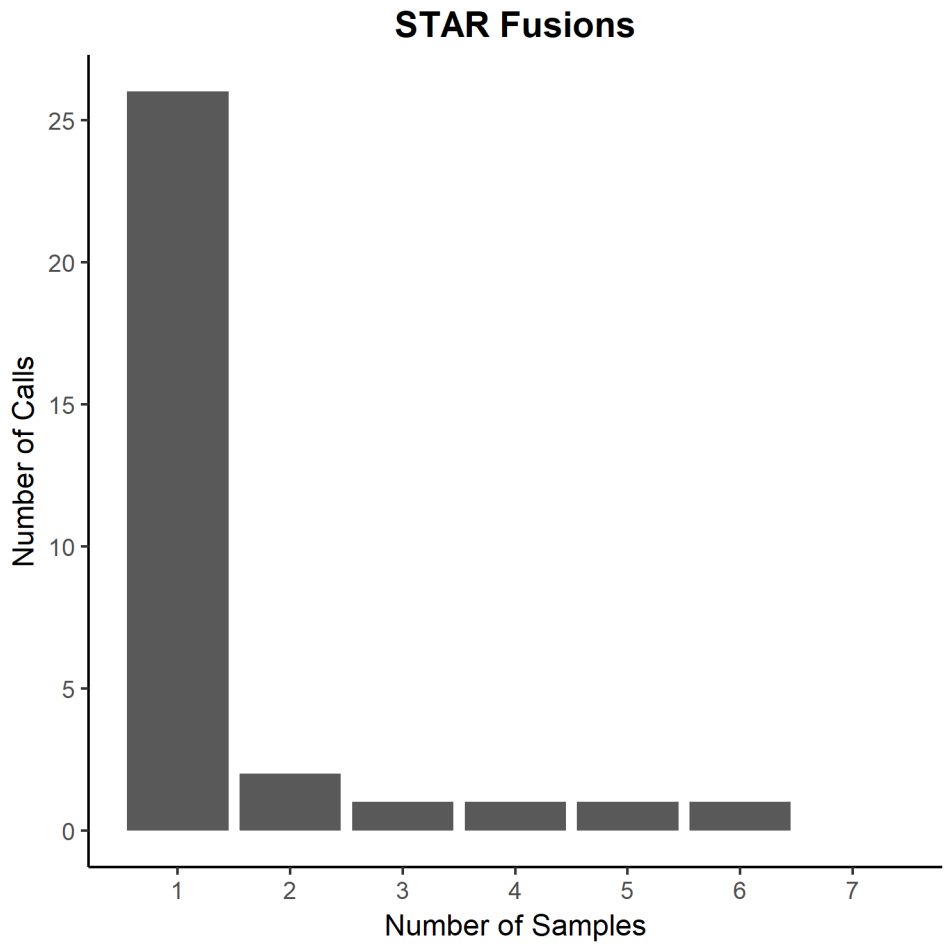

Figure 4.3 Paired-end library support of STAR-FUSION events. The number of events supported by at least one paired-end mRNA-Seq library, and those supported by multiple mRNA-Seq libraries.

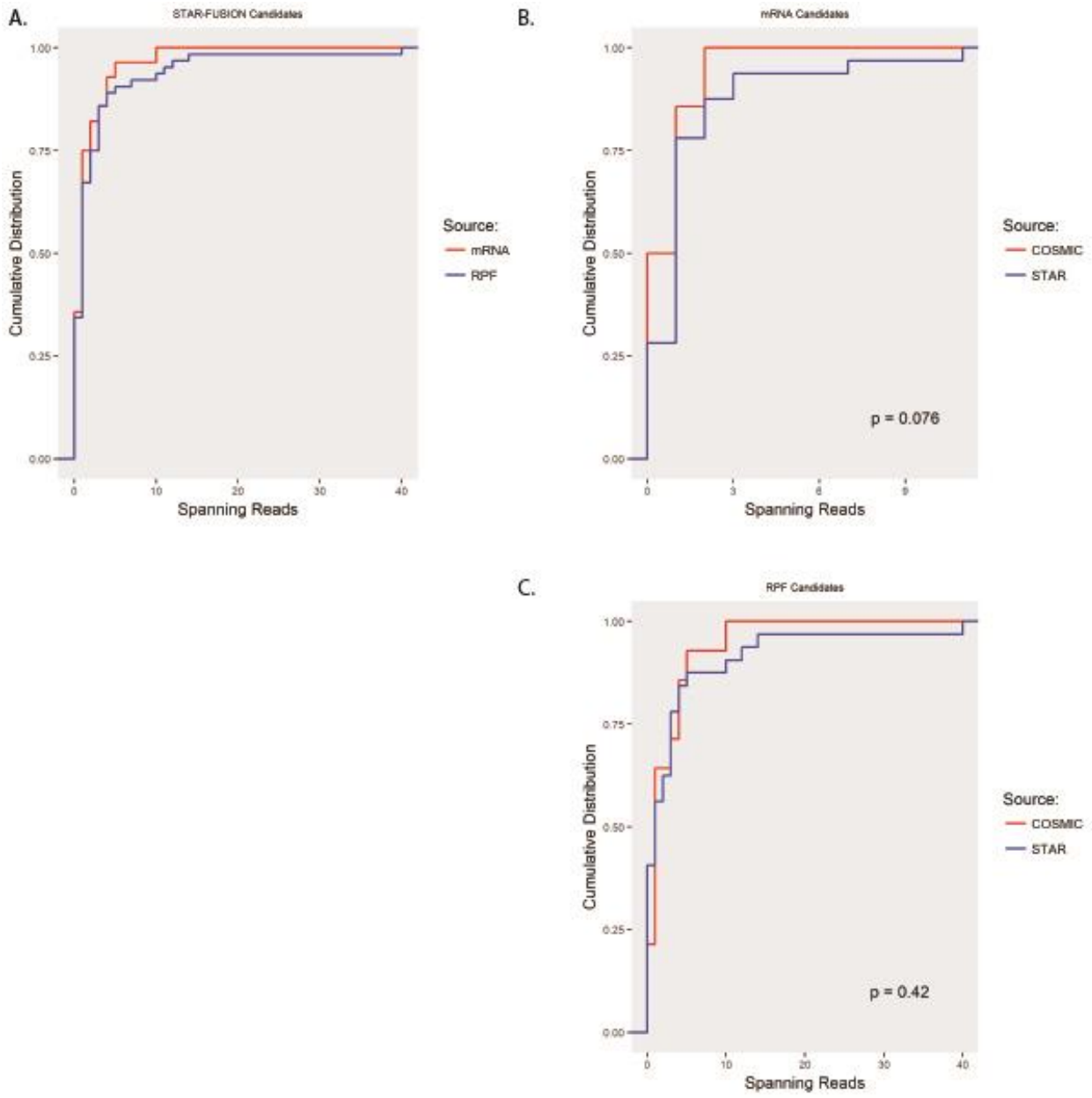

Figure 4.4 Number of spanning reads by breakpoint source and profiling method. A) Cumulative distribution of junction spanning reads profiled by mRNA (red), and RPF (blue). B) Cumulative distribution of junction spanning reads over COSMIC integrated breakpoints (red), and novel breakpoints (blue) identified by mRNA-Seq alignment by STAR-FUSION. C) Cumulative distribution of junction spanning RPFs over COSMIC integrated breakpoints (red) and novel STAR-FUSION breakpoints (blue).

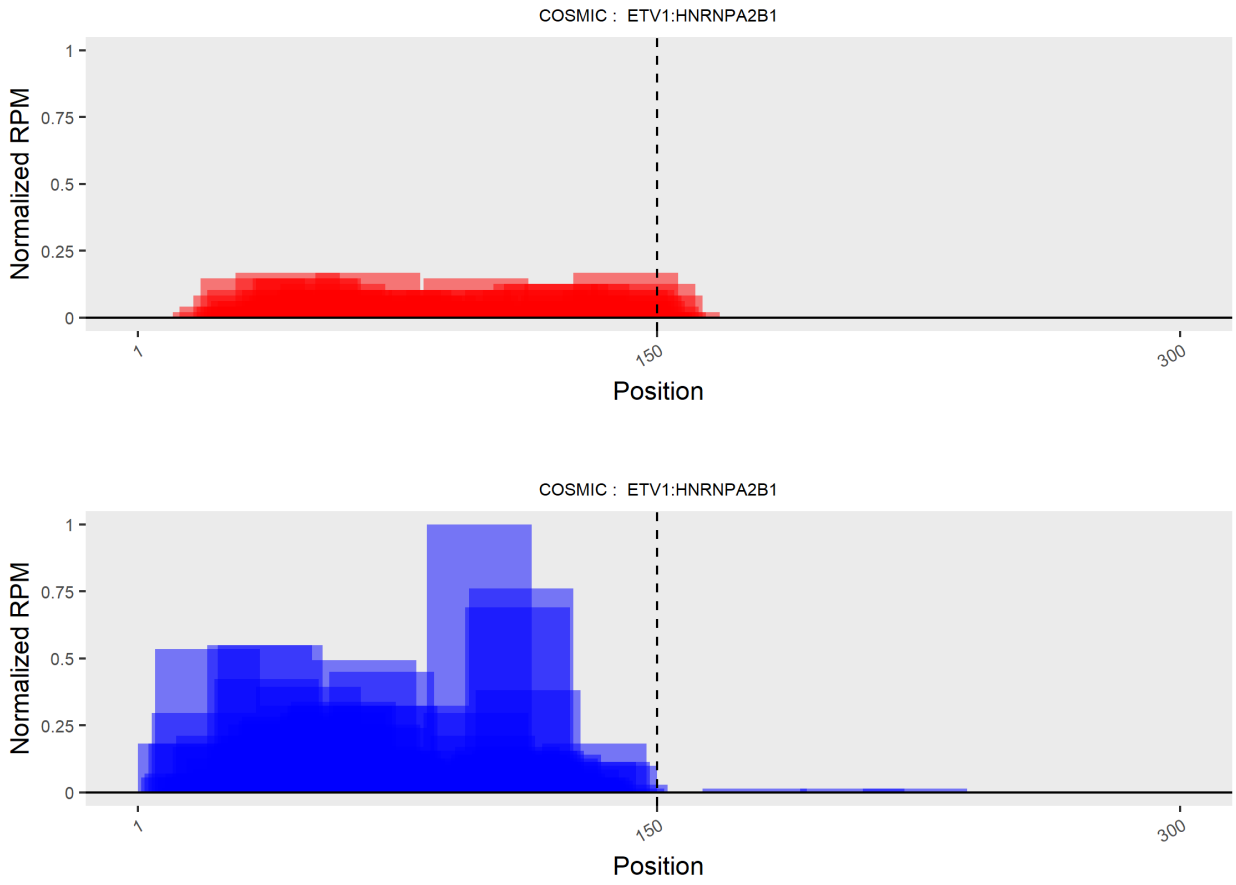

Figure 4.5 Coverage over the ETV-HNRNPA2B1 breakpoint junction. Normalized read coverage profile derived from mRNA-Seq (top, red) and ribosome profiling (bottom, blue). Read coverage is normalized by library size and then divided by the maximum depth in mRNA-Seq or Ribo-Seq to generate relative abundance over each position.

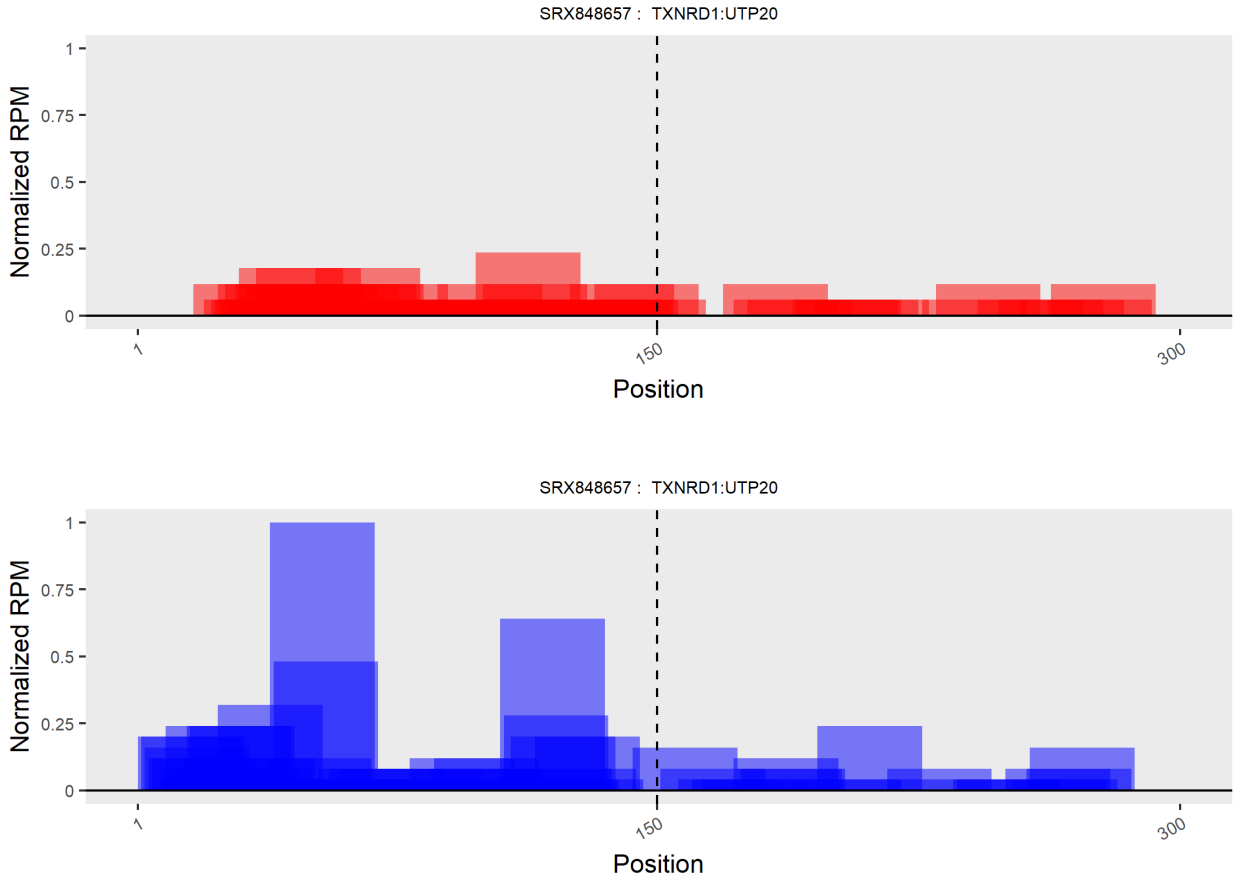

Figure 4.6 Coverage over the TXRND1-UTP20 breakpoint junction. Normalized read coverage profile derived from mRNA-Seq (top, red) and ribosome profiling (bottom, blue). Read coverage is normalized by library size and then divided by the maximum depth in mRNA-Seq or Ribo-Seq to generate relative abundance over each position.

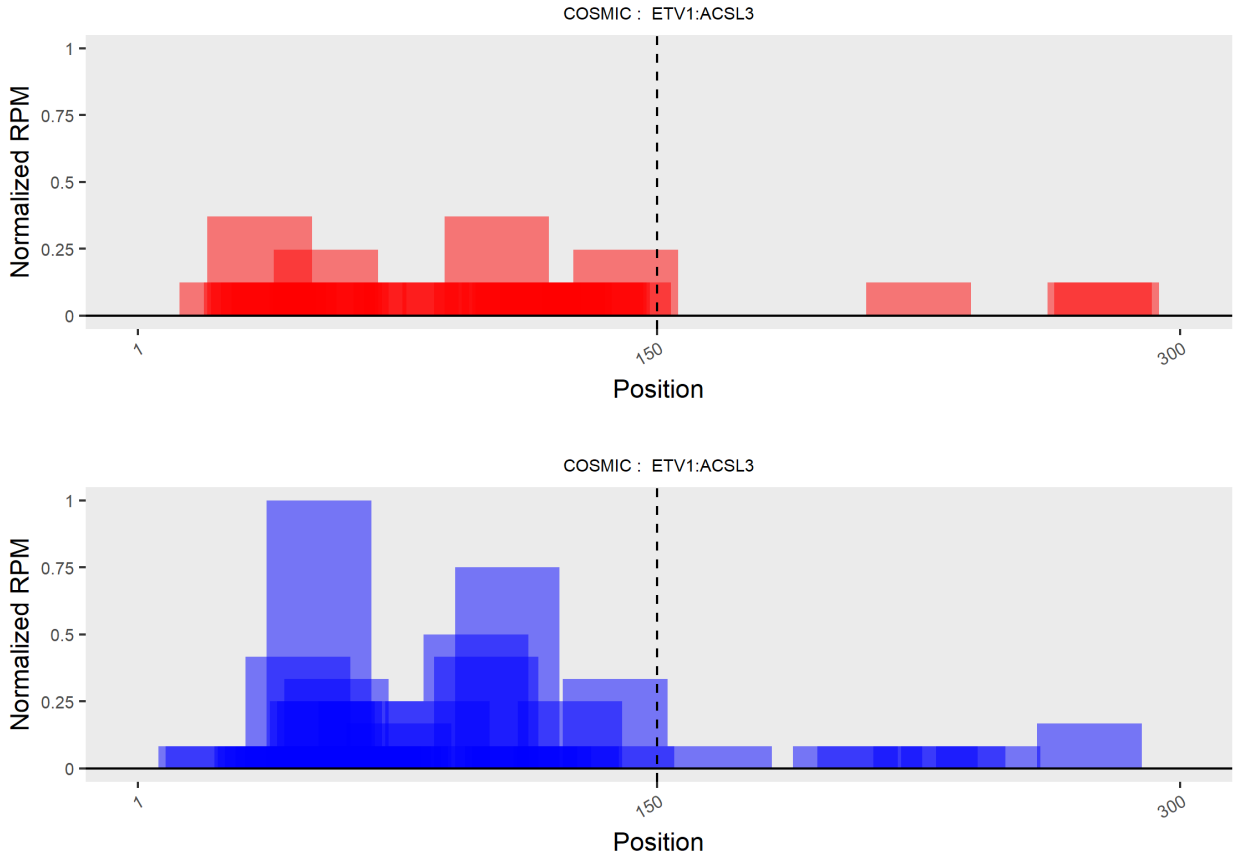

Figure 4.7 Coverage over the ETV1-ACSL3 breakpoint junction. Normalized read coverage profile derived from mRNA-Seq (top, red) and ribosome profiling (bottom, blue). Read coverage is normalized by library size and then divided by the maximum depth in mRNA-Seq or Ribo-Seq to generate relative abundance over each position.

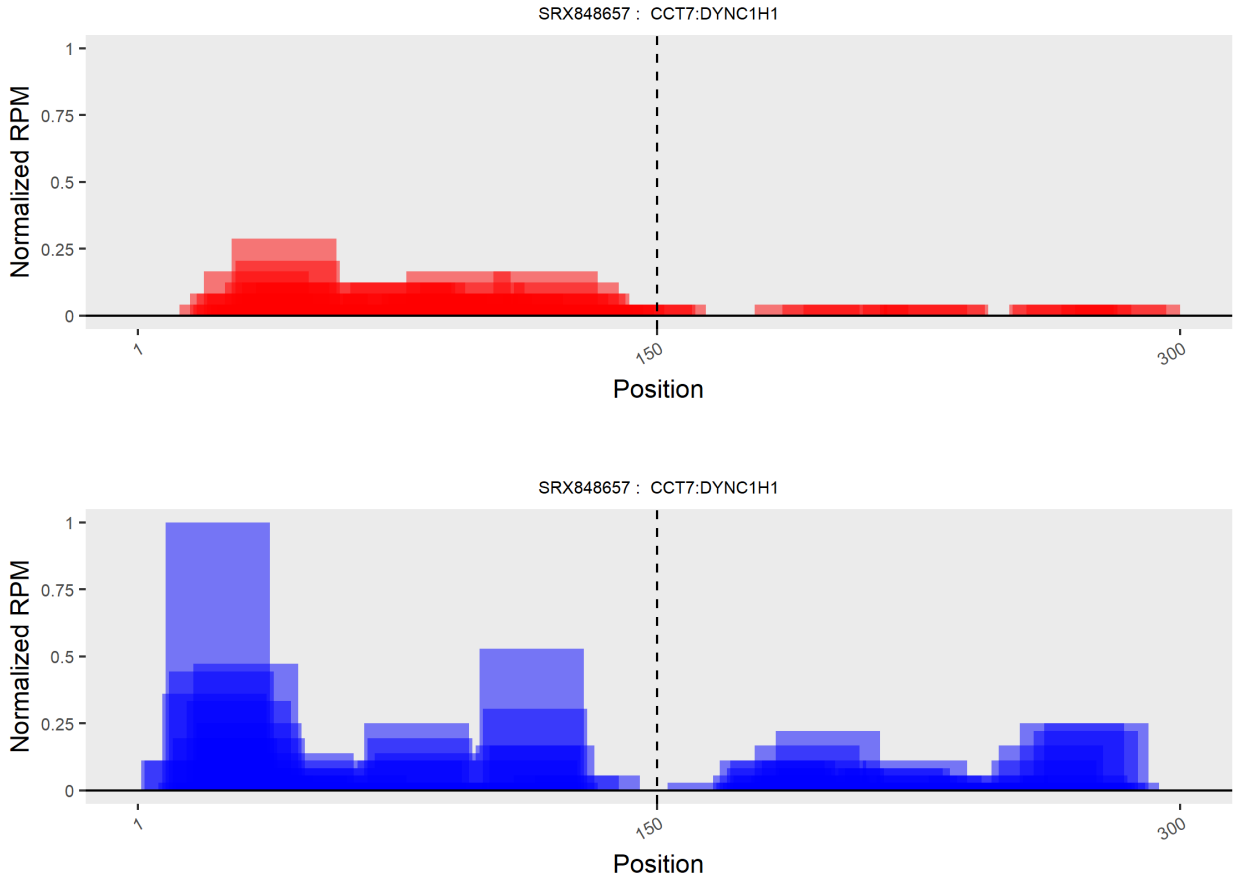

Figure 4.8 Coverage over the CCT7-DYNC1H1 breakpoint junction. Normalized read coverage profile derived from mRNA-Seq (top, red) and ribosome profiling (bottom, blue). Read coverage is normalized by library size and then divided by the maximum depth in mRNA-Seq or Ribo-Seq to generate relative abundance over each position.

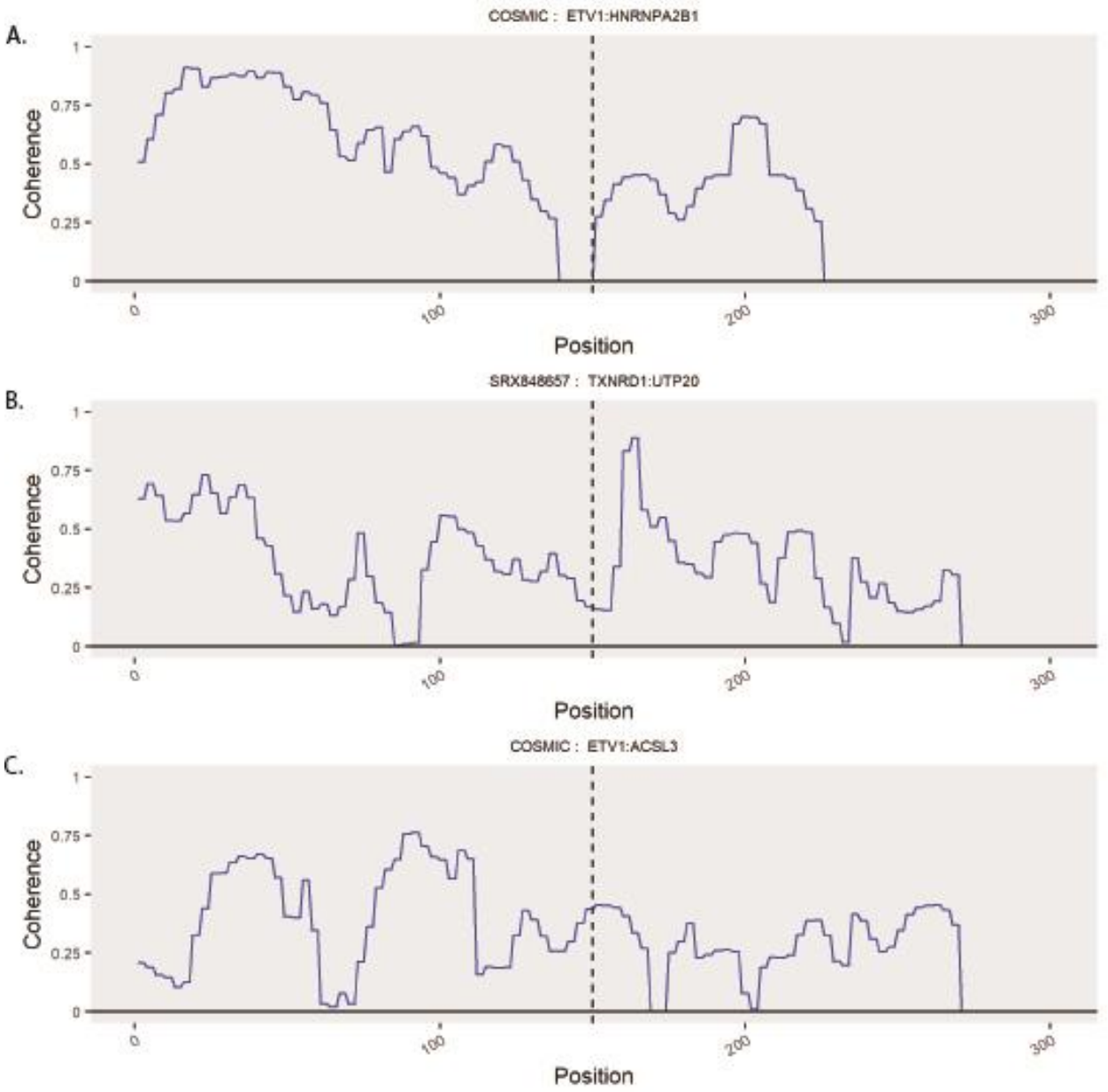

Figure 4.9 Ribosome profiling validation of junction translation. A) Spectral coherence profile over the COSMIC breakpoint ETV-HNRNPA2B1. Spectral coherence measures the strength of the tri-nucleotide periodicity inherent to active translation. B) Coherence profile over the STAR-FUSION chimeric junction candidate TNXRD1-UTP20 demonstrates decreased, but continuous, tri-nucleotide periodic signal across and through the nominated breakpoint. C) Continuous tri-nucleotide periodicity is observed across the COSMIC breakpoint ETV4-ACSL3 indicative of translational potential through the junction.

CHAPTER 5

Concluding Remarks and Future Directions

5.1 Translational Profiling and Protein Synthesis

Estimates of mRNA abundance have often been used as a proxy for protein synthesis, however direct comparison if complicated by both biological and technical considerations; multiple regulatory mechanisms may alter the abundance, structure, or localization of an mRNA or protein, and comprehensive characterization of the mRNA or proteomic content of a cell remains elusive due to technological or algorithmic limitations of the survey methodology employed. For instance, the short length and half-life of many uORFs may make their detection by mass spectrometry at the proteomic level difficult. Further, previous studies have demonstrated that direct comparison of mRNA and protein abundance, although positively correlated, are only moderately so. Since ribosome profiling involves the sequencing and analysis of mRNA directly engaged by actively translating ribosomes, it is strategically positioned to more intimately monitor the dynamics of protein synthesis than transcriptome sequencing. Indeed, early ribosome profiling experiments demonstrated that the abundance of ribosome-protected fragments of mRNA was more highly correlated with mass spectrometry estimates of protein abundance. However, given its relatively recent development, and specific limitations regarding library preparation and analysis, further

work may be required to more fully entrench ribosome profiling as a standard component of integrative transcriptional and translational studies.

Although protein synthesis estimates derived from the abundance of ribosome-protected fragments of mRNA are more highly correlated with mass spectrometry readouts of protein expression, technical limitations of the ribosome profiling methodology persist that prevent it from supplanting mRNA-Seq. In particular, ribosome profiling experiments are contaminated by higher levels of ribosomal RNA than comparable mRNA-Seq experiments; in practice, this means that typical ribosome profiling experiments capture less of the transcriptome at lower average depth of coverage. This reduced depth and diversity of transcriptome coverage by ribosome profiling relative to mRNA-Seq may be partially mitigated through a combination of additional sequencing investment and experimental design. In spite of this reduced depth and diversity of transcriptome coverage relative to mRNA-Seq, ribosome profiling estimates of protein synthesis are more highly aligned with protein expression. Furthermore, as a bridge between the transcriptome and proteome, ribosome profiling enables increasing insight into the mechanism of translation and translational regulation.

For instance, various ribosome profiling studies have been published that demonstrate engagement of the translational machinery outside of canonically annotate protein-coding regions, including uORFs. Since these uORFs are typically short in length and half-life, they are generally difficult to detect by mass spectrometry without size selection or N-terminal enrichment. Although uORFs had been shown previously to be important regulators of protein synthesis in limited contexts,

ribosome profiling is helping to further define their prevalence and contribution to the regulatory landscape of cellular control and diversity. In addition to the engagement of ribosomes over uORFs, several experiments have observed ribosomes engaged over long non-coding RNA loci with conflicting proteomic evidence. Although one group detected ribosomes engaged on long non-coding RNA regions, they were unable to find evidence of their translation by mass spectrometry, and concluded that their engagement by ribosomes was not indicative over active translation. Later, other groups detected a subset of long non-coding RNAs by mass spectrometry, and further defined their translational status by shared characteristics of ribosome-protected fragment coverage with protein-coding loci. Taken together, these results underscore the utility of ribosome profiling as an intermediate methodology to study translation and protein synthesis, as well as the need for the continued development and application of algorithms described in this thesis to more comprehensively characterize both the mechanism of translation and the global cellular impact of translational regulation.

5.2 Spectral Coherence Profiling

In chapter 2 of this dissertation, I presented SPECtre: an algorithm and software package to reliably identify regions under active translation in ribosome profiling sequence data using an approach that leverages the tri-nucleotide periodicity of codon-dependent peptide elongation. Spectral coherence is a signal processing algorithm originally developed for pattern recognition in time series data, and measures the similarity of two signals over the frequency domain. Although foundational aspects of this profiling methodology were preceded by another group using a similar

approach, this underscores the utility and potential advantages of translational profiling using a coherence-based approach. However, the translational coherence classification algorithm and software package presented in this differs from the earlier established method in two key aspects: codebase optimization resulted in more efficient usage of computational resources, but more significantly, SPECtre achieves comparable sensitivity and specificity in translational classification in the absence of matched mRNA-Seq data. Although deeper insight is gained by the tandem, and integrative analysis of transcriptional and ribosome profiling data, obviation of matched mRNA-Seq samples for translational analysis grants researchers wider latitude in experimental design.

5.3 Computational Prediction of Regulatory uORFs

In the third chapter of this dissertation, we extended the spectral coherence approach for classification of active translation to identify upstream-initiated open-reading frames with context-specific regulatory potential. A subset of computationally predicted uORFs that negatively regulate their corresponding protein-coding gene were experimentally validation in differentiated SY-SY5Y cells. Furthermore, we investigate multiple factors that might contribute to the regulatory impact of these ORFs. In addition to the translational activity of the uORF, sequence conservation and content, as well as spatial proximity to the annotated protein-coding region are predictive of its potential regulatory impact on protein synthesis. Importantly, we found that many of these translationally repressed genes are related to cell cycle control, regulation of cell division,

and chromosome organization. Finally, a limited subset of these uORFs was experimentally validated to negatively regulate protein synthesis in a cellular model of neuronal differentiation.

With the advent of translational profiling approaches like the next-generation sequencing of ribosome-protected fragments of mRNA, efforts to more comprehensively map the intricate network of protein synthesis regulation have intensified. Methods for the analysis of translational efficiency and dynamics using ribosome profiling data are under continued, and rigorous, development. Like SPECtre, some approaches have leveraged the tri-nucleotide periodic signal inherent to the codon-dependent mechanism of peptide elongation; however, I note several areas in which translational classification algorithms, including SPECtre, could be extended to more sensitively and specifically identify regions under active translation and explore their context-specific regulation implications. The work presented in this chapter of the dissertation underscores the contribution of multiple sequence and spatial parameters that underlie the regulatory potential of uORFs. However, additional parameters that were not assessed include mRNA secondary structure, k-mer content, tRNA abundance, and codon usage bias. Previous studies have demonstrated the significance of mRNA secondary structure on non-AUG translation initiation in certain contexts. Furthermore, tRNA abundance and codon usage bias have been shown to moderate the translational efficiency of transcripts in a resource-dependent manner. Finally, uORFs were predicted in a supervised manner; given the number of factors included (and not included) in our multiple regression model, the robust feature selection of unsupervised machine learning approaches may be well-suited for the sensitive and accurate prediction of translated uORFs with regulatory potential.

5.4 Translational Profiling of Chimeric Junctions

In chapter 4 of this dissertation, I proposed a framework for the integrative profiling of the transcriptional and translational landscape of gene fusions in cancer. High-throughput, massively parallel next-generation sequencing has empowered deeper insight into the genomic and transcriptional regulatory landscape that underlies the development of cancer and its progression. Immunoassay-based approaches to detect chimeric gene fusion events in tumor samples are limited by the availability and specificity of high-quality antibodies. High-throughput mass spectrometry and database search methods have become increasingly sensitive in their detection of gene fusion events, but some of these chimeric events may be difficult to capture due to their altered size, structure, or localization. Moreover, some gene fusion events may not be detected in the proteome due post-transcriptional regulation or decreased stability.

This work lays the initial computational and analytical framework for the integrative investigation of the transcriptional and translational landscape of chimeric gene fusion events in a context-specific manner across cancer. Based on preliminary alignment and analysis of mRNA-Seq and ribosome profiling reads, we observed differential patterns of transcriptional and translational coverage over a limited set of previously annotated and novel chimeric breakpoint junctions. Aspects in which this preliminary work could be immediately extended include the aggregation of additional curated databases of chimeric gene fusions, and the integration of multiple gene fusion calling pipelines to more comprehensively catalog potentially pathogenic chimeric events in cancer. Moreover, increasingly robust analysis of the transcriptional and translational coverage

over these chimeric junctions would account for copy number variations across the two fusion partners. However, these preliminary findings suggest the importance of accounting for translational regulation in the pathogenic characterization of gene fusion events, as well as for their suitability as potential biomarkers or therapeutic targeting. Tandem and integrative analysis of mRNA-Seq and ribosome profiling data could be readily adapted to existing clinical sequencing protocols, and offer deeper insight into the context-specific changes in the transcriptional and translational landscape that underlies cancer.

5.5 Concluding Note on Translational Profiling

As a relatively recent entrant into the field of next-generation sequencing, ribosome profiling and corresponding methods for its analysis are under continued and rigorous development. Through these efforts, we have gained deeper understanding of the diverse network of translational regulation that underlies gene expression and protein synthesis. However, early efforts were focused, and continue to be focused on assessing the prevalence and significance of uORF translation. Although continued development of analytical and computational algorithms is needed to meet these challenges, it seems to me that these advancements and the product of these efforts must be appropriately contextualized against the background of dynamic cell states and conditions. The field, including the work presented in this dissertation, has begun to shift from identifying the global prevalence of uORF translation to dissecting their role across cellular states. Likewise, the compendium of genomic and transcriptomic aberrations in cancer continues to grow as larger cohorts of tumor samples are sequenced and analyzed. The continued development of rigorous

analytical algorithms, and the growing computational resources to drive their application is leading the way for personalized medicine, including targeted cancer therapeutics. These therapeutic targets or diagnostic markers may be more efficiently identified if the underlying mechanism of their transcriptional and translational regulation is more fully understood. The work of integrating and analyzing this data will present exciting challenges, and empower deeper understanding of the dynamic changes that define cellular states.

APPENDIX A

SUPPLEMENTAL MATERIAL FOR CHAPTER 2

A.1 Supplemental Methods

Data access and pre-processing

HEK293 ribosome profiling alignments in BAM format, were downloaded from the data repository hosted by the authors of RiboTaper (https://ohlerlab.mdc-berlin.de/files/RiboTaper/alignment_files.tar.gz).[100] No further pre-processing for the HEK293 alignments was required. Zebrafish ribosome profiling libraries) were downloaded from the NCBI Gene Expression Omnibus (accession GSE53693).[99] Ribosome profiling data of mouse embryonic stem cells treated with cycloheximide was downloaded from GEO (accession GSE60095, sample GSM1464901).[95]

Mouse embryonic stem cell and zebrafish ribosome profiling sequence libraries were converted from SRA format to FASTQ. mESC ribosome profiling reads were trimmed of adapters according to previously published methods.[95] Adapter sequences were removed from zebrafish ribosome profiling reads and further trimmed based on base quality using *fastq-mcf* (https://github.com/ExpressionAnalysis/ea-utils). For both trimming methods, a minimum read

length of 24 nucleotides was required after adapter removal and supplemental trimming. The minimum threshold for base quality trimming using *fastq-mcf* was set to 10 over a consecutive window of 4 nucleotides. All sequence libraries were then aligned to their respective UCSC ribosomal RNA contaminant database; mm10 for mESC, and Zv9 for zebrafish embryos.[181] Sequence reads were aligned to their respective ribosomal RNA contaminant database using Bowtie version 1.1.2 with a seed length of 22 nucleotides, and allowing no mismatches in the seed alignment.[62]

Alignment and post-alignment processing

Mouse and zebrafish ribosome profiling reads that did not map to their respective ribosomal RNA contaminant database were aligned using TopHat version 2.0.[182] Zebrafish ribosome profiling reads were aligned to the Ensembl v78 genome and transcriptome references.[183] Mouse embryonic stem cell ribosome profiling sequence reads were aligned to the Ensembl v72 genome and transcriptome reference. All ribosome profiling sequence reads were aligned with TopHat parameters that required Bowtie v1.0.0, Solexa quality scores, no novel junctions to be generated, with a forward/unstranded library type designated.

Aligned mESC and zebrafish reads were filtered based on a minimum mapping quality of 10 using SAMTools, and then sorted by genomic position using Picard version 1.114 (http://broadinstitute.github.io/picard).[184] For meta-analysis of the zebrafish ribosome profiling data, aligned reads from the sixteen available zebrafish samples and replicates were merged into a

single BAM alignment file using Picard. Shell scripts and code specific to each experiment are reproduced in the proceeding sections.

A.2 SPECtre

Read coverage normalization

For a given transcript with coordinates defined by the set *C*, the A- or P-site adjusted read positions overlapping those coordinates are extracted from a BAM alignment file. The coverage over each coordinate position in the set is summed, then normalized to the highest coverage such that all coordinate positions defined by set *C* range from zero (no coverage) to one (highest coverage).

Spectral coherence

In signal processing, coherence measures the power relationship between two signals as a function of frequency. Coherence estimates range from zero, where two signals are fully independent of each other, to one, where one signal may be perfectly predicted by the other. Assuming a sampling interval **Δ** over time interval *T*, signal X_j and its Fast Fourier Transform (FFT) X_j^* define the power spectrum of signal *X* at frequency *j* as:

(1)

$$S_{XX,j} = \left(\frac{2\Delta^2}{T}\right) X_j X_j^*$$

The cross-power spectrum of signal X_j and Y_j at frequency *j,* is calculated as the mean of the product of signal X_j and the Fast Fourier Transform of signal $_j$ over *K* trials:

(2)

$$< S_{XY,j} > = \left(\frac{2\Delta^2}{T}\right)\left(\frac{1}{K}\right)\sum_{k=1}^{K} X_{j,k}\, Y_{j,k}^{*}$$

Coherence is defined as the magnitude of the cross-power spectrum between signal *X* and *Y* at frequency *j* divided by the product of the square roots of the power spectrum of signal *X* at frequency *j*, and the power spectrum of signal *Y* at frequency *j*.

(3)

$$Coh_{XY,j} = \frac{|< S_{XY,j} >|}{\sqrt{< S_{XX,j} >}\sqrt{< S_{YY,j} >}}$$

SPECtre scoring

The default SPECtre score is calculated as the average coherence over *N* nucleotide sliding windows across a normalized coverage region against an idealized tri-nucleotide control signal of the same length.[104] Welch's coherence decreases the variance of the coherence estimate at the expense of resolution. Alternatively, modified Welch's coherence estimates over a region may be calculated using the median, maximum, or the non-zero mean or median. The SPECtre score of a normalized coverage region *R* with coordinates *C*, at frequency *j* against an idealized tri-nucleotide signal *S*, over adjacent *N* nucleotide windows is given by:

(4)

$$Spec_{RS,j} = \frac{1}{M}\sum_{m=1}^{M} Coh_{R_{m,m+N}S_N,j} \;\; for\ all\ m + N \in C$$

The number of sliding windows over the coordinate set *C* may be modified based on the step size between each window. Given a coordinate set *C*, and step size of *L:*

(5)

$$W_n = C_{Ln}, for\ n \geq 1\ and\ L \geq 1$$

Therefore, the default SPECtre score of a normalized coverage region *R*, at frequency *j* against an idealized tri-nucleotide signal *S*, over *N* nucleotide windows with a step size of *L* is:

(6)

$$Spec_{RS,j} = \frac{1}{M}\sum_{m=1}^{M} Coh_{R_{m,m+N}S_N,j}\ \ for\ all\ m \in W_n\ and\ all\ m+N \in C$$

A.3 Analysis of Read Length Bias

Treatment with cycloheximide typically isolates ribosome-protected fragments 28 to 30 nucleotides in length, which align with high fidelity to regions annotated to protein-coding transcripts.[85] However, in the absence of cycloheximide, conformational changes in the ribosomal complex may enrich for a shorter range of RPFs that also map with high-fidelity to regions annotated to protein-coding transcripts.[92] It is possible that these shorter length RPFs may obscure the tri-nucleotide signal of longer length RPFs that may cause coherence-based classifiers, like SPECtre, to under-estimate the number of actively translated ORFs in a ribosome profiling experiment. Ideally, this could be tested using simulated data as is done for whole genome sequencing using *wgsim* (https://github.com/lh3/wgsim) or RNA-Seq.[185] However, unlike RNA-Seq, simulation of RPFs would have to account for the distribution of RPFs protected by

ribosomes (Figure A.1, see "All Reads") as well as variance in the tri-nucleotide periodicity signal once those RPFs are aligned to the transcriptome. Instead of simulating an entire ribosome profiling experiment, we have examined the robustness of SPECtre scoring as a function of increased variance in RPF lengths outside of the expected 28-30 nt range. We have simulated this by randomly sampling 10,000 RPFs from the over 500,000 mESC RPFs aligned to the housekeeping gene ACTB using a weighted biased probability function.

Given a distribution of aligned RPF lengths, *D*, in a ribosome profiling experiment, with the RPF lengths defined by the set, $L = \{18, 19, 20, \ldots, 38, 39, 40\}$, and the relative frequency of each RPF length given by p_{Ln}, we define the weighted bias for a given RPF length to be randomly sampled as:

(7)

$$W_{Ln} = \frac{p_{Ln}}{p_{Ln}^{b}}$$

Where *b* is the bias assigned to the sampling distribution, such that if $b = 1$ the weighted bias for a given RPF length to be randomly sampled would be defined by the experimental RPF length frequencies. In contrast, if $b = 2$, the weighted bias for a given RPF length to be randomly sampled would be defined by the inverse of the experimental RPF length frequencies. The effect of increasing *b* from 1.0 to 2.0 may be seen in Supplemental Figure 1; starting with $b = 1$, the random sampling (with replacement) of 10,000 reads from the ~500,000 RPFs aligned to ACTB closely conforms to the experimental RPF length distribution (Supplemental Figure 1, see "All Reads"). As *b* is increased from 1.0 to 2.0, the RPF length distribution demonstrates increased variance in

RPF lengths outside of the expected enrichment of 28-30 nt fragments to the extent that the RPF length distribution progressively resembles a uniform distribution.

Incrementing *b* from 1.0 to 2.0, we sampled 10,000 RPFs from ~500,000 aligned to ACTB with replacement using the *sample()* function in R. Sampling was performed with replacement due to the low number of RPFs at the low and high extremes, and to simulate the persistence of sequence duplication. This biased re-sampling was done over 10,000 trials, and in each trial the normalized read coverage over ACTB was calculated then scored using SPECtre against an idealized tri-nucleotide periodic signal of the same length. The results of these biased sampling simulations are shown in Supplemental Figure A.2; as *b* is increased from 1.0 to 2.0, the distribution of SPECtre scores is plotted with the median score denoted by the dark black inside each box, and the extremities depicted by the ends of each whisker. The horizontal black line represents the mean SPECtre score over all simulations; the dashed lines above and below mark the boundaries of the extreme outlier cutoff as defined by Tukey. Tukey's outlier cutoffs are defined as 1.5 times the inter-quartile range. Base on this outlier analysis of 10,000 trials over an increasing weighted bias for RPF length selection, SPECtre is robust against increasing variance in RPF lengths outside of the expected (28-30 nt) range.

A.4 Experimental Scripts

Adapter removal in read quality trimming

```
# For mESC libraries:
fastx_clipper -Q33 -a CTGTAGGCACCATCAAT -l 24 -c -n –v –i /path/to/FASTQ > clipped.fq
fastx_trimmer -Q33 -f 2 -m 24 -i /path/to/clipped.fq > trimmed.fq

# For zebrafish ribosome profiling library:
```

```
fastq-mcf -o /path/to/trimmed.fq -l 24 -q 10 -w 4 -t 0 /path/to/adapter.fa /path/to/FASTQ
```

Alignment to rRNA contaminant database

Ribosomal rRNA sequences may be downloaded as part of the iGenomes annotation (https://support.illumina.com/sequencing/sequencing_software/igenome.html) package, or from the UC-Santa Cruz Table Browser (https://genome.ucsc.edu/cgi-bin/hgTables).

```
bowtie -l 22 -n 0 -S --un ribo-rRNA.fq /path/to/rRNA_index /path/to/trimmed.fq > rRNA.sam
```

Alignment to reference transcriptome and genome

```
tophat --bowtie1 \
        --solexa-quals \
        --no-novel-juncs \
        --library-type fr-unstranded \
        --GTF /path/to/GTF \
        –o /path/to/alignments \
        /path/to/genome_index \
        /path/to/ribo-rRNA.fq
```

Alignment post-processing

```
samtools view –b –q 10 /path/to/alignments/accepted_hits.bam > filtered_hits.bam

samtools index filtered_hits.bam
```

A.5 Example Analysis

Test data

The test data available on the SPECtre GitHub repository consists of human SH-SY5Y neuroblastoma ribosome profiling (treated with cycloheximide) sequence alignments limited to

the Ensembl v78 human chromosome 3 genome and transcriptome references. Likewise, the annotation database is limited to human chromosome 3 for testing purposes.

Test analysis

```
python /path/to/SPECTre.py \
        --input /path/to/test.bam \
        --output /path/to/spectre_test.txt \
        --log /path/to/spectre_test.log \
        --gtf /path/to/Homo_sapiens.GRCh38.78.test.gtf \
        --fpkm /path/to/isoforms.fpkm_tracking \
        --len 30 \
        --fdr 0.05 \
        --min 3.0 \
        --type mean \
        --floss \
        --orfscore
```

Cluster script

For faster runtime, SPECtre may be parallelized and submitted to a compute cluster. A sample PBS script is provided below:

```
#!/bin/bash

#PBS -N spectre_test
#PBS -l nodes=8,mem=32gb,walltime=96:00:00
#PBS -m abe
#PBS -M stonyc\@umich.edu
#PBS -d .
#PBS -V

#PBS -o spectre_test.out
#PBS -e spectre_test.err

python /path/to/SPECTre.py \
--input /path/to/test.bam \
--output /path/to/spectre_test.txt \
--log /path/to/spectre_test.log \
--gtf /path/to/Homo_sapiens.GRCh38.78.test.gtf \
--fpkm /path/to/isoforms.fpkm_tracking \
--nt 8 \ # use up to 8 processors
--len 30 \
--fdr 0.05 \
--min 3.0 \
--type mean \
--floss \
--orfscore
```

Weighted bias sampling

For increasing bias, *b*, RPFs aligned to ACTB are sampled, converted into a normalized coverage vector, and then scored using SPECtre. The R script for a single bias is shown below, which can be run in parallel with other biases for faster runtime and efficiency:

```
# LOAD READS INTO R:
reads <- read.delim("/dir/to/ACTB_reads.txt", stringsAsFactors=FALSE)

read_coverage <- function(positions.vector) {
        coverage <- rep(0, times=1128)
        coverage.table <- table(positions.vector)
        positions = as.numeric(names(coverage.table))
        depth = as.vector(coverage.table)
        for (i in 1:length(positions)) {
                coverage[positions[i]] <- depth[i]
        }
        return(coverage)
}

normalized_coverage <- function(cov) {
        return(cov/max(cov))
}

roundup <- function(x, to=3) {
        to*(x%/%to + as.logical(x%%to))
        }

calculate_spectre_score <- function(normalized.coverage, window.size, step.size) {
        coherences <- numeric()
        coding.coverage <- rep(c(4/6,1/6,1/6),
        times=roundup(length(normalized.coverage)/3))[1:length(normalized.coverage)]
        for (i in seq(1, length(normalized.coverage)-30, 3)) {
                j = i + window.size
                if (sum(normalized.coverage[i:j]) == 0 || is.na(sum(normalized.coverage[i:j]))) {
                        coherences <- c(coherences, 0.0)
                } else {
                        test.spec <- spec.pgram(data.frame(normalized.coverage[i:j],
                                coding.coverage[i:j]), spans=c(3,3), plot=FALSE)
                        coherences <- c(coherences, test.spec$coh[which(abs(test.spec$freq-1/3) ==
                                min(abs(test.spec$freq-1/3)))])
                }
        }
        return(mean(coherences))
}

sample_reads <- function(df, sample.size, weight) {
        reads <- data.frame(name=rep(NA, times=sample.size), len=rep(NA, times=sample.size), pos=rep(NA,
times=sample.size))
        names <- sample(df$read, size=sample.size, replace=TRUE, prob=df$p/(df$p^weight))
        for (i in 1:length(names)) {
                read.name <- unlist(strsplit(names[i], split="\\|"))[1]
                read.len <- as.numeric(unlist(strsplit(names[i], split="\\|"))[2])
                read.pos <- as.numeric(unlist(strsplit(names[i], split="\\|"))[3])
                reads[i,"name"] <- read.name
                reads[i,"len"] <- read.len
                reads[i,"pos"] <- read.pos
        }
        return(reads)
}
```

```
reads.dist <- data.frame(c(18:40))
colnames(reads.dist) <- c("len")
# CALCULATE DISTRIBUTION OF READ LENGTHS:
reads.dist$num <- NA
reads.dist$ratio <- NA
for (i in 1:length(reads.dist[,1])) {
        reads.dist[i,"num"] <- length(reads$len[reads$len==reads.dist[i,"len"]])
        reads.dist[i,"ratio"] <- length(reads$len[reads$len==reads.dist[i,"len"]])/length(reads$len)
}

# INITIALIZE THE SCORING MATRIX:
scores <- numeric()
weight <- 1.0

# CALCULATE SPECTRE SCORE FOR VARIABLY BIASED SAMPLING OF READ LENGTHS OVER ACTB:
n.trials = 10000 # Number of trials.
n.sample = 10000 # Sample size.
for (i in 1:n.trials) {
        sampled.reads <- sample_reads(reads, n.sample, weight)
        sampled.coverage <- read_coverage(sampled.reads$pos)
        sampled.coverage.nlz <- normalized_coverage(sampled.coverage)
        sampled.spec <- calculate_spectre_score(sampled.coverage.nlz, 30, 3)
        scores[i] <- sampled.spec
        print(paste(Sys.time(), paste(weight, i, sep=": "), sep=" "))
}
```

A.6 Data Access

mESC, GSE53693[95]

Zebrafish, GSE60095, sample GSM1464901[99]

HEK293, https://ohlerlab.mdc-berlin.de/files/RiboTaper/alignment_files.tar.gz[100]

A.7 Usage and Implementation

Usage

The files required for SPECtre analysis are an indexed alignment file in BAM format, an isoform-level expression tracking file output from Cufflinks, and a transcript annotation file in the form of

a Gene Transfer Format (GTF, version 2.2+) file.[58] GTF annotation files may be downloaded from the UCSC Genome Browser or the Ensembl archive. User-defined arguments to specify the SPECtre scoring method (mean, median, maximum, etc.), the length of the windows over which to calculate the spectral coherence, minimum FPKM cutoffs, and FDR thresholds to calculate the Bayesian posterior probability of translation for each transcript are provided. Implementations of the FLOSS metric and ORFscore have also been made available as optional command-line arguments. Finally, an option to calculate the un-windowed spectral coherence of the full length of a transcript has been provided.

Output

Depending on the detail requested, the end-user will be provided with a tab-delimited text document with annotation information relevant to each transcript tested, including a unique identifier, genomic coordinates of the CDS and UTR regions, transcript abundance, the normalized read coverage over each region, the user-defined spectral coherence metric, and the Bayesian posterior probability of each transcript to be classified as actively translated. Optionally, the respective fragment length distribution and FLOSS metric, and the total number of reads over each frame and ORFscore may be calculated and output for each transcript. The spectral coherence score distribution for translated versus non-translated transcripts, and summary ROC and AUC plots are generated for user review. All plots generated by the SPECtre analytical package are output in PDF format.

Implementation

SPECtre is a standalone analytical package written in Python, and is intended to run on a wide range of platforms. Therefore, installation of only a limited number of non-standard modules is required. SPECtre requires the following non-standard modules: NumPy and RPy2 (http://rpy.sourceforge.net), and HTSeq.[186,187] HTSeq is used to convert alignments from the BAM or SAM input into transcript coverage, hash transcript intervals, and check for overlaps. Prior installation of R and the ROCR package are required to perform the ROC analyses and generate summary plots.[188] Shell scripts and SPECtre analysis for typical single sample and multi-sample comparative analyses are available as Supplementary Material and via the SPECtre GitHub repository.

Ingolia (2014)		Bazzini (2014)		
Number	Percent	Number	Percent	Processing Step
310,854,993	n/a	1,902,337,857	n/a	Raw
268,719,953	86.4	1,746,870,724	91.8	After adapter removal and trimming
128,734,835	47.9	1,276,335,857	73.1	After alignment to rRNA database
122,290,796	95.0	1,113,387,047	88.8	After alignment to transcriptome + genome
94,225,088	77.1	591,916,525	52.2	After MAPQ filtering

Table A.1. Number of reads remaining at each stage of pre-processing, alignment and quality filtering of ribosome profiling libraries derived from mESC and zebrafish. The percentages listed are relative to the previous number of reads.

		FPKM Cutoff				
	Method	0.5	1.0	3.0	5.0	10.0
Bazzini	FLOSS	0.779	0.773	0.759	0.755	0.747
	ORFScore	0.584	0.591	0.606	0.619	0.634
	SPECtre (30nt)	0.877	0.876	0.881	0.888	0.897
	SPECtre (60nt)	0.852	0.850	0.859	0.867	0.870
	SPECtre (90nt)	0.837	0.838	0.848	0.853	0.849
Ingolia	FLOSS	0.941	0.925	0.889	0.864	0.837
	ORFScore	0.573	0.580	0.582	0.589	0.594
	SPECtre (30nt)	0.955	0.936	0.900	0.885	0.872
	SPECtre (60nt)	0.938	0.911	0.874	0.860	0.844
	SPECtre (90nt)	0.913	0.889	0.859	0.844	0.825

Table A.2. Translational classification accuracy in mESC and zebrafish. AUC for each classification algorithm at various minimum FPKM cutoffs for ribosome profiling derived from zebrafish, and mESC. In addition to the default length of 30 nt sliding windows, 60 nt and 90 nt windows were also tested.

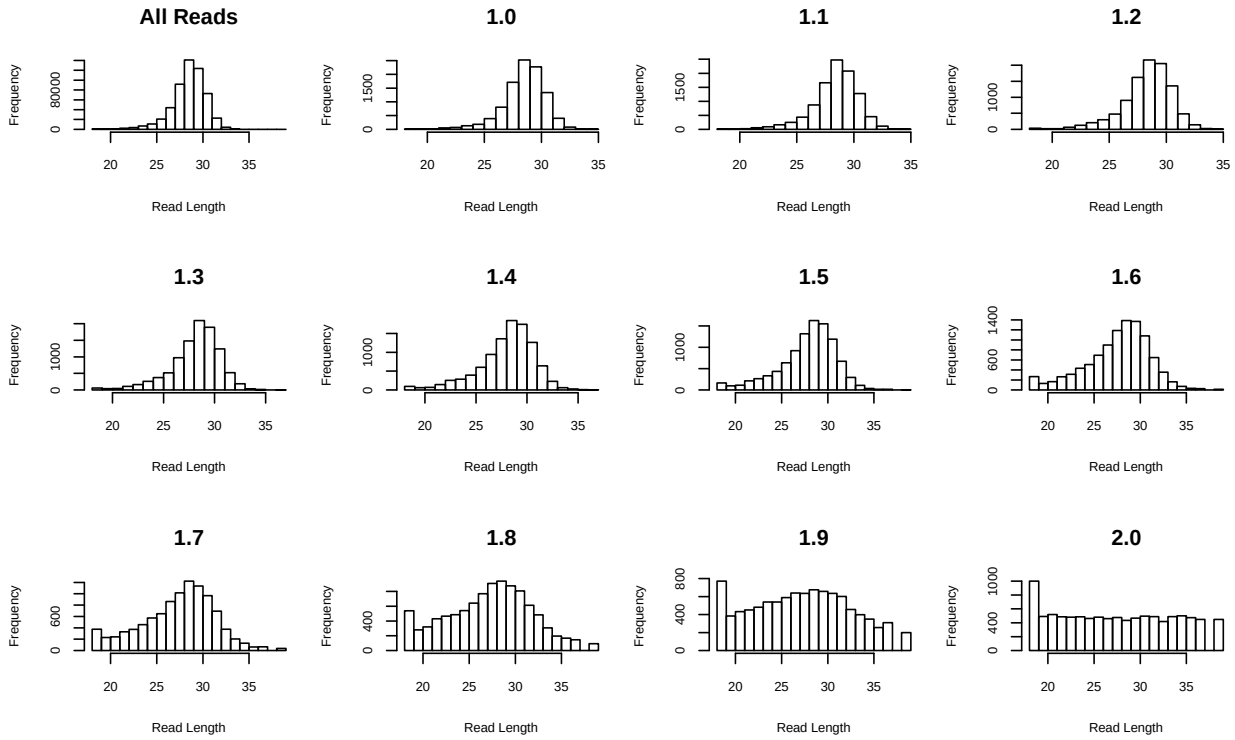

Figure A.1. Read length distribution of RPFs aligned to ACTB in mESC. The distribution of all 18 to 40 nt RPFs aligned to the housekeeping gene ACTB in ribosome profiling of mESC, top left, which depicts the enrichment of 28-30 nt reads indicative of RPFs protected by ribosomes during cycloheximide treatment. Also shown are example read length distribution profiles at various weighted biases (1.0 to 2.0) after sampling 10,000 RPFs from ACTB. As the weighted bias increases from 1.0 to 2.0, the distribution of read lengths increases in variance relative to the experimental read length distribution profile ("All Reads") and adopts a progressively uniform distribution.

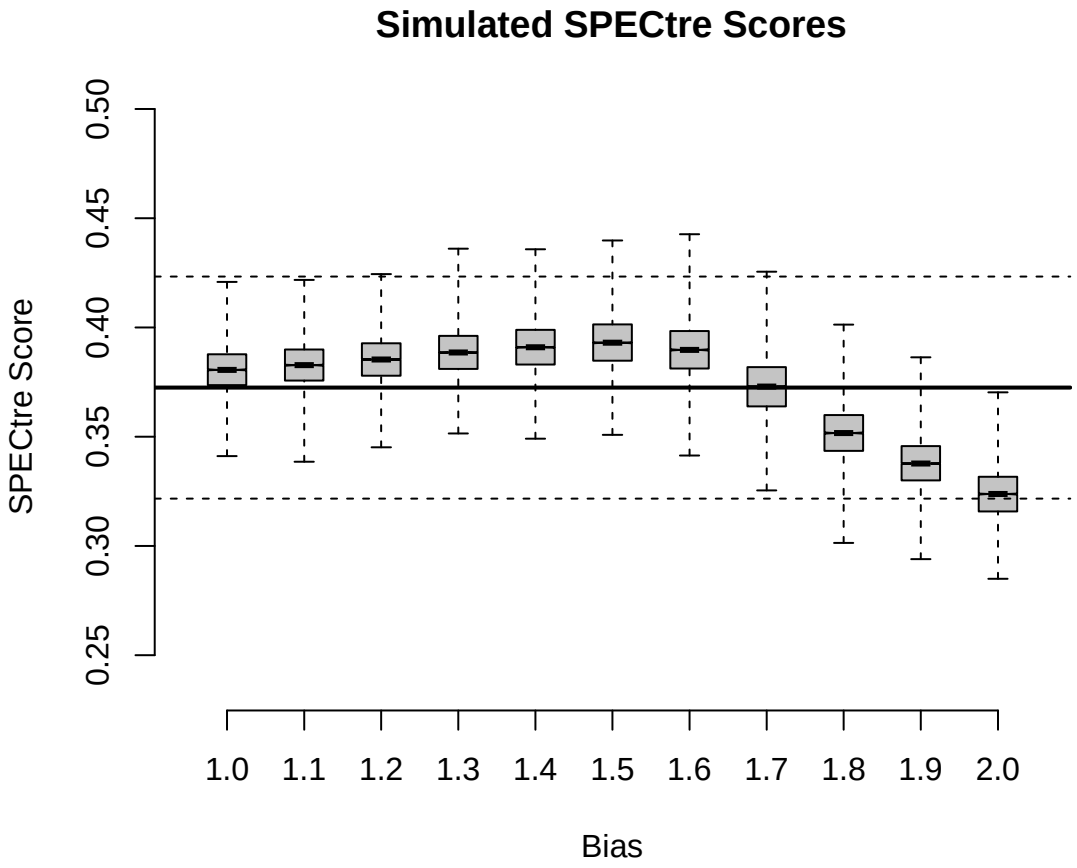

Figure A.2. Distribution of SPECtre scores over ACTB after weighted re-sampling. Re-sampling analysis of 10,000 RPFs with various weighted biases (from 1.0 to 2.0). Dark line inside boxes denotes the median of the SPECtre scores at each weighted bias level, and whiskers depict the minimum and maximum scores. Solid horizontal line denotes the mean SPECtre score over all sampling simulations. The dashed, horizontal lines depict the extreme outlier boundaries as defined by Tukey, which is 1.5 times the IQR.

APPENDIX B

SUPPLEMENTAL MATERIAL FOR CHAPTER 3

A.1 Materials and Methods

Alignment to the human genome and transcriptome (GRCh38 Ensembl version 78)

Ribosome profiling and mRNA-Seq reads were trimmed of adapters, and then by quality using *fasqt-mcf* from the *ea-utils* package (https://github.com/ExpressionAnalysis/ea-utils). Ribosome profiling and mRNA-Seq reads in FASTQ format were trimmed based on quality if four consecutive nucleotides were observed with Phred scores of 10 or below. The minimum read length required after trimming was 25 nucleotides. Trimmed sequences were then aligned to a ribosomal RNA sequence index using Bowtie v1.1.2 to deplete them of contaminant sequences.[62] Alignment to the rRNA sequence contaminant index was performed using the following parameters: seed alignment length of 22 nucleotides, no mismatches in the seed alignment were allowed, with the unmapped reads written to a separate FASTQ file.

```
bowtie -l 22 -n 0 -S --un /path/to/depleted_reads.fq \
        /path/to/rRNA_index \
        /path/to/trimmed_reads.fq
```

Ribosome profiling and mRNA-Seq reads depleted of rRNA contaminant sequences were aligned to the human genome and transcriptome using TopHat v.2.0.0.[64,183] The trimmed and rRNA-depleted reads were aligned to the human genome and transcriptome with the parameters

specifying standard Illumina reads, with un-gapped Bowtie 1 alignment (using a seed alignment length of 22 nucleotides, with no mismatches in the seed alignment allowed), to annotated junctions only, using Solexa quality scores:

```
tophat2 -p 4 –bowtie1 \
        --no-novel-juncs \
        --library-type fr-unstranded \
        --solexa-quals \
        -G /path/to/ensembl.gtf \
        /path/to/bowtie_index \
        /path/to/depleted_reads.fq
```

Sequence alignment quality filtering and merging

Ribosome profiling and mRNA-Seq reads aligned to the human genome and transcriptome by TopHat2 were output to BAM format, and then sorted by chromosomal coordinate. Reads were then filtered by mapping quality using SAMTools; read alignments were required to have minimum mapping quality of 10, or higher, to be retained for subsequent analyses. Unique read group identifiers were assigned to each technical and biological replicate, and then the alignments were merged by technical replicates and subsequently as biological replicates using Picard (http://broadinstitute.github.io/picard/).[184]

Metagene profile generation and alignment offset calculation

For counting reads over transcript isoforms, metagene profiles were generated from the Ensembl (version 78) transcript annotation database using Plastid.[93] A- and P-site offsets for harringtonine and cycloheximide ribosome profiling reads, respectively, were determined by pooling all reads that overlapped canonical AUG translation initiation start sites from annotated

protein-coding genes. The most common (mode) distance from the 5’ ends of reads of a given length to the position of the AUG in those reads was accepted as the A- or P-site offset distance.

Calculation of transcript abundance

Read counts over each transcript isoform, or region (5’UTR, CDS, and 3’UTR), were normalized by length, summed, and reported as transcripts per million as described previously.[68] At the time of analysis, Cufflinks was required for initial transcript quality control, and was run with the following parameters:[58]

```
cufflinks -p 8 -o /path/to/output \
        -G /path/to/ensemble.gtf \
        /path/to/tophat/alignments
```

SPECtre analysis of transcripts in non-differentiated and RA-differentiated libraries

SPECtre profiling measures the strength of the tri-nucleotide periodicity inherent to the alignment of ribosome protected fragments to protein-coding genes in a reference transcriptome.[96] SPECtre analysis was applied in two stages: 1) to score the translational potential of annotated transcripts to build a background protein-coding reference model, and 2) to score the translational potential of predicted upstream-initiation ORFs. In this way, the translational potential of predicted upstream and overlapping ORFs are score against a background model of translation derived from annotated protein-coding transcripts. Annotated protein-coding transcripts were profiled by SPECtre using the following parameters:

```
python /path/to/SPECtre.py \
        --input /path/to/tophat/alignments \
        --output /path/to/output \
        --log /path/to/logfile \
        --gtf /path/to/ensemble.gtf \
        --fpkm /path/to/cufflinks/isoforms.fpkm_tracking \
```

```
        --len 30 \
        --fdr 0.05 \
        --min 3.0 \
        --nt 8 \
        --type mean \
        --target <chromosome_id>
```

Where the minimum FPKM required for a transcript to be considered as translated for generation of the background model was specified as 3.0, and the length of the sliding SPECtre windows was set to encompass 30 nucleotides. The SPECtre score for a transcript was defined as the mean of the scores over these sliding windows, and a 5% false discovery rate was established to set the minimum SPECtre translational score threshold. In addition, SPECtre profiling was split by chromosome to speed computation, and the results were merged afterwards using a custom Python script. Finally, prior to analysis of predicted upstream-initiated ORFs by SPECtre profiling, the minimum SPECtre translational threshold was re-calculated using TPM instead of FPKM using a minimum cutoff of 10 transcripts per million.

Computational prediction of upstream-initiated open reading frames

Open reading frames were computationally predicted from annotated 5'UTR sequences (Ensembl, version 78) using AUG, and near-cognate non-AUG translation initiation site sequences. Open reading frame sequences were generated based on these predicted initiation site sequences and read through to the first in-frame termination codon encountered in the annotated CDS. These predicted ORFs were then used to generate coordinates over which they would be profiled and scored by SPECtre. Identical parameters to the annotated transcript SPECtre analysis were employed for consistency across analyses:

```
python /path/to/SPECtre-uORFs.py \
        --input /path/to/alignments \
        --output /path/to/output \
        --results /path/to/spectre/transcript_results \
        --log /path/to/logfile \
```

```
--fpkm /path/to/cufflinks/isoforms.fpkm_tracking \
--len 30 \
--fdr 0.05 \
--min 3.0 \
--nt 8 \
--type mean \
--target <chromosome_id>
```

Three alternative inputs are required for the SPECtre analysis of predicted ORFs: 1) the annotated transcript GTF database was not required and removed, 2) the results of the annotated transcript analysis, and 3) a genomic sequence file in FASTA format. The results of the annotated transcript analysis were used to identify the set of transcripts from which to predict upstream-initiated ORFs, and the FASTA sequence file was used to generated the ORF sequences for output.

Supplemental annotation of non-AUG translation initiation sites

Upstream sequences of predicted non-AUG translation initiation sites were examined for possible in-frame AUG initiation start sites; 5'UTR sequences of predicted non-AUG sites were extracted, and then searched for the presence of in-frame AUG sites. These non-AUG sites were then re-annotated according to the proximity of upstream AUG initiation sites: those with an in-frame AUG site within 30 nucleotides of the predicted TIS, and those with an in-frame AUG site in-frame, but beyond 30 nucleotides upstream of the predicted site.

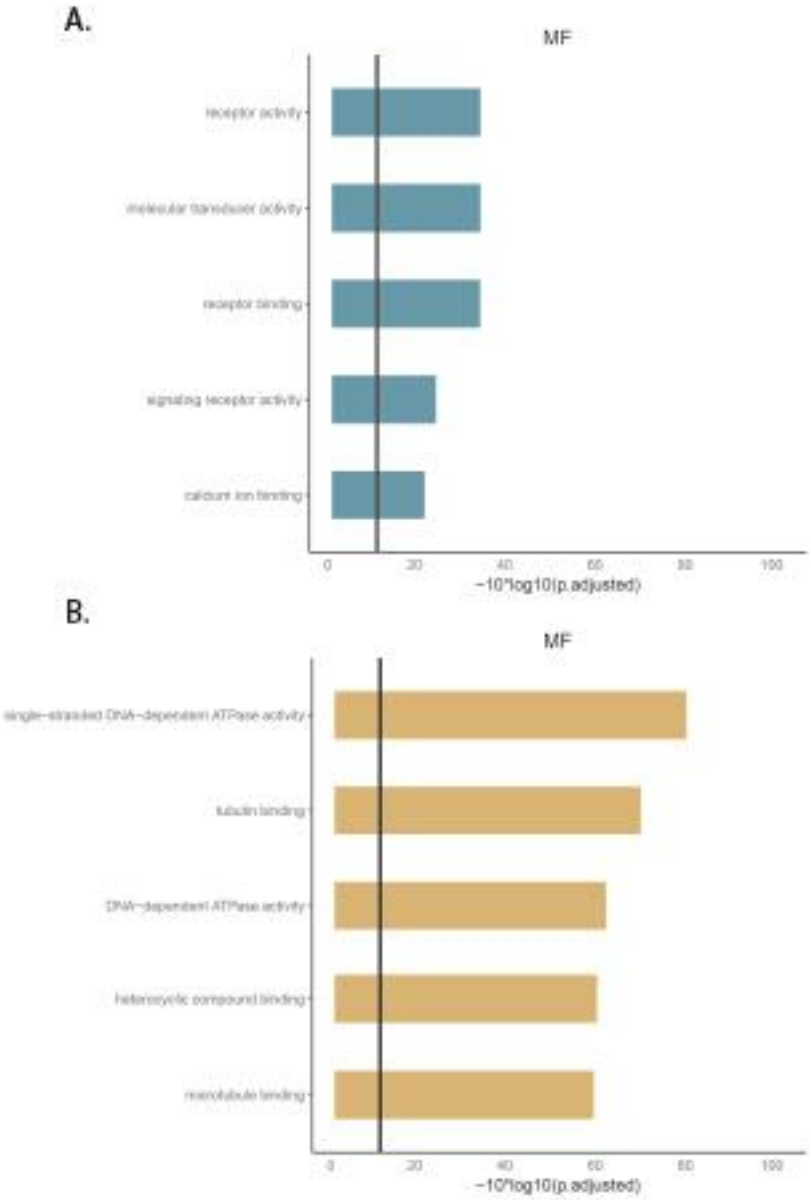

Figure B.1 Molecular function gene set enrichment based on mRNA rank-change analysis. A) Up-regulated gene sets in RA differentiated include those related to receptor binding, and signaling. B) Down-regulated gene sets include those related to microtubule binding and ATPase activity.

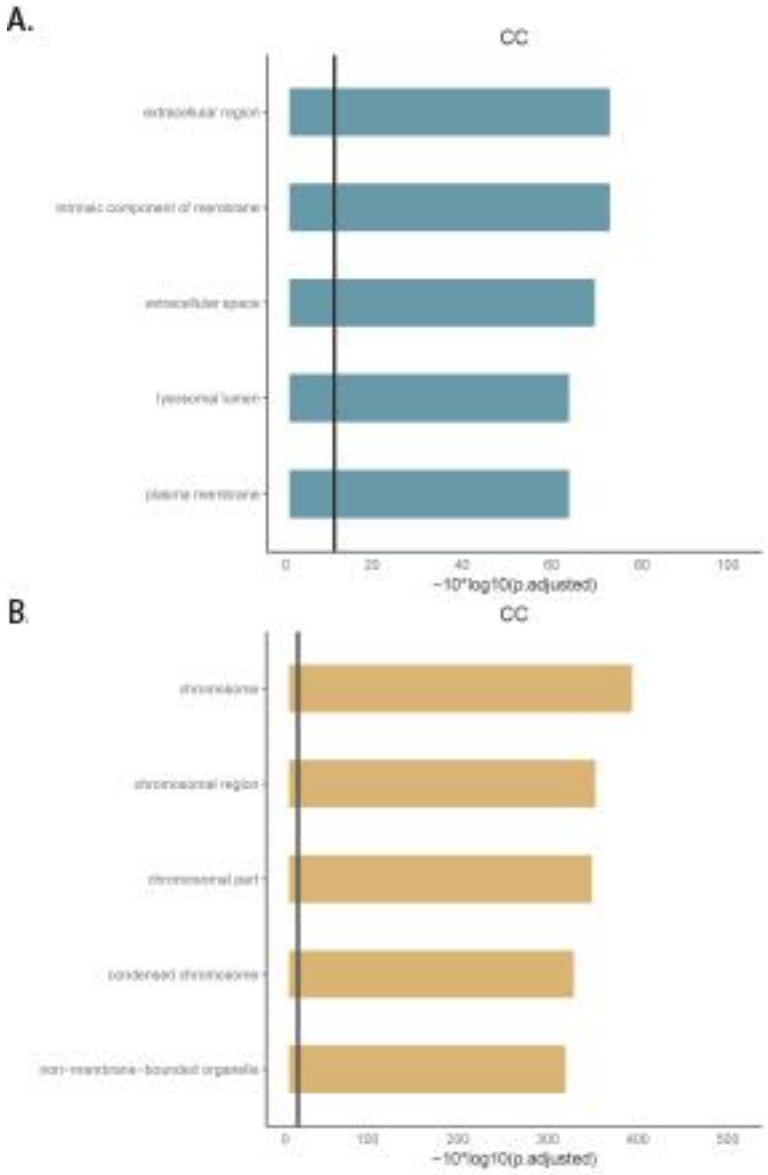

Figure B.2 Cellular component gene set enrichment based on mRNA rank-change analysis. A) Enriched terms in up-regulated gene sets include those related to extracellular space. B) Down-regulated gene sets include terms related to chromosome structure.

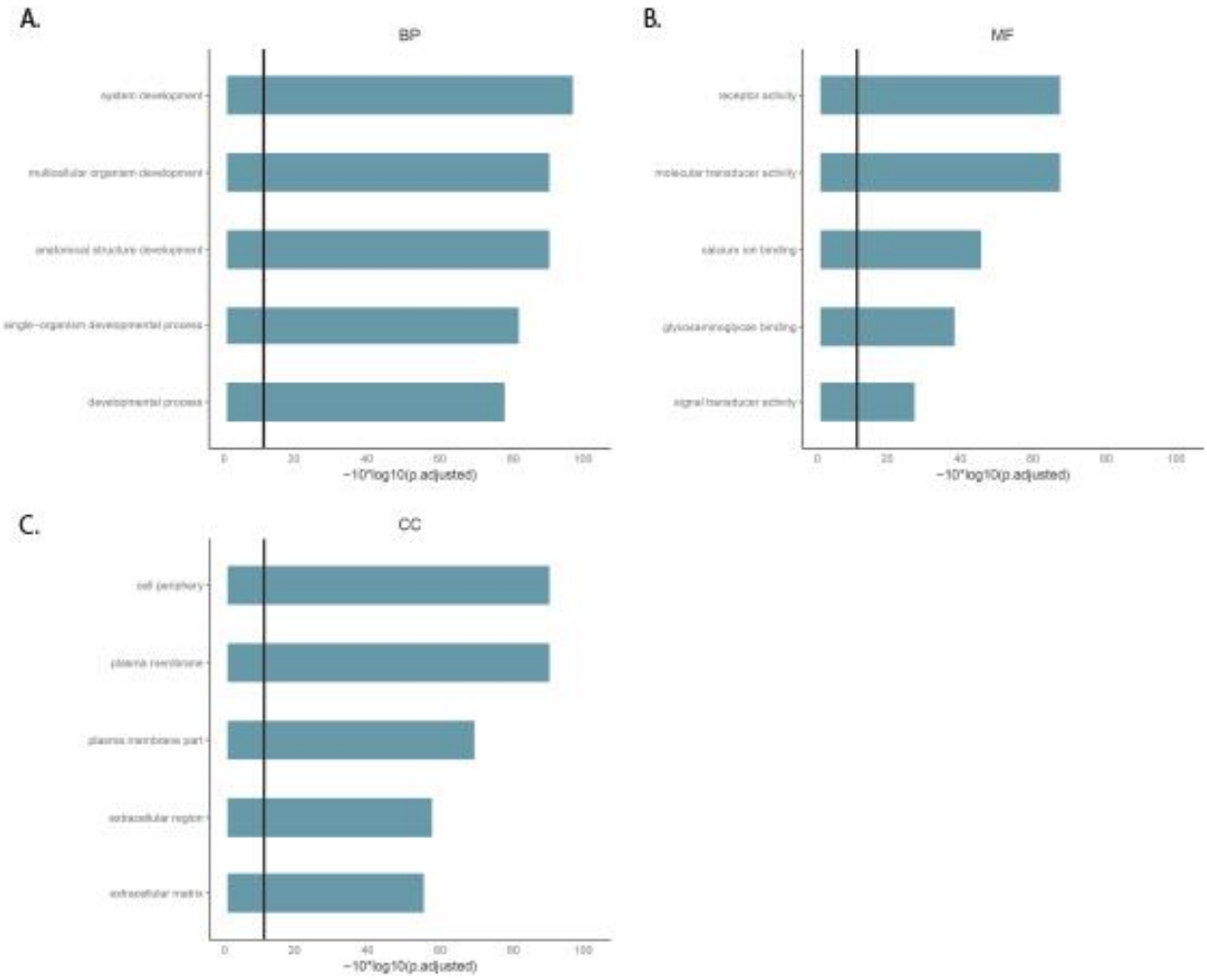

Figure B.3 Enrichment of up-regulated gene sets based on DE analysis of RPF counts. A) Enriched biological processes include those related to development. B) Signal transduction and receptor activity gene sets up-regulated in RA differentiated cells. C) Extracellular components are enriched in RA differentiated cells compared to non-differentiated cells.

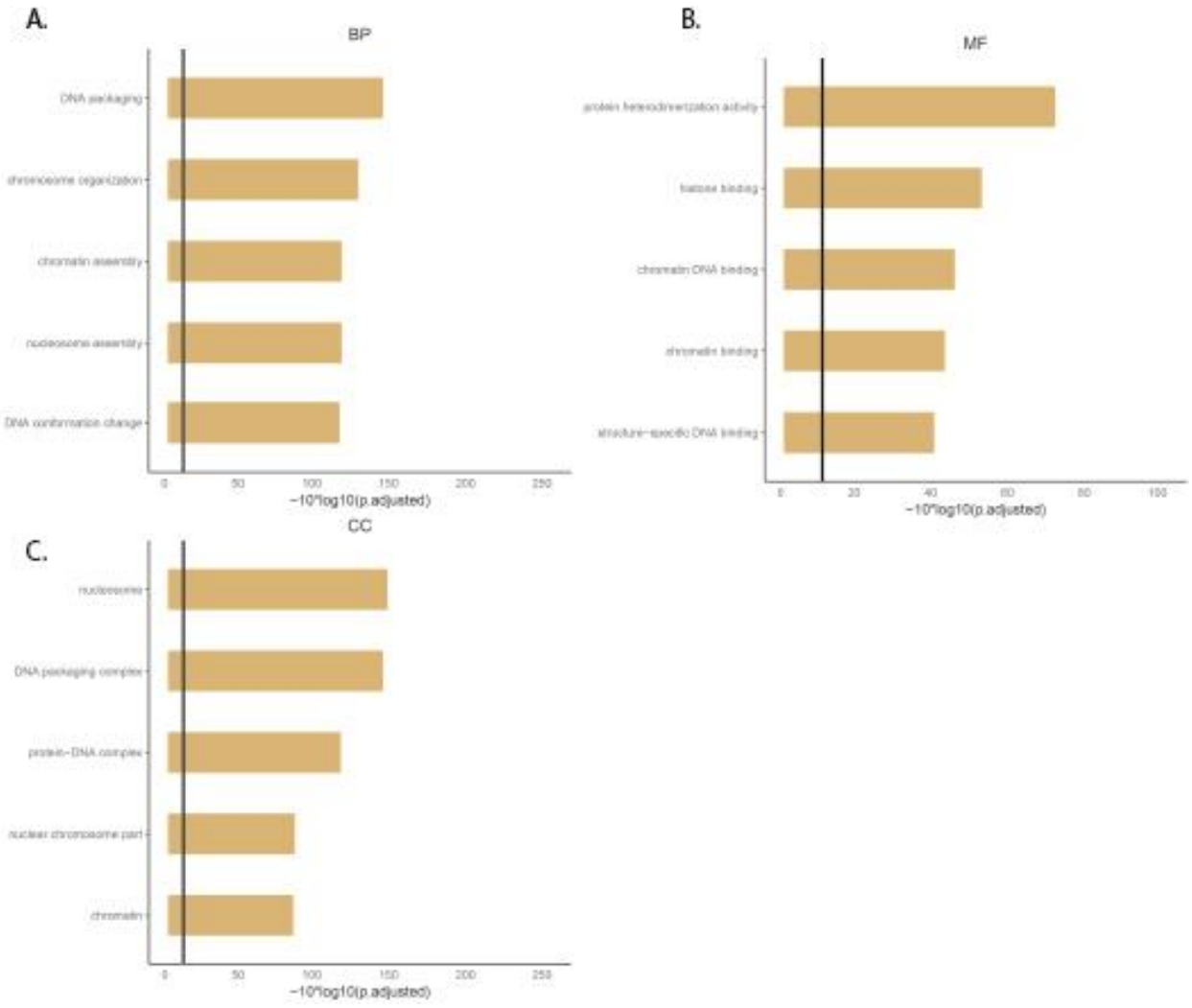

Figure B.4 Enrichment of down-regulated gene sets based on DE analysis of RPF counts. A) Biological processes are enriched in terms related to chromosome structure and assembly. B) Molecular function terms are enriched for chromatin and structural binding. C) Cellular component terms are enriched for chromosome structure and DNA packaging.

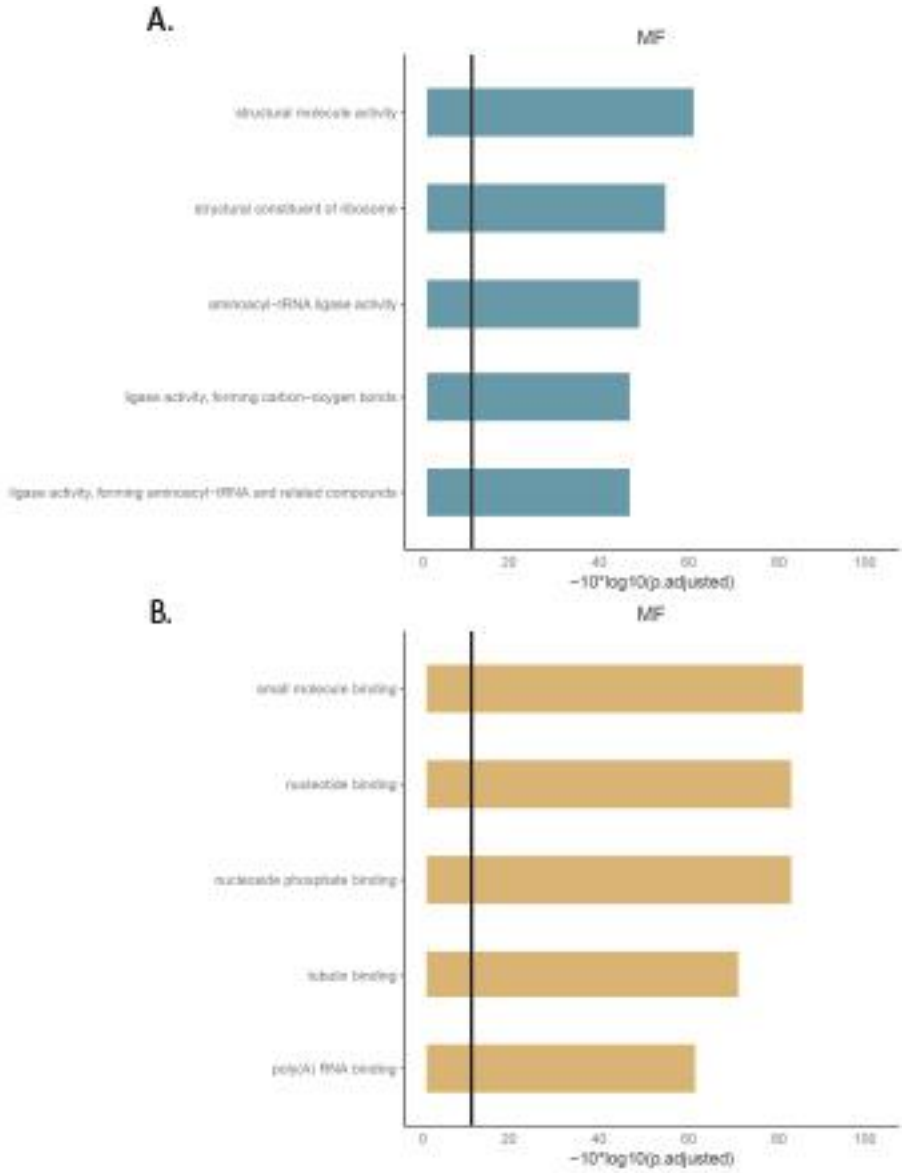

Figure B.5 Molecular function gene set enrichment based on translational efficiency. A) Molecular function terms are enriched for structural components of the ribosome and tRNA binding. B) Down-regulated molecular function gene sets are enriched for small molecule and structural binding terms.

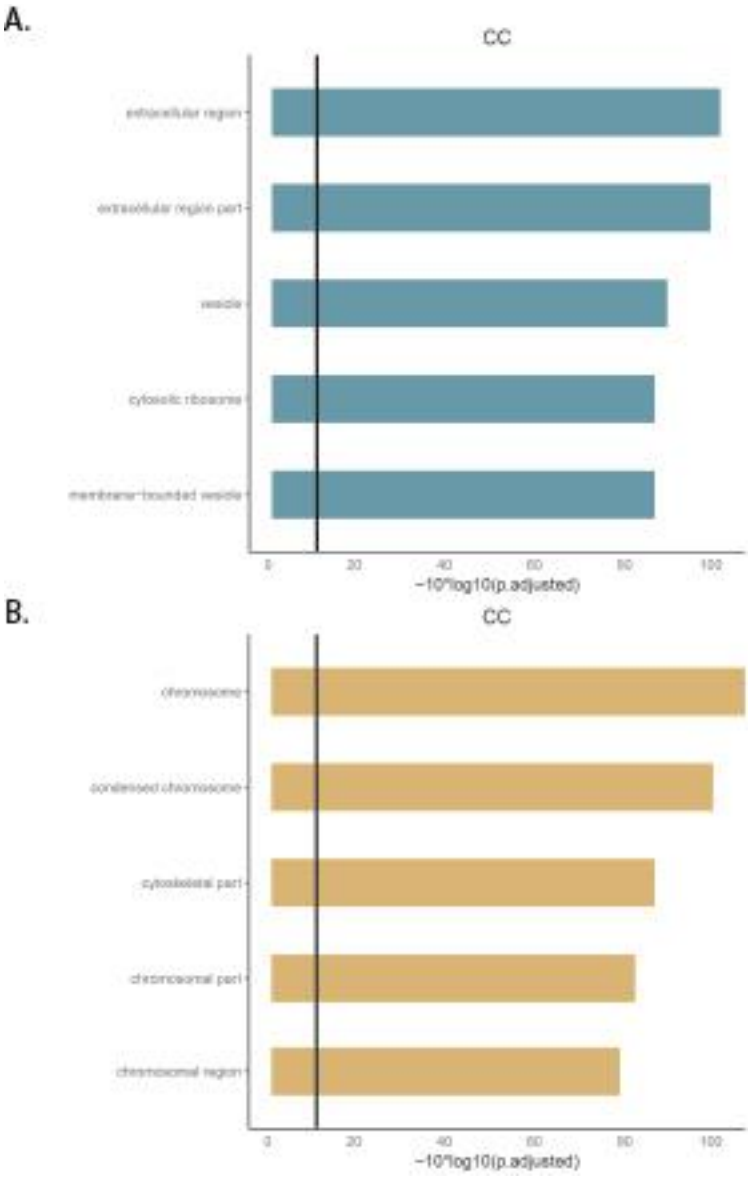

Figure B.6 Cellular component gene set enrichment based on translational efficiency. A) Up-regulated gene sets are enriched for terms relate to extracellular space. B) Down-regulated gene sets are enriched for terms related to chromosome structure.

gene_id	transcript_id	gene_name	uorf_id	dist	mean_residual
ENSG00000203814	ENST00000545683	HIST2H2BF	ENST00000545683.1	-138	214.009660627775
ENSG00000164919	ENST00000520468	COX6C	ENST00000520468.56	-75	7.48415663679263
ENSG00000131174	ENST00000481445	COX7B	ENST00000481445.12	-128	6.05530073370224
ENSG00000155090	ENST00000285407	KLF10	ENST00000285407.3	-300	4.6977592616836
ENSG00000133639	ENST00000256015	BTG1	ENST00000256015.27	-238	4.37266835988406
ENSG00000204954	ENST00000549478	C12orf73	ENST00000549478.43	-166	4.1335802850581
ENSG00000135535	ENST00000413644	CD164	ENST00000413644.5	-250	4.0339389968682
ENSG00000124596	ENST00000373154	OARD1	ENST00000373154.18	-263	3.94226656417065
ENSG00000109099	ENST00000312280	PMP22	ENST00000312280.20	-110	3.75254432924282
ENSG00000166482	ENST00000395592	MFAP4	ENST00000395592.5	-290	3.50409809285365
ENSG00000111897	ENST00000339697	SERINC1	ENST00000339697.8	-92	3.05084168188823
ENSG00000220205	ENST00000316509	VAMP2	ENST00000316509.2	-234	2.7827129178595
ENSG00000074695	ENST00000251047	LMAN1	ENST00000251047.80	-247	2.78027152995272
ENSG00000132406	ENST00000254742	TMEM128	ENST00000254742.75	-115	2.7544412241449
ENSG00000115128	ENST00000233468	SF3B6	ENST00000233468.18	-47	2.72897736886105
ENSG00000169020	ENST00000304312	ATP5I	ENST00000304312.11	-101	2.67511502668913
ENSG00000165389	ENST00000298130	SPTSSA	ENST00000298130.13	-203	2.46326924453185
ENSG00000052723	ENST00000369528	SIKE1	ENST00000369528.6	-196	2.36628068847243
ENSG00000006468	ENST00000420159	ETV1	ENST00000420159.40	-155	2.33053818614439
ENSG00000127540	ENST00000591899	UQCR11	ENST00000591899.2	-158	2.30110461157947
ENSG00000213420	ENST00000292377	GPC2	ENST00000292377.6	-170	2.20948360358882
ENSG00000168002	ENST00000301788	POLR2G	ENST00000301788.10	-138	2.06215949129986
ENSG00000115128	ENST00000233468	SF3B6	ENST00000233468.20	-165	2.02599419353475
ENSG00000189043	ENST00000339600	NDUFA4	ENST00000339600.22	-46	1.92194154519564
ENSG00000129055	ENST00000510994	ANAPC13	ENST00000510994.101	-104	1.91685522041752
ENSG00000085063	ENST00000351554	CD59	ENST00000351554.2	-186	1.90037524822165
ENSG00000179085	ENST00000368400	DPM3	ENST00000368400.3	-196	1.86546022606256
ENSG00000079950	ENST00000367941	STX7	ENST00000367941.14	-180	1.86084708595664
ENSG00000271601	ENST00000604000	LIX1L	ENST00000604000.7	-295	1.78284169077653
ENSG00000205208	ENST00000379205	C4orf46	ENST00000379205.26	-215	1.76679586370253
ENSG00000111237	ENST00000447578	VPS29	ENST00000447578.54	-227	1.53321434735956
ENSG00000128609	ENST00000355749	NDUFA5	ENST00000355749.57	-199	1.52085271342457
ENSG00000079150	ENST00000424785	FKBP7	ENST00000424785.4	-105	1.50839437465663
ENSG00000060762	ENST00000621630	MPC1	ENST00000621630.11	-265	1.3516320650817
ENSG00000154582	ENST00000520210	TCEB1	ENST00000520210.15	-113	1.34924892563321
ENSG00000006468	ENST00000403685	ETV1	ENST00000403685.26	-207	1.33192548256211
ENSG00000124172	ENST00000243997	ATP5E	ENST00000243997.8	-107	1.24867441318077
ENSG00000091136	ENST00000222399	LAMB1	ENST00000222399.4	-269	1.1749046682958
ENSG00000176454	ENST00000617710	LPCAT4	ENST00000617710.55	-232	1.10552942043999
ENSG00000134825	ENST00000537328	TMEM258	ENST00000537328.3	-176	1.0859091551629
ENSG00000163191	ENST00000271638	S100A11	ENST00000271638.2	-183	1.02524457965488
ENSG00000155957	ENST00000556010	TMBIM4	ENST00000556010.3	-179	1.02038839594092
ENSG00000196628	ENST00000354452	TCF4	ENST00000354452.67	-152	1.003257734837
ENSG00000127452	ENST00000247977	FBXL12	ENST00000247977.19	-171	0.985331204079753
ENSG00000117122	ENST00000375535	MFAP2	ENST00000375535.19	-65	0.982748606719541

ENSG00000184047	ENST00000443649	DIABLO	ENST00000443649.70	-115	0.981005829474638
ENSG00000125827	ENST00000246024	TMX4	ENST00000246024.17	-274	0.912146644073427
ENSG00000132432	ENST00000352861	SEC61G	ENST00000352861.1	-185	0.911388696044815
ENSG00000085832	ENST00000371733	EPS15	ENST00000371733.1	-277	0.906480145192362
ENSG00000150093	ENST00000396033	ITGB1	ENST00000396033.2	-85	0.875993375496484
ENSG00000198668	ENST00000544280	CALM1	ENST00000544280.24	-35	0.844908638059924
ENSG00000105438	ENST00000330720	KDELR1	ENST00000330720.3	-190	0.830959779619959
ENSG00000151552	ENST00000281243	QDPR	ENST00000281243.3	-169	0.8234164379142
ENSG00000100906	ENST00000216797	NFKBIA	ENST00000216797.1	-254	0.804658104953933
ENSG00000244754	ENST00000357505	N4BP2L2	ENST00000357505.24	-91	0.797754000974177
ENSG00000164054	ENST00000612611	SHISA5	ENST00000612611.40	-186	0.790012045975735
ENSG00000166228	ENST00000299299	PCBD1	ENST00000299299.18	-174	0.757990861931869
ENSG00000196628	ENST00000537578	TCF4	ENST00000537578.70	-92	0.738709289790308
ENSG00000130731	ENST00000614890	C16orf13	ENST00000614890.3	-229	0.73678431471042
ENSG00000025796	ENST00000369002	SEC63	ENST00000369002.16	-233	0.735344058921464
ENSG00000114503	ENST00000321256	NCBP2	ENST00000321256.4	-83	0.733913690913015
ENSG00000168374	ENST00000303436	ARF4	ENST00000303436.17	-112	0.727015552079459
ENSG00000133983	ENST00000389912	COX16	ENST00000389912.14	-76	0.676840640464102
ENSG00000136240	ENST00000258739	KDELR2	ENST00000258739.8	-179	0.65800456276373
ENSG00000075142	ENST00000394641	SRI	ENST00000394641.1	-82	0.609081354350244
ENSG00000113575	ENST00000481195	PPP2CA	ENST00000481195.21	-107	0.597347838969469
ENSG00000127922	ENST00000248566	SHFM1	ENST00000248566.1	-95	0.564001102458755
ENSG00000155380	ENST00000538576	SLC16A1	ENST00000538576.98	-245	0.546894116837497
ENSG00000110090	ENST00000376618	CPT1A	ENST00000376618.8	-69	0.536384072657107
ENSG00000105223	ENST00000409281	PLD3	ENST00000409281.25	-60	0.507684483232725
ENSG00000001630	ENST00000450723	CYP51A1	ENST00000450723.26	-194	0.486296029848002
ENSG00000134153	ENST00000256545	EMC7	ENST00000256545.8	-263	0.454876567454237
ENSG00000138069	ENST00000398529	RAB1A	ENST00000398529.22	-49	0.430983967101631
ENSG00000132388	ENST00000396981	UBE2G1	ENST00000396981.2	-88	0.409590512755403
ENSG00000140307	ENST00000396060	GTF2A2	ENST00000396060.16	-57	0.357284091058508
ENSG00000185088	ENST00000330964	RPS27L	ENST00000330964.35	-257	0.324956326333918
ENSG00000131238	ENST00000433473	PPT1	ENST00000433473.59	-164	0.315578296532823
ENSG00000119655	ENST00000555619	NPC2	ENST00000555619.5	-204	0.300166259008394
ENSG00000154723	ENST00000284971	ATP5J	ENST00000284971.4	-185	0.294405818784438
ENSG00000063241	ENST00000438389	ISOC2	ENST00000438389.79	-245	0.289505438369753
ENSG00000147255	ENST00000370910	IGSF1	ENST00000370910.13	-194	0.265933110942999
ENSG00000135404	ENST00000420846	CD63	ENST00000420846.2	-85	0.262074002549052
ENSG00000172270	ENST00000545507	BSG	ENST00000545507.44	-129	0.246795080303511
ENSG00000112514	ENST00000374496	CUTA	ENST00000374496.30	-211	0.229477885271077
ENSG00000109065	ENST00000357814	NAT9	ENST00000357814.1	-275	0.21946453747034
ENSG00000035862	ENST00000262768	TIMP2	ENST00000262768.12	-155	0.183231194207834
ENSG00000146425	ENST00000367089	DYNLT1	ENST00000367089.1	-96	0.168063833021454
ENSG00000104763	ENST00000262097	ASAH1	ENST00000262097.10	-298	0.167143529270821
ENSG00000170540	ENST00000304414	ARL6IP1	ENST00000304414.12	-101	0.162134795098452
ENSG00000182117	ENST00000328848	NOP10	ENST00000328848.2	-62	0.157873730955841
ENSG00000183010	ENST00000329875	PYCR1	ENST00000329875.1	-91	0.147765411511009

ENSG00000116791	ENST00000370871	CRYZ	ENST00000370871.9	-179	0.146769664775778
ENSG00000123933	ENST00000337190	MXD4	ENST00000337190.17	-223	0.139839506501833
ENSG00000100219	ENST00000611155	XBP1	ENST00000611155.1	-290	0.118515717745031
ENSG00000100528	ENST00000216416	CNIH1	ENST00000216416.3	-273	0.10299321646171
ENSG00000137876	ENST00000260443	RSL24D1	ENST00000260443.12	-57	0.101922613166617
ENSG00000127184	ENST00000247655	COX7C	ENST00000247655.4	-47	0.0990643451928337
ENSG00000143158	ENST00000271373	MPC2	ENST00000271373.2	-227	0.0953577701619
ENSG00000118181	ENST00000527673	RPS25	ENST00000527673.41	-138	0.0930416477215578
ENSG00000148248	ENST00000545297	SURF4	ENST00000545297.1	-257	0.0817327574752148
ENSG00000137868	ENST00000323940	STRA6	ENST00000323940.25	-215	0.0807719741579371
ENSG00000141378	ENST00000393038	PTRH2	ENST00000393038.25	-201	0.0730939487486646
ENSG00000008283	ENST00000392975	CYB561	ENST00000392975.10	-239	0.0707678951588921
ENSG00000175768	ENST00000544379	TOMM5	ENST00000544379.1	-48	0.0603761698132674
ENSG00000171497	ENST00000307720	PPID	ENST00000307720.3	-245	0.054817771060998
ENSG00000004779	ENST00000007516	NDUFAB1	ENST00000007516.2	-223	0.0533917499232271
ENSG00000175334	ENST00000312175	BANF1	ENST00000312175.47	-83	0.0523633662762182
ENSG00000188486	ENST00000530167	H2AFX	ENST00000530167.4	-184	0.050800413281085
ENSG00000138071	ENST00000377982	ACTR2	ENST00000377982.9	-56	0.0311450513673084
ENSG00000132254	ENST00000423813	ARFIP2	ENST00000423813.22	-279	0.0201223202987942
ENSG00000080644	ENST00000348639	CHRNA3	ENST00000348639.27	-228	-0.0161047158214547
ENSG00000065518	ENST00000184266	NDUFB4	ENST00000184266.2	-237	-0.0236689183432249
ENSG00000197696	ENST00000360476	NMB	ENST00000360476.21	-260	-0.0280671285411302
ENSG00000138175	ENST00000260746	ARL3	ENST00000260746.3	-186	-0.0313722248469262
ENSG00000117519	ENST00000370206	CNN3	ENST00000370206.27	-109	-0.0326596808454877
ENSG00000132635	ENST00000360652	PCED1A	ENST00000360652.46	-222	-0.0611921218654977
ENSG00000132635	ENST00000356872	PCED1A	ENST00000356872.25	-213	-0.0628914166351539
ENSG00000141425	ENST00000589050	RPRD1A	ENST00000589050.18	-194	-0.0707040015394569
ENSG00000164733	ENST00000353047	CTSB	ENST00000353047.4	-236	-0.111293748701263
ENSG00000140264	ENST00000409960	SERF2	ENST00000409960.5	-47	-0.113307644512048
ENSG00000167476	ENST00000300961	JSRP1	ENST00000300961.2	-251	-0.128486638479777
ENSG00000129084	ENST00000396394	PSMA1	ENST00000396394.35	-202	-0.147381257217403
ENSG00000117519	ENST00000545882	CNN3	ENST00000545882.16	-64	-0.163556663489164
ENSG00000120889	ENST00000347739	TNFRSF10B	ENST00000347739.11	-276	-0.190360894329478
ENSG00000179348	ENST00000487848	GATA2	ENST00000487848.36	-238	-0.20776396821143
ENSG00000122566	ENST00000356674	HNRNPA2B1	ENST00000356674.1	-49	-0.209723429231841
ENSG00000182004	ENST00000414487	SNRPE	ENST00000414487.1	-112	-0.210668985977755
ENSG00000185825	ENST00000345046	BCAP31	ENST00000345046.25	-230	-0.214869710580321
ENSG00000055950	ENST00000342071	MRPL43	ENST00000342071.2	-154	-0.244590604054466
ENSG00000183648	ENST00000329559	NDUFB1	ENST00000329559.1	-236	-0.249336665428592
ENSG00000099800	ENST00000215570	TIMM13	ENST00000215570.29	-66	-0.256667910963101
ENSG00000105341	ENST00000417807	ATP5SL	ENST00000417807.2	-224	-0.273074886882529
ENSG00000108774	ENST00000346213	RAB5C	ENST00000346213.16	-200	-0.291555465067221
ENSG00000136238	ENST00000356142	RAC1	ENST00000356142.11	-48	-0.309149894800538
ENSG00000162704	ENST00000294742	ARPC5	ENST00000294742.2	-251	-0.310841352790972
ENSG00000173726	ENST00000366607	TOMM20	ENST00000366607.19	-177	-0.31629214760063
ENSG00000044574	ENST00000324460	HSPA5	ENST00000324460.22	-226	-0.324808659740707

ENSG00000198728	ENST00000425280	LDB1	ENST00000425280.21	-205	-0.328098825393004
ENSG00000127054	ENST00000545578	CPSF3L	ENST00000545578.32	-34	-0.346771974303562
ENSG00000257727	ENST00000273308	CNPY2	ENST00000273308.50	-115	-0.361479253612223
ENSG00000100320	ENST00000449924	RBFOX2	ENST00000449924.28	-178	-0.377114632250786
ENSG00000184983	ENST00000498737	NDUFA6	ENST00000498737.4	-263	-0.387230673983785
ENSG00000165609	ENST00000537776	NUDT5	ENST00000537776.46	-243	-0.389163058867658
ENSG00000166595	ENST00000422424	FAM96B	ENST00000422424.2	-164	-0.416418250402919
ENSG00000141759	ENST00000355491	TXNL4A	ENST00000355491.4	-117	-0.418189757568447
ENSG00000105968	ENST00000308153	H2AFV	ENST00000308153.1	-240	-0.424636284455856
ENSG00000116521	ENST00000355379	SCAMP3	ENST00000355379.25	-216	-0.437214601833621
ENSG00000127054	ENST00000540437	CPSF3L	ENST00000540437.25	-159	-0.439294660545869
ENSG00000138801	ENST00000265174	PAPSS1	ENST00000265174.1	-206	-0.439849329893503
ENSG00000163634	ENST00000469584	THOC7	ENST00000469584.7	-116	-0.444284451057211
ENSG00000112514	ENST00000374496	CUTA	ENST00000374496.33	-293	-0.453853258843619
ENSG00000115944	ENST00000234301	COX7A2L	ENST00000234301.4	-213	-0.462732554413591
ENSG00000069275	ENST00000367142	NUCKS1	ENST00000367142.29	-72	-0.463025123513495
ENSG00000124795	ENST00000397239	DEK	ENST00000397239.17	-296	-0.47428935467215
ENSG00000162704	ENST00000294742	ARPC5	ENST00000294742.11	-270	-0.481661270747767
ENSG00000135052	ENST00000388711	GOLM1	ENST00000388711.8	-196	-0.483779734693854
ENSG00000143870	ENST00000381611	PDIA6	ENST00000381611.43	-285	-0.490962260426038
ENSG00000099817	ENST00000615234	POLR2E	ENST00000615234.2	-85	-0.495810750555621
ENSG00000126756	ENST00000335890	UXT	ENST00000335890.17	-225	-0.508973607855719
ENSG00000074201	ENST00000525428	CLNS1A	ENST00000525428.7	-185	-0.516719181922892
ENSG00000198937	ENST00000373408	CCDC167	ENST00000373408.4	-170	-0.520165640348613
ENSG00000198728	ENST00000361198	LDB1	ENST00000361198.54	-293	-0.523296105130001
ENSG00000075945	ENST00000367767	KIFAP3	ENST00000367767.1	-216	-0.525185520231444
ENSG00000241837	ENST00000290299	ATP5O	ENST00000290299.16	-114	-0.530476937321087
ENSG00000116521	ENST00000302631	SCAMP3	ENST00000302631.5	-162	-0.544239900223973
ENSG00000095059	ENST00000210060	DHPS	ENST00000210060.9	-262	-0.549621709551695
ENSG00000237190	ENST00000458198	CDKN2AIPNL	ENST00000458198.2	-259	-0.585034855183767
ENSG00000109332	ENST00000343106	UBE2D3	ENST00000343106.56	-32	-0.590322600719257
ENSG00000106153	ENST00000395422	CHCHD2	ENST00000395422.18	-67	-0.592354933593252
ENSG00000141551	ENST00000392334	CSNK1D	ENST00000392334.2	-175	-0.599562601212801
ENSG00000147123	ENST00000276062	NDUFB11	ENST00000276062.39	-230	-0.617319760348337
ENSG00000136709	ENST00000322313	WDR33	ENST00000322313.6	-254	-0.621102190485708
ENSG00000151694	ENST00000310823	ADAM17	ENST00000310823.14	-107	-0.623032840210605
ENSG00000255526	ENST00000534348	NEDD8-MDP1	ENST00000534348.6	-41	-0.637645524984602
ENSG00000137726	ENST00000614497	FXYD6	ENST00000614497.37	-241	-0.6428314299647
ENSG00000177889	ENST00000318066	UBE2N	ENST00000318066.27	-217	-0.643060374623743
ENSG00000082898	ENST00000401558	XPO1	ENST00000401558.73	-173	-0.662760870411819
ENSG00000003096	ENST00000540167	KLHL13	ENST00000540167.41	-201	-0.675720176705533
ENSG00000122565	ENST00000396386	CBX3	ENST00000396386.12	-60	-0.683349890817454
ENSG00000120742	ENST00000239944	SERP1	ENST00000239944.38	-202	-0.707968706666024
ENSG00000174886	ENST00000418389	NDUFA11	ENST00000418389.1	-103	-0.719015849826662
ENSG00000171858	ENST00000343986	RPS21	ENST00000343986.3	-104	-0.72813144001002
ENSG00000115524	ENST00000414963	SF3B1	ENST00000414963.6	-166	-0.731510873150678

ENSG00000164885	ENST00000618146	CDK5	ENST00000618146.1	-231	-0.732739593101115
ENSG00000177885	ENST00000316804	GRB2	ENST00000316804.24	-98	-0.75006859660494
ENSG00000143543	ENST00000271843	JTB	ENST00000271843.35	-220	-0.750328469205788
ENSG00000119977	ENST00000614499	TCTN3	ENST00000614499.11	-278	-0.751308528604164
ENSG00000089693	ENST00000203630	MLF2	ENST00000203630.49	-113	-0.760283418051521
ENSG00000160075	ENST00000291386	SSU72	ENST00000291386.22	-267	-0.778292785332658
ENSG00000119977	ENST00000614499	TCTN3	ENST00000614499.19	-235	-0.782267935996206
ENSG00000142192	ENST00000359726	APP	ENST00000359726.15	-85	-0.790181138564981
ENSG00000153048	ENST00000614449	CARHSP1	ENST00000614449.11	-248	-0.802418347039104
ENSG00000135404	ENST00000546939	CD63	ENST00000546939.19	-187	-0.803368215438916
ENSG00000130255	ENST00000347512	RPL36	ENST00000347512.9	-101	-0.811787342946931
ENSG00000177733	ENST00000314940	HNRNPA0	ENST00000314940.12	-99	-0.834038425010562
ENSG00000003096	ENST00000545703	KLHL13	ENST00000545703.38	-157	-0.856318603633122
ENSG00000085662	ENST00000285930	AKR1B1	ENST00000285930.4	-85	-0.858164998065364
ENSG00000157593	ENST00000615337	SLC35B2	ENST00000615337.30	-62	-0.872562096514651
ENSG00000170315	ENST00000302182	UBB	ENST00000302182.46	-47	-0.881379772837001
ENSG00000255245	ENST00000532984	FXYD6-FXYD2	ENST00000532984.2	-189	-0.906592968898827
ENSG00000143320	ENST00000368222	CRABP2	ENST00000368222.9	-194	-0.920319738044272
ENSG00000175203	ENST00000434715	DCTN2	ENST00000434715.1	-264	-0.924330527292718
ENSG00000154518	ENST00000284727	ATP5G3	ENST00000284727.403	-107	-0.929179765605121
ENSG00000139116	ENST00000544797	KIF21A	ENST00000544797.4	-282	-0.948718767014162
ENSG00000143570	ENST00000537590	SLC39A1	ENST00000537590.18	-245	-0.961637016549463
ENSG00000107223	ENST00000371649	EDF1	ENST00000371649.1	-221	-0.995249121854717
ENSG00000123144	ENST00000242784	C19orf43	ENST00000242784.9	-168	-1.01048357917033
ENSG00000242485	ENST00000344843	MRPL20	ENST00000344843.8	-245	-1.03841132995427
ENSG00000143612	ENST00000368521	C1orf43	ENST00000368521.16	-176	-1.05801058344335
ENSG00000088247	ENST00000398148	KHSRP	ENST00000398148.2	-289	-1.0731766875609
ENSG00000198931	ENST00000378364	APRT	ENST00000378364.3	-194	-1.08043123202743
ENSG00000167397	ENST00000354895	VKORC1	ENST00000354895.17	-279	-1.08501853897882
ENSG00000136942	ENST00000348462	RPL35	ENST00000348462.1	-55	-1.11092721218343
ENSG00000258947	ENST00000554444	TUBB3	ENST00000554444.48	-47	-1.13801419643665
ENSG00000115694	ENST00000535007	STK25	ENST00000535007.48	-177	-1.13963411524601
ENSG00000034510	ENST00000233143	TMSB10	ENST00000233143.7	-44	-1.14096665432478
ENSG00000115053	ENST00000322723	NCL	ENST00000322723.20	-73	-1.16138760059223
ENSG00000166794	ENST00000300026	PPIB	ENST00000300026.1	-103	-1.16953472834137
ENSG00000110492	ENST00000617138	MDK	ENST00000617138.4	-189	-1.17000997680216
ENSG00000143321	ENST00000537739	HDGF	ENST00000537739.3	-115	-1.18796584105409
ENSG00000177576	ENST00000318240	C18orf32	ENST00000318240.10	-215	-1.20279837697525
ENSG00000117362	ENST00000414276	APH1A	ENST00000414276.29	-266	-1.21853461513679
ENSG00000108424	ENST00000540627	KPNB1	ENST00000540627.44	-89	-1.21856829614583
ENSG00000178952	ENST00000313511	TUFM	ENST00000313511.1	-254	-1.22149508261608
ENSG00000168066	ENST00000334944	SF1	ENST00000334944.30	-97	-1.25158240344226
ENSG00000163479	ENST00000295702	SSR2	ENST00000295702.1	-166	-1.25725161846665
ENSG00000100902	ENST00000622405	PSMA6	ENST00000622405.28	-158	-1.2804061012882
ENSG00000110700	ENST00000525634	RPS13	ENST00000525634.11	-211	-1.28300049594478
ENSG00000108106	ENST00000264552	UBE2S	ENST00000264552.8	-188	-1.2879586388961

ENSG00000104964	ENST00000327141	AES	ENST00000327141.18	-188	-1.29481151581135
ENSG00000161016	ENST00000394920	RPL8	ENST00000394920.11	-70	-1.29782642237
ENSG00000142676	ENST00000374550	RPL11	ENST00000374550.2	-101	-1.31087487320222
ENSG00000196531	ENST00000393891	NACA	ENST00000393891.32	-174	-1.31186245873421
ENSG00000109971	ENST00000534624	HSPA8	ENST00000534624.29	-236	-1.3264140797818
ENSG00000144381	ENST00000345042	HSPD1	ENST00000345042.4	-187	-1.33061886541826
ENSG00000117362	ENST00000369109	APH1A	ENST00000369109.7	-284	-1.34175935016826
ENSG00000196683	ENST00000621567	TOMM7	ENST00000621567.6	-151	-1.36980326886019
ENSG00000111669	ENST00000535434	TPI1	ENST00000535434.54	-77	-1.38001713319878
ENSG00000163682	ENST00000295955	RPL9	ENST00000295955.1	-61	-1.38088773408104
ENSG00000143321	ENST00000537739	HDGF	ENST00000537739.1	-92	-1.39061940279209
ENSG00000186468	ENST00000296674	RPS23	ENST00000296674.10	-174	-1.44965277767312
ENSG00000204628	ENST00000512805	GNB2L1	ENST00000512805.39	-113	-1.45038996618863
ENSG00000167552	ENST00000295766	TUBA1A	ENST00000295766.34	-231	-1.45062087065742
ENSG00000009307	ENST00000358528	CSDE1	ENST00000358528.55	-203	-1.49839302412736
ENSG00000103202	ENST00000620944	NME4	ENST00000620944.26	-95	-1.53451110959796
ENSG00000159335	ENST00000309083	PTMS	ENST00000309083.13	-267	-1.5606760064114
ENSG00000233276	ENST00000620890	GPX1	ENST00000620890.7	-278	-1.56560616045304
ENSG00000197111	ENST00000359462	PCBP2	ENST00000359462.9	-273	-1.56729601762157
ENSG00000067225	ENST00000389093	PKM	ENST00000389093.23	-182	-1.57705878064094
ENSG00000169100	ENST00000381401	SLC25A6	ENST00000381401.59	-281	-1.67459365341014
ENSG00000110321	ENST00000396525	EIF4G2	ENST00000396525.6	-262	-1.68307949354514
ENSG00000110700	ENST00000525634	RPS13	ENST00000525634.4	-93	-1.7321324821891
ENSG00000142937	ENST00000396651	RPS8	ENST00000396651.12	-157	-1.76170933476306
ENSG00000134419	ENST00000322989	RPS15A	ENST00000322989.5	-152	-1.78851234605872
ENSG00000131051	ENST00000528062	RBM39	ENST00000528062.29	-132	-1.84378825807836
ENSG00000166441	ENST00000314138	RPL27A	ENST00000314138.38	-103	-1.9364870969166
ENSG00000114942	ENST00000392221	EEF1B2	ENST00000392221.11	-117	-1.95726819118581
ENSG00000161016	ENST00000262584	RPL8	ENST00000262584.17	-209	-1.99567572901688
ENSG00000125691	ENST00000479035	RPL23	ENST00000479035.15	-89	-2.16195520179977
ENSG00000186591	ENST00000473814	UBE2H	ENST00000473814.1	-141	-2.21344944622502
ENSG00000071553	ENST00000619046	ATP6AP1	ENST00000619046.172	-130	-3.08107352094565
ENSG00000143761	ENST00000540651	ARF1	ENST00000540651.19	-130	-3.09310292609271
ENSG00000143621	ENST00000615950	ILF2	ENST00000615950.16	-143	-4.30466650102223
ENSG00000143256	ENST00000368010	PFDN2	ENST00000368010.6	-137	-4.98448083590993
ENSG00000096092	ENST00000211314	TMEM14A	ENST00000211314.14	-147	-5.09234161017485
ENSG00000183291	ENST00000611507	SEP15	ENST00000611507.44	-143	-5.18003689109536
ENSG00000074800	ENST00000234590	ENO1	ENST00000234590.7	-134	-5.49295117979661
ENSG00000170759	ENST00000302418	KIF5B	ENST00000302418.41	-146	-6.19660816181915
ENSG00000100612	ENST00000557185	DHRS7	ENST00000557185.8	-149	-6.22478403761334
ENSG00000003096	ENST00000545703	KLHL13	ENST00000545703.41	-122	-6.66550073978611
ENSG00000175130	ENST00000329421	MARCKSL1	ENST00000329421.12	-134	-7.56901600568944
ENSG00000156261	ENST00000540844	CCT8	ENST00000540844.43	-128	-7.95648179447574
ENSG00000179010	ENST00000382581	MRFAP1	ENST00000382581.7	-127	-8.26393702815768
ENSG00000145817	ENST00000513112	YIPF5	ENST00000513112.38	-140	-8.48938875645551
ENSG00000140612	ENST00000559729	SEC11A	ENST00000559729.29	-137	-8.76337578986518

ENSG00000137409	ENST00000538808	MTCH1	ENST00000538808.1	-140	-10.1088360945316
ENSG00000197457	ENST00000370053	STMN3	ENST00000370053.2	-128	-10.2126643760176
ENSG00000136758	ENST00000375972	YME1L1	ENST00000375972.14	-146	-10.2370870700578
ENSG00000062582	ENST00000317534	MRPS24	ENST00000317534.2	-134	-10.6356602026815
ENSG00000105825	ENST00000222543	TFPI2	ENST00000222543.19	-121	-10.836637711152
ENSG00000171603	ENST00000361311	CLSTN1	ENST00000361311.14	-136	-12.3761242667669
ENSG00000173457	ENST00000309318	PPP1R14B	ENST00000309318.12	-139	-13.331635373616
ENSG00000138382	ENST00000260953	METTL5	ENST00000260953.23	-134	-13.3901605910573
ENSG00000075142	ENST00000265729	SRI	ENST00000265729.2	-149	-13.5492239730267
ENSG00000077147	ENST00000371142	TM9SF3	ENST00000371142.12	-136	-13.5495964783895
ENSG00000196305	ENST00000375643	IARS	ENST00000375643.28	-134	-16.3148281988894
ENSG00000197958	ENST00000361436	RPL12	ENST00000361436.3	-132	-19.3191399380345

Table B.1 Positive and negative residuals from multiple regression analysis of oORFs in RA differentiated SH-SY5Y cells. Genes are subset to those that terminate within 300 nt of the annotated TIS start site.

APPENDIX C

SUPPLEMENTAL MATERIAL FOR CHAPTER 4

C.1 Methods

Integration of known fusion breakpoints from COSMIC:

```
# Input files:
cosmic_file = sys.argv[1]
fasta_file = sys.argv[2]

# Load FASTA sequence of gene transcripts:
def load_fasta_file(infile):
    fasta = dict()
    for line in open(infile):
        if line.startswith(">"):
            transcript_id = line.strip().split(" ")[1]
            fasta[transcript_id] = str()
        else:
            fasta[transcript_id] += line.strip()
    return fasta

seq = load_fasta_file(fasta_file)

for line in open(cosmic_file):
    fusion = line.strip()
    # Extract gene names and transcript ids of fusion partners:
    genes = [gene[:-1] for gene in re.findall("[a-zA-Z0-9]+{", fusion)]
    transcripts = [transcript[:-1] for transcript in re.findall("[a-zA-Z0-9_.]+}", fusion)]
    # Extract the breakpoint coordinates from HGVS annotation:
    coordinates = re.findall("r.[0-9]+_[0-9]+", fusion)
    fusion_sequences = list()
    for x in xrange(len(transcripts)):
        if transcripts[x] in seq:
            fasta = seq[transcripts[x]]
            # Start must be de-cremented by because Python is 0-based and HGVS coordinates are 1-based:
            start = int(re.findall("[0-9]+", coordinates[x])[0]) - 1
            end = int(re.findall("[0-9]+", coordinates[x])[-1])
            if x == 0:
                fusion_sequences = [fasta[start:end]]
            else:
                fusion_sequences.append(fasta[start:end])
    # By default the breakpoint occurs between the first and second gene:
    break_index = 1
    last_gene = transcripts[0]
    for x in xrange(len(transcripts)):
        if transcripts[x] == last_gene:
```

```
        pass
    else:
        break_index = x
left_sequences = "".join(seq for seq in fusion_sequences[:break_index])
right_sequences = "".join(seq for seq in fusion_sequences[break_index:])
# Trim left sequence to last 30 nucleotides:
left_seq = left_sequences[len(left_sequences)-150:]
# Trim right sequence to first 30 nucleotides:
right_seq = right_sequences[:150]
# Generate the name of the fusion:
fusion_genes = ":".join(gene for gene in list(set(genes)))
fusion_transcripts = ":".join(transcript for transcript in list(set(transcripts)))
fusion_coordinates = ":".join(pos for pos in coordinates)
fusion_name = “>{COSMIC}|” + “|”.join(str(field) for field in [fusion_genes, fusion_transcripts, \
     fusion_coordinates, len(left_seq), len(right_seq)])
#if len(left_seq + right_seq) == 60:
print fusion_name
print left_seq.upper() + right_seq.upper()
```

SRA ID	Condition	Method	Paired	Length	Number of Reads	Total Reads
SRX848657	untreated	mRNA	Yes	102	41,064,374	97,394,360
SRX848658	untreated	mRNA	Yes	102	56,329,986	
SRX1100334	untreated	mRNA	Yes	100	24,207,325	66,933,654
SRX1100335	untreated	mRNA	Yes	100	26,548,537	
SRX1100336	untreated	mRNA	Yes	100	16,177,792	
SRX209073	untreated	mRNA	Yes	102	9,690,701	21,641,312
SRX209074	untreated	mRNA	Yes	102	11,950,611	
SRX118285	vehicle-treated	mRNA	No	40	22,153,049	46,628,839
SRX118291	vehicle-treated	mRNA	No	40	24,475,790	
SRX118287	rapamycin	mRNA	No	40	21,878,806	40,996,269
SRX118293	rapamycin	mRNA	No	40	19,117,463	
SRX118289	PP242	mRNA	No	40	19,655,874	45,411,611
SRX118295	PP242	mRNA	No	40	25,755,737	
SRX118286	vehicle-treated	RPF	No	40	20,762,604	45,160,552
SRX118292	vehicle-treated	RPF	No	40	24,397,948	
SRX118288	rapamycin	RPF	No	40	19,742,795	45,197,863
SRX118294	rapamycin	RPF	No	40	25,455,068	
SRX118290	PP242	RPF	No	40	19,706,845	44,998,520
SRX118296	PP242	RPF	No	40	25,291,675	

Table C.1 Paired-end and single-end sequencing PC-3 mRNA and ribosome profiling libraries used in this study.

			mRNA Reads		RPFs	
Fusion Partners	**Source**	**Breakpoint Coordinates**	**Total**	**Junction**	**Total**	**Junction**
TPM4:ALK	COSMIC	r.1_905:r.4080_6220	36	2	115	3
ABLIM3:RP11-337C18.10	SRX848658	chr5:149142108:+^chr1:147252623:+	11	1	75	12
TXNRD1:UTP20	SRX848657	chr12:104327671:+^chr12:101352055:+	101	11	230	40
FUT8:NUMB	SRX848658	chr14:65413214:+^chr14:73410068:-	20	2	14	3
ETV1:ACSL3	COSMIC	r.1_491:r.1034_6740	35	0	92	5
ETV1:HNRNPA2B1	COSMIC	r.1_175:r.714_6740	334	2	1280	10
PRIM2:ZNF451	SRX848657	chr6:57382168:+^chr6:57161084:+	19	0	48	2
RPTOR:RP11-290K4.2	SRX1100335	chr17:80545791:+^chr3:158016443:-	5	0	28	3
ABLIM3:CHD1L	SRX848658	chr5:149142108:+^chr1:147252623:+	17	0	85	14
LAMA3:PANK2	SRX848657	chr18:23714072:+^chr20:39916927:+	15	0	44	10
RP11-96H19.1:RP11-446N19.1	SRX1100334|SRX209073|SRX848657|SRX1100335	chr12:46387972:+^chr12:46652390:+	20	0	18	4
CCT7:DYNC1H1	SRX848657	chr2:73251432:+^chr14:101999989:+	138	1	369	5
ARMCX5-GPRASP2	SRX209073|SRX1100335	chrX:102605653:+^chrX:102713782:+	9	1	19	3
ETV4:CANT1	COSMIC	r.1_154:r.518_2391	30	1	20	4

Table C.2 Filtered set of previously annotated and novel STAR-FUSION gene fusions.

LITERATURE CITED

1. Crick, F.H., *On protein synthesis.* Symp Soc Exp Biol, 1958. **12**: p. 138-63.
2. Ambros, V., *The functions of animal microRNAs.* Nature, 2004. **431**(7006): p. 350-5.
3. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.
4. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.
5. Farazi, T.A., J.I. Spitzer, *et al.*, *miRNAs in human cancer.* J Pathol, 2011. **223**(2): p. 102-15.
6. Franks, A., E. Airoldi, and N. Slavov, *Post-transcriptional regulation across human tissues.* PLoS Comput Biol, 2017. **13**(5): p. e1005535.
7. Friedman, R.C., K.K. Farh, *et al.*, *Most mammalian mRNAs are conserved targets of microRNAs.* Genome Res, 2009. **19**(1): p. 92-105.
8. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* J Mol Biol, 1961. **3**: p. 318-56.
9. Kozak, M., *Initiation of translation in prokaryotes and eukaryotes.* Gene, 1999. **234**(2): p. 187-208.
10. Beck-Sickinger, A. and K. Mörl, *Posttranslational Modification of Proteins. Expanding Nature's Inventory. By Christopher T. Walsh*. Vol. 45. 2006. 1020-1020.
11. Hinnebusch, A.G., *Molecular mechanism of scanning and start codon selection in eukaryotes.* Microbiol Mol Biol Rev, 2011. **75**(3): p. 434-67, first page of table of contents.
12. Kozak, M., *Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes.* Cell, 1986. **44**(2): p. 283-92.
13. Feng, Y., F. Zhang, *et al.*, *Translational suppression by trinucleotide repeat expansion at FMR1.* Science, 1995. **268**(5211): p. 731-4.
14. Sans, M.D., Q. Xie, and J.A. Williams, *Regulation of translation elongation and phosphorylation of eEF2 in rat pancreatic acini.* Biochem Biophys Res Commun, 2004. **319**(1): p. 144-51.
15. Nakamura, Y., K. Ito, *et al.*, *Regulation of translation termination: conserved structural motifs in bacterial and eukaryotic polypeptide release factors.* Biochem Cell Biol, 1995. **73**(11-12): p. 1113-22.
16. Chou, T., *Ribosome recycling, diffusion, and mRNA loop formation in translational regulation.* Biophys J, 2003. **85**(2): p. 755-73.
17. Hinnebusch, A.G., I.P. Ivanov, and N. Sonenberg, *Translational control by 5'-untranslated regions of eukaryotic mRNAs.* Science, 2016. **352**(6292): p. 1413-6.
18. Shine, J. and L. Dalgarno, *Determinant of cistron specificity in bacterial ribosomes.* Nature, 1975. **254**(5495): p. 34-8.
19. Loughran, G., M.S. Sachs, *et al.*, *Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5.* Nucleic Acids Res, 2012. **40**(7): p. 2898-906.
20. Fijalkowska, D., S. Verbruggen, *et al.*, *eIF1 modulates the recognition of suboptimal translation initiation sites and steers gene expression via uORFs.* Nucleic Acids Res, 2017. **45**(13): p. 7997-8013.
21. Pestova, T.V. and V.G. Kolupaeva, *The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection.* Genes Dev, 2002. **16**(22): p. 2906-22.
22. Arribere, J.A. and W.V. Gilbert, *Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing.* Genome Res, 2013. **23**(6): p. 977-87.

23. Kozak, M., *Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes.* Proc Natl Acad Sci U S A, 1990. **87**(21): p. 8301-5.
24. Weingarten-Gabbay, S., S. Elias-Kirma, *et al.*, *Comparative genetics. Systematic discovery of cap-independent translation sequences in human and viral genomes.* Science, 2016. **351**(6270).
25. Calvo, S.E., D.J. Pagliarini, and V.K. Mootha, *Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans.* Proc Natl Acad Sci U S A, 2009. **106**(18): p. 7507-12.
26. Hsieh, A.C., Y. Liu, *et al.*, *The translational landscape of mTOR signalling steers cancer initiation and metastasis.* Nature, 2012. **485**(7396): p. 55-61.
27. Mamane, Y., E. Petroulakis, *et al.*, *mTOR, translation initiation and cancer.* Oncogene, 2006. **25**(48): p. 6416-22.
28. Pelletier, J., J. Graff, *et al.*, *Targeting the eIF4F translation initiation complex: a critical nexus for cancer development.* Cancer Res, 2015. **75**(2): p. 250-63.
29. Schwanhausser, B., D. Busse, *et al.*, *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-42.
30. Sonenberg, N. and A.G. Hinnebusch, *Regulation of translation initiation in eukaryotes: mechanisms and biological targets.* Cell, 2009. **136**(4): p. 731-45.
31. Barbosa, C., I. Peixeiro, and L. Romao, *Gene expression regulation by upstream open reading frames and human disease.* PLoS Genet, 2013. **9**(8): p. e1003529.
32. Holcik, M. and N. Sonenberg, *Translational control in stress and apoptosis.* Nat Rev Mol Cell Biol, 2005. **6**(4): p. 318-27.
33. Harding, H.P., Y. Zhang, *et al.*, *Perk is essential for translational regulation and cell survival during the unfolded protein response.* Mol Cell, 2000. **5**(5): p. 897-904.
34. Scheuner, D., B. Song, *et al.*, *Translational control is required for the unfolded protein response and in vivo glucose homeostasis.* Mol Cell, 2001. **7**(6): p. 1165-76.
35. Vattem, K.M. and R.C. Wek, *Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells.* Proc Natl Acad Sci U S A, 2004. **101**(31): p. 11269-74.
36. Occhi, G., D. Regazzo, *et al.*, *A novel mutation in the upstream open reading frame of the CDKN1B gene causes a MEN4 phenotype.* PLoS Genet, 2013. **9**(3): p. e1003350.
37. Pellegata, N.S., L. Quintanilla-Martinez, *et al.*, *Germ-line mutations in p27Kip1 cause a multiple endocrine neoplasia syndrome in rats and humans.* Proc Natl Acad Sci U S A, 2006. **103**(42): p. 15558-63.
38. Kearse, M.G., K.M. Green, *et al.*, *CGG Repeat-Associated Non-AUG Translation Utilizes a Cap-Dependent Scanning Mechanism of Initiation to Produce Toxic Proteins.* Mol Cell, 2016. **62**(2): p. 314-22.
39. The, E.P.C., *An integrated encyclopedia of DNA elements in the human genome.* 2012. **489**: p. 57.
40. Lytle, J.R., T.A. Yario, and J.A. Steitz, *Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR.* Proc Natl Acad Sci U S A, 2007. **104**(23): p. 9667-72.
41. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.
42. Li, S., L. Liu, *et al.*, *MicroRNAs inhibit the translation of target mRNAs on the endoplasmic reticulum in Arabidopsis.* Cell, 2013. **153**(3): p. 562-74.
43. Brogna, S. and J. Wen, *Nonsense-mediated mRNA decay (NMD) mechanisms.* Nat Struct Mol Biol, 2009. **16**(2): p. 107-13.
44. Mann, M., N.A. Kulak, *et al.*, *The coming age of complete, accurate, and ubiquitous proteomes.* Mol Cell, 2013. **49**(4): p. 583-90.
45. Lander, E.S., L.M. Linton, *et al.*, *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
46. Venter, J.C., M.D. Adams, *et al.*, *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.

47. Sanger, F. and A.R. Coulson, *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.* J Mol Biol, 1975. **94**(3): p. 441-8.
48. Staden, R., *A strategy of DNA sequencing employing computer programs.* Nucleic Acids Res, 1979. **6**(7): p. 2601-10.
49. Anderson, S., *Shotgun DNA sequencing using cloned DNase I-generated fragments.* Nucleic Acids Res, 1981. **9**(13): p. 3015-27.
50. International Human Genome Sequencing, C., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
51. Gygi, S.P., B. Rist, *et al.*, *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.* Nat Biotechnol, 1999. **17**(10): p. 994-9.
52. Link, A.J., J. Eng, *et al.*, *Direct analysis of protein complexes using mass spectrometry.* Nat Biotechnol, 1999. **17**(7): p. 676-82.
53. Washburn, M.P., D. Wolters, and J.R. Yates, 3rd, *Large-scale analysis of the yeast proteome by multidimensional protein identification technology.* Nat Biotechnol, 2001. **19**(3): p. 242-7.
54. Melton, L., *Protein arrays: proteomics in multiplex.* Nature, 2004. **429**(6987): p. 101-7.
55. Bentley, D.R., S. Balasubramanian, *et al.*, *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
56. Morin, R., M. Bainbridge, *et al.*, *Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing.* Biotechniques, 2008. **45**(1): p. 81-94.
57. Fullwood, M.J., C.L. Wei, *et al.*, *Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses.* Genome Res, 2009. **19**(4): p. 521-32.
58. Trapnell, C., A. Roberts, *et al.*, *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.
59. Williams, C.R., A. Baccarella, *et al.*, *Trimming of sequence reads alters RNA-Seq gene expression estimates.* BMC Bioinformatics, 2016. **17**: p. 103.
60. Fox, E.J., K.S. Reid-Bayliss, *et al.*, *Accuracy of Next Generation Sequencing Platforms.* Next Gener Seq Appl, 2014. **1**.
61. Burrows, M., Wheeler, D. J., *A block sorting lossless data compression algorithm.* Technical Report, 1994. **124**: p. 18.
62. Langmead, B., C. Trapnell, *et al.*, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
63. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
64. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
65. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
66. Dobin, A., C.A. Davis, *et al.*, *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
67. Patro, R., S.M. Mount, and C. Kingsford, *Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms.* Nat Biotechnol, 2014. **32**(5): p. 462-4.
68. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.
69. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.
70. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
71. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

72. Burnette, W.N., *"Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate--polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A.* Anal Biochem, 1981. **112**(2): p. 195-203.
73. Williams, T.M., J.E. Burlein, *et al.*, *Advantages of firefly luciferase as a reporter gene: application to the interleukin-2 gene promoter.* Anal Biochem, 1989. **176**(1): p. 28-32.
74. Arun, K.H., C.L. Kaul, and P. Ramarao, *Green fluorescent proteins in receptor research: an emerging tool for drug discovery.* J Pharmacol Toxicol Methods, 2005. **51**(1): p. 1-23.
75. Lequin, R.M., *Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA).* Clin Chem, 2005. **51**(12): p. 2415-8.
76. Kodadek, T., *Protein microarrays: prospects and problems.* Chem Biol, 2001. **8**(2): p. 105-15.
77. Perkins, D.N., D.J. Pappin, *et al.*, *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.
78. Craig, R. and R.C. Beavis, *TANDEM: matching proteins with tandem mass spectra.* Bioinformatics, 2004. **20**(9): p. 1466-7.
79. Ma, K., O. Vitek, and A.I. Nesvizhskii, *A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet.* BMC Bioinformatics, 2012. **13 Suppl 16**: p. S1.
80. Deutsch, E.W., L. Mendoza, *et al.*, *A guided tour of the Trans-Proteomic Pipeline.* Proteomics, 2010. **10**(6): p. 1150-9.
81. Kong, A.T., F.V. Leprevost, *et al.*, *MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics.* Nat Methods, 2017. **14**(5): p. 513-520.
82. McIlwain, S., M. Mathews, *et al.*, *Estimating relative abundances of proteins from shotgun proteomics data.* BMC Bioinformatics, 2012. **13**: p. 308.
83. Ning, K., D. Fermin, and A.I. Nesvizhskii, *Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data.* J Proteome Res, 2012. **11**(4): p. 2261-71.
84. Nesvizhskii, A.I., *Proteogenomics: concepts, applications and computational strategies.* Nat Methods, 2014. **11**(11): p. 1114-25.
85. Ingolia, N.T., S. Ghaemmaghami, *et al.*, *Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.* Science, 2009. **324**(5924): p. 218-23.
86. Obrig, T.G., W.J. Culp, *et al.*, *The mechanism by which cycloheximide and related glutarimide antibiotics inhibit peptide synthesis on reticulocyte ribosomes.* J Biol Chem, 1971. **246**(1): p. 174-81.
87. Poehlsgaard, J. and S. Douthwaite, *The bacterial ribosome as a target for antibiotics.* Nat Rev Microbiol, 2005. **3**(11): p. 870-81.
88. McGlincy, N.J. and N.T. Ingolia, *Transcriptome-wide measurement of translation by ribosome profiling.* Methods, 2017. **126**: p. 112-129.
89. Ingolia, N.T., G.A. Brar, *et al.*, *The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments.* Nat Protoc, 2012. **7**(8): p. 1534-50.
90. Becker, A.H., E. Oh, *et al.*, *Selective ribosome profiling as a tool for studying the interaction of chaperones and targeting factors with nascent polypeptide chains and ribosomes.* Nat Protoc, 2013. **8**(11): p. 2212-39.
91. Chhangawala, S., G. Rudy, *et al.*, *The impact of read length on quantification of differentially expressed genes and splice junction detection.* Genome Biol, 2015. **16**: p. 131.
92. Lareau, L.F., D.H. Hite, *et al.*, *Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments.* Elife, 2014. **3**: p. e01257.
93. Dunn, J.G. and J.S. Weissman, *Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data.* BMC Genomics, 2016. **17**(1): p. 958.
94. Pestova, T.V. and C.U. Hellen, *Translation elongation after assembly of ribosomes on the Cricket paralysis virus internal ribosomal entry site without initiation factors or initiator tRNA.* Genes Dev, 2003. **17**(2): p. 181-6.

95. Ingolia, N.T., G.A. Brar, *et al.*, *Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes.* Cell Rep, 2014. **8**(5): p. 1365-79.
96. Chun, S.Y., C.M. Rodriguez, *et al.*, *SPECtre: a spectral coherence--based classifier of actively translated transcripts from ribosome profiling sequence data.* BMC Bioinformatics, 2016. **17**(1): p. 482.
97. Vasquez, J.J., C.C. Hon, *et al.*, *Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages.* Nucleic Acids Res, 2014. **42**(6): p. 3623-37.
98. Mumtaz, M.A. and J.P. Couso, *Ribosomal profiling adds new coding sequences to the proteome.* Biochem Soc Trans, 2015. **43**(6): p. 1271-6.
99. Bazzini, A.A., T.G. Johnstone, *et al.*, *Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation.* EMBO J, 2014. **33**(9): p. 981-93.
100. Calviello, L., N. Mukherjee, *et al.*, *Detecting actively translated open reading frames in ribosome profiling data.* Nat Methods, 2016. **13**(2): p. 165-70.
101. Fields, A.P., E.H. Rodriguez, *et al.*, *A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation.* Mol Cell, 2015. **60**(5): p. 816-827.
102. White, L.B., Boashash, B., *Cross Spectral Analysis of Nonstationary Processes.* IEEE Transactions on Information Theory, 1990. **36**(4): p. 830-835.
103. Bendat, J.S. and A.G. Piersol, *Random data : analysis and measurement procedures*. 2nd ed. 1986, New York: Wiley. xvii, 566 p.
104. Welch, P., *The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms.* IEEE Transactions on Audio and Electroacoustics, 1967. **15**(2): p. 70-73.
105. Trapnell, C., B.A. Williams, *et al.*, *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.
106. Harrow, J., A. Frankish, *et al.*, *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.
107. Zhang, B., J. Wang, *et al.*, *Proteogenomic characterization of human colon and rectal cancer.* Nature, 2014. **513**(7518): p. 382-7.
108. Brar, G.A., M. Yassour, *et al.*, *High-resolution view of the yeast meiotic program revealed by ribosome profiling.* Science, 2012. **335**(6068): p. 552-7.
109. Ingolia, N.T., L.F. Lareau, and J.S. Weissman, *Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes.* Cell, 2011. **147**(4): p. 789-802.
110. Stern-Ginossar, N., B. Weisburd, *et al.*, *Decoding human cytomegalovirus.* Science, 2012. **338**(6110): p. 1088-93.
111. Andreev, D.E., P.B. O'Connor, *et al.*, *Translation of 5' leaders is pervasive in genes resistant to eIF2 repression.* Elife, 2015. **4**: p. e03971.
112. Arias, C., B. Weisburd, *et al.*, *KSHV 2.0: a comprehensive annotation of the Kaposi's sarcoma-associated herpesvirus genome using next-generation sequencing reveals novel genomic and functional features.* PLoS Pathog, 2014. **10**(1): p. e1003847.
113. Bazzini, A.A., M.T. Lee, and A.J. Giraldez, *Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish.* Science, 2012. **336**(6078): p. 233-7.
114. Caro, F., V. Ahyong, *et al.*, *Genome-wide regulatory dynamics of translation in the Plasmodium falciparum asexual blood stages.* Elife, 2014. **3**.
115. Jensen, B.C., G. Ramasamy, *et al.*, *Extensive stage-regulation of translation revealed by ribosome profiling of Trypanosoma brucei.* BMC Genomics, 2014. **15**: p. 911.
116. Juntawong, P., T. Girke, *et al.*, *Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis.* Proc Natl Acad Sci U S A, 2014. **111**(1): p. E203-12.
117. Oh, E., A.H. Becker, *et al.*, *Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo.* Cell, 2011. **147**(6): p. 1295-308.

118. Schafer, S., E. Adami, *et al.*, *Translational regulation shapes the molecular landscape of complex disease phenotypes.* Nat Commun, 2015. **6**: p. 7200.
119. Stadler, M. and A. Fire, *Wobble base-pairing slows in vivo translation elongation in metazoans.* RNA, 2011. **17**(12): p. 2063-73.
120. Chew, G.L., A. Pauli, *et al.*, *Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs.* Development, 2013. **140**(13): p. 2828-34.
121. Chew, G.L., A. Pauli, and A.F. Schier, *Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish.* Nat Commun, 2016. **7**: p. 11663.
122. Crappe, J., W. Van Criekinge, *et al.*, *Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs.* BMC Genomics, 2013. **14**: p. 648.
123. Ji, Z., R. Song, *et al.*, *Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins.* Elife, 2015. **4**: p. e08890.
124. Johnstone, T.G., A.A. Bazzini, and A.J. Giraldez, *Upstream ORFs are prevalent translational repressors in vertebrates.* EMBO J, 2016. **35**(7): p. 706-23.
125. Pauli, A., M.L. Norris, *et al.*, *Toddler: an embryonic signal that promotes cell movement via Apelin receptors.* Science, 2014. **343**(6172): p. 1248636.
126. Smith, J.E., J.R. Alvarez-Dominguez, *et al.*, *Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae.* Cell Rep, 2014. **7**(6): p. 1858-66.
127. Raj, A., S.H. Wang, *et al.*, *Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling.* Elife, 2016. **5**.
128. Chiocchetti, A.G., D. Haslinger, *et al.*, *Transcriptomic signatures of neuronal differentiation and their association with risk genes for autism spectrum and related neuropsychiatric disorders.* Transl Psychiatry, 2016. **6**(8): p. e864.
129. Korecka, J.A., R.E. van Kesteren, *et al.*, *Phenotypic characterization of retinoic acid differentiated SH-SY5Y cells by transcriptional profiling.* PLoS One, 2013. **8**(5): p. e63862.
130. Crappe, J., E. Ndah, *et al.*, *PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration.* Nucleic Acids Res, 2015. **43**(5): p. e29.
131. Guttman, M., P. Russell, *et al.*, *Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins.* Cell, 2013. **154**(1): p. 240-51.
132. Blank, H.M., R. Perez, *et al.*, *Translational control of lipogenic enzymes in the cell cycle of synchronous, growing yeast cells.* EMBO J, 2017. **36**(4): p. 487-502.
133. Janich, P., A.B. Arpat, *et al.*, *Ribosome profiling reveals the rhythmic liver translatome and circadian clock regulation by upstream open reading frames.* Genome Res, 2015. **25**(12): p. 1848-59.
134. Sendoel, A., J.G. Dunn, *et al.*, *Translation from unconventional 5′ start sites drives tumour initiation.* Nature, 2017. **541**(7638): p. 494-499.
135. Blair, J.D., D. Hockemeyer, *et al.*, *Widespread Translational Remodeling during Human Neuronal Differentiation.* Cell Rep, 2017. **21**(7): p. 2005-2016.
136. Kearse, M.G. and P.K. Todd, *Repeat-associated non-AUG translation and its impact in neurodegenerative disease.* Neurotherapeutics, 2014. **11**(4): p. 721-31.
137. Zu, T., B. Gibbens, *et al.*, *Non-ATG-initiated translation directed by microsatellite expansions.* Proc Natl Acad Sci U S A, 2011. **108**(1): p. 260-5.
138. Forster, J.I., S. Koglsberger, *et al.*, *Characterization of Differentiated SH-SY5Y as Neuronal Screening Model Reveals Increased Oxidative Vulnerability.* J Biomol Screen, 2016. **21**(5): p. 496-509.
139. Nicolini, G., M. Miloso, *et al.*, *Retinoic acid differentiated SH-SY5Y human neuroblastoma cells: an in vitro model to assess drug neurotoxicity.* Anticancer Res, 1998. **18**(4A): p. 2477-81.
140. Shipley, M.M., C.A. Mangold, and M.L. Szpara, *Differentiation of the SH-SY5Y Human Neuroblastoma Cell Line.* J Vis Exp, 2016(108): p. 53193.
141. Amaratunga, D. and J. Cabrera, *Analysis of Data From Viral DNA Microchips.* Journal of the American Statistical Association, 2001. **96**(456): p. 1161-1170.

142. Bolstad, B.M., R.A. Irizarry, *et al.*, *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
143. Tukey, J.W., *Comparing individual means in the analysis of variance.* Biometrics, 1949. **5**(2): p. 99-114.
144. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
145. Young, M.D., M.J. Wakefield, *et al.*, *Gene ontology analysis for RNA-seq: accounting for selection bias.* Genome Biol, 2010. **11**(2): p. R14.
146. Li, W., W. Wang, *et al.*, *Riborex: fast and flexible identification of differential translation from Ribo-seq data.* Bioinformatics, 2017. **33**(11): p. 1735-1737.
147. Lin, M.F., I. Jungreis, and M. Kellis, *PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions.* Bioinformatics, 2011. **27**(13): p. i275-82.
148. Peabody, D.S., *Translation initiation at non-AUG triplets in mammalian cells.* J Biol Chem, 1989. **264**(9): p. 5031-5.
149. Clement, K., C. Vaisse, *et al.*, *A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction.* Nature, 1998. **392**(6674): p. 398-401.
150. Marti, A., J.L. Santos, *et al.*, *Association between leptin receptor (LEPR) and brain-derived neurotrophic factor (BDNF) gene variants and obesity: a case-control study.* Nutr Neurosci, 2009. **12**(4): p. 183-8.
151. Phillips, C.M., L. Goumidi, *et al.*, *Leptin receptor polymorphisms interact with polyunsaturated fatty acids to augment risk of insulin resistance and metabolic syndrome in adults.* J Nutr, 2010. **140**(2): p. 238-44.
152. Plaza, S., G. Menschaert, and F. Payre, *In Search of Lost Small Peptides.* Annu Rev Cell Dev Biol, 2017. **33**: p. 391-416.
153. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer Statistics, 2017.* CA Cancer J Clin, 2017. **67**(1): p. 7-30.
154. Draisma, G., R. Etzioni, *et al.*, *Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context.* J Natl Cancer Inst, 2009. **101**(6): p. 374-83.
155. Hayes, J.H. and M.J. Barry, *Screening for prostate cancer with the prostate-specific antigen test: a review of current evidence.* JAMA, 2014. **311**(11): p. 1143-9.
156. Martin, R.M., *Commentary: prostate cancer is omnipresent, but should we screen for it?* Int J Epidemiol, 2007. **36**(2): p. 278-81.
157. Rich, A.R., *Classics in oncology. On the frequency of occurrence of occult carcinoma of the prostate: Arnold Rice Rich, M.D., Journal of Urology 33:3, 1935.* CA Cancer J Clin, 1979. **29**(2): p. 115-9.
158. Friedlander, T.W., R. Roy, *et al.*, *Common structural and epigenetic changes in the genome of castration-resistant prostate cancer.* Cancer Res, 2012. **72**(3): p. 616-25.
159. Giordano, T.J., *The cancer genome atlas research network: a sight to behold.* Endocr Pathol, 2014. **25**(4): p. 362-5.
160. Kim, J.H., S.M. Dhanasekaran, *et al.*, *Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer.* Genome Res, 2011. **21**(7): p. 1028-41.
161. Robinson, D., E.M. Van Allen, *et al.*, *Integrative clinical genomics of advanced prostate cancer.* Cell, 2015. **161**(5): p. 1215-1228.
162. Robinson, D.R., Y.M. Wu, *et al.*, *Integrative clinical genomics of metastatic cancer.* Nature, 2017. **548**(7667): p. 297-303.
163. Mehra, R., S.A. Tomlins, *et al.*, *Comprehensive assessment of TMPRSS2 and ETS family gene aberrations in clinically localized prostate cancer.* Mod Pathol, 2007. **20**(5): p. 538-44.
164. Mehra, R., S.A. Tomlins, *et al.*, *Characterization of TMPRSS2-ETS gene aberrations in androgen-independent metastatic prostate cancer.* Cancer Res, 2008. **68**(10): p. 3584-90.

165. Mertz, K.D., S.R. Setlur, *et al.*, *Molecular characterization of TMPRSS2-ERG gene fusion in the NCI-H660 prostate cancer cell line: a new perspective for an old model.* Neoplasia, 2007. **9**(3): p. 200-6.
166. Tomlins, S.A., B. Laxman, *et al.*, *Role of the TMPRSS2-ERG gene fusion in prostate cancer.* Neoplasia, 2008. **10**(2): p. 177-88.
167. Tomlins, S.A., D.R. Rhodes, *et al.*, *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. **310**(5748): p. 644-8.
168. Bozic, I., T. Antal, *et al.*, *Accumulation of driver and passenger mutations during tumor progression.* Proc Natl Acad Sci U S A, 2010. **107**(43): p. 18545-50.
169. Haas, B., A. Dobin, *et al.*, *STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq.* bioRxiv, 2017.
170. Hsieh, G., R. Bierman, *et al.*, *Statistical algorithms improve accuracy of gene fusion detection.* Nucleic Acids Res, 2017. **45**(13): p. e126.
171. Kim, D. and S.L. Salzberg, *TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.* Genome Biol, 2011. **12**(8): p. R72.
172. Mertens, F., B. Johansson, *et al.*, *The emerging complexity of gene fusions in cancer.* Nat Rev Cancer, 2015. **15**(6): p. 371-81.
173. Mitelman, F., B. Johansson, and F. Mertens, *The impact of translocations and gene fusions on cancer causation.* Nat Rev Cancer, 2007. **7**(4): p. 233-45.
174. He, J., X. Sun, *et al.*, *Antibody-independent targeted quantification of TMPRSS2-ERG fusion protein products in prostate cancer.* Mol Oncol, 2014. **8**(7): p. 1169-80.
175. den Dunnen, J.T., R. Dalgleish, *et al.*, *HGVS Recommendations for the Description of Sequence Variants: 2016 Update.* Hum Mutat, 2016. **37**(6): p. 564-9.
176. Forbes, S.A., D. Beare, *et al.*, *COSMIC: somatic cancer genetics at high-resolution.* Nucleic Acids Res, 2017. **45**(D1): p. D777-D783.
177. Schlattl, A., S. Anders, *et al.*, *Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions.* Genome Res, 2011. **21**(12): p. 2004-13.
178. Seim, I., P.L. Jeffery, *et al.*, *Whole-Genome Sequence of the Metastatic PC3 and LNCaP Human Prostate Cancer Cell Lines.* G3 (Bethesda), 2017. **7**(6): p. 1731-1741.
179. Lee, M., K. Lee, *et al.*, *ChimerDB 3.0: an enhanced database for fusion genes from cancer transcriptome and literature data mining.* Nucleic Acids Res, 2017. **45**(D1): p. D784-D789.
180. Zhao, X., S.B. Emery, *et al.*, *Resolving complex structural genomic rearrangements using a randomized approach.* Genome Biol, 2016. **17**(1): p. 126.
181. Rosenbloom, K.R., J. Armstrong, *et al.*, *The UCSC Genome Browser database: 2015 update.* Nucleic Acids Res, 2015. **43**(Database issue): p. D670-81.
182. Kim, D., G. Pertea, *et al.*, *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biol, 2013. **14**(4): p. R36.
183. Flicek, P., M.R. Amode, *et al.*, *Ensembl 2014.* Nucleic Acids Res, 2014. **42**(Database issue): p. D749-55.
184. Li, H., B. Handsaker, *et al.*, *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
185. Frazee, A.C., A.E. Jaffe, *et al.*, *Polyester: simulating RNA-seq datasets with differential transcript expression.* Bioinformatics, 2015. **31**(17): p. 2778-84.
186. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.
187. Walt, S.v.d., S.C. Colbert, and G. Varoquaux, *The NumPy Array: A Structure for Efficient Numerical Computation.* Computing in Science & Engineering, 2011. **13**(2): p. 22-30.
188. Sing, T., O. Sander, *et al.*, *ROCR: visualizing classifier performance in R.* Bioinformatics, 2005. **21**(20): p. 3940-1.