Title

Untranslated regions and the regulation of transcript specific translation

Author

Arake deTacca, Luisa Mayumi

Publication Date

2020

Peer reviewed|Thesis/dissertation

Untranslated regions and the regulation of transcript specific translation

By

Luisa Mayumi Arake de Tacca

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Comparative Biochemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jamie H. Doudna Cate, Chair
Professor Jennifer Doudna
Professor Daniel Nomura

Summer 2020

Abstract

Untranslated regions and the regulation of transcript specific translation

by

Luisa Mayumi Arake de Tacca

Doctor of Philosophy in Comparative Biochemistry

University of California, Berkeley

Associate Professor Jamie Doudna-Cate, Chair

Proteins are considered the "workhorses" of the cell and they are produced according to the central dogma of biology. The general rule is that the genetic information hard-wired into DNA is transcribed into a messenger RNA (mRNA) molecule which contains the program for protein synthesis through the translation process. The regulation of protein production can happen in many ways. The process of mRNA production (transcriptional regulation) has been studied extensively and we have a good understanding of how it works. More recently, other forms of regulation have been gaining attention, particularly translation initiation regulation. This is the rate limiting step during translation and it is an important gatekeeper of protein synthesis. This regulation occurs by both the *cis*−regulatory elements, which are located in the 5′− and 3′−UTRs (untranslated regions), and by *trans*−acting factors. Translational control of mRNA provides the cell with a rapid way to control changes in protein concentration and thus acts to assist in maintaining homeostasis while also having a role in modulating more persistent physiological changes towards cell fate (Sonnenberg and Hinnebusch, 2009). A large proportion of the energy budget of a living cell is funneled into protein synthesis making it intimately integrated with cell metabolism. For this reason, misregulation of translation results in aberrations and several disease phenotypes (Silvera et al., 2010). It is therefore of great value to understand detailed aspects of translational control when studying cell homeostasis and disease. Untranslated regions of messenger RNAs are populated with a variety of regulatory structures such as stem-loop structures, upstream initiation codons and open reading frames, internal ribosome entry sites and *cis*-acting elements that interact with RNA-binding proteins. In the present work, I will discuss the importance of untranslated elements on the 5' and the 3' ends of specific transcripts and how interactions with these regions alter the interplay between the RNA and the translation machinery, focusing on eukaryotic translation initiation factor eIF3,the largest translation initiation factor.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

As I reach the end of this journey, this was by far the hardest piece I had to write.

I have met a lot of inspiring people early in my career and I would not be here without them. I will be forever in debt with Carlos Bloch Jr, Guilherme Brandt, Maura Prates and Diego Pires. We can't ever forget where we come from.

My PhD advisor, Jamie Cate, certainly changed the way I do science. He welcomes anyone that seeks to become better. He is patient and understanding. He works with people's personal story and he provides everything you need to excel. I started this PhD with a 6 month old son. Jamie never doubted my abilities as a scientist or as a student because of my responsibilities with my family. Thank you for believing in me and allowing me to do science at my own pace. I learned to be generous, or at least a little bit more generous, because of you, Jamie. Thank you for teaching me that everybody is different and will get to where they need to get at their own time.

I believe that everybody invested in a PhD reaches their breaking moment at some point. I was very fortunate to find a way out of that. Stephen Floor was my co-advisor for the last year of this PhD and probably the reason I made it through to the end. I had just had my second baby and was commuting to UCSF-Parnassus to spend the day learning a completely new technique and talking to amazing people and contributing to amazing science. It re-energized me. Thank you, Stephen. I learned a bit of computer magic from you and learned a whole lot about what it means to be a scientist and a citizen.

My direct lab advisors, Duanne and Amy, you were so important in this PhD. I was very overwhelmed when I started. I knew literally nothing. Thank you for showing me the way and for letting me straight up copy the way you guys did things until I found my own way.

Thank you, Liz Mo, for being my complaining buddy. We would sit across from each-other, not actually seeing one another, and would just complain about the most diverse topics. From science to kids, from potty-training to doctor appointments. I deeply missed you when you left and now I am glad we are keeping in touch and I can call you a friend. Thank you, Arto. You were a great bay mate and taught me so much. When I moved to your bay, Annsea had just left. I was so sad and yet being your bay mate made everything better. Thank you, Audre, for being the senior post doc, the one we would ask so many things and confide our concerns. Thank you, WF. We were friends before we realized. It was so precious watching you work and be the best in everything you decided to do. I admire your work ethic so much! Thank you for welcoming me as a friend and teaching me so much in lab. I hope we can still be friends no matter where we both end up. Thank you, Fred. It was so great to be your bay mate and lab friend. It is so great to be your life friend and outside of the lab friend. Thank you for showing me that some people can be messy as hell and yet have everything together.

Special thank you to Mia and Annsea. We were so tight. You guys helped me through a lot. There are no words to describe how much the two of you are a big part of this PhD. Thank you for teaching me how to be a good lab mate and a good friend and how the two can go together.

Thank you everybody in the Cate lab no matter what level of contact we had. You will always be part of my memory.

Oh, the friends I made. Raissa was the first one. It was instant. We never had to try. I am so happy to be able to call her a true life friend. We had babies together, our families are friends. We know each others parents. That's the level of friendship right there. Thank you for making this jungle feel like home!

Daphne, Tim, Boyan, Yasmin, Dielly and Todd. We stuck together. Once daycare friends, always daycare friends. Look at how big these kids are. Thank you for being friends with us since the start. Thank you for providing our son with a trust circle. He knows his best friends are Eleanor, Nura and Joaquin. He knows it's who he grew up with. Thank you for being family to us. I hope this will carry on forever and these kids will grow up together.

Ella, Seb, David and Elena. You are family away from home. I cannot express how much love I feel for you and how important our group was during these last 6 years. I carry you in my heart forever and expect we will be friends for a long time. We had our girls together. Elsa, Olivia and Evelyn will always have that connection and I am so thankful this connection if with you guys. Thank you from the bottom of my heart.

Thank you to my program and to MCB for letting me attend all fo their units.

I would like to thank my committee. Prof. Doudna and Prof. Nomura for taking the time to evaluate me throughout these 6 years while suggesting new venues for my research and better ways to carry experiments and projects. You are a personal role model of great scientists and change-makers. Thank you.

Thank you to my family for putting up with never seeing the kids or us. For being so proud of me and also saying I am not a real doctor. Thank you to my in-laws, Maria Carmen Tacca and Luiz Tacca Jr., for taking care of my kids whenever I was so busy with work and for raising a wonderful man I can call husband. Thank you, Madivó for taking such good care of Sammy and Olivia for such a long time. For taking care of us too. For loving us as her own. Thank you, mom. For winning against all odds and coming all the way here to visit us and spend time with us and thank you, Dad, for complete and unconditional support, always.

And lastly, thank you to my amazing husband Guilherme. I could have never done this without you. We went through really hard times so I could finish my PhD and you were always standing by me; raising our son Samuel and now our daughter Olivia with me. It was never easy, but we are making it. We are winning together and it is because it's you, because it's us. Thank you, Sammy, for teaching me that no matter how hard I work; you matter the most. You are more important than any of this and knowing this has always made me put everything into perspective and be able to be the best mom I can before I can be a scientist. Thank you, Olivia, for turning everything upside down and bringing more love into our world and pushing me into trying to be better for you. You guys are everything and will always come first, so thank you for allowing and helping me to successfully finish the second place on the list.

CHAPTER 1: Introduction to translation initiation regulation

RNA is one of the most versatile molecules in biology. It performs a remarkable range of roles in all types of living cells. Not only does it carry the message from DNA to proteins, but it is also involved in regulating transcription through long non-coding RNA (lncRNA); regulating translation through 5'–UTR structures and 3'–UTR elements in messenger RNAs (mRNAs) and by microRNAs (miRNA), while also playing a part in chromosome end-maintenance and dosage compensation. Its ability to use direct read-out through base-pairing and finding specific zip-codes within any genetic code and the sophisticated three dimensional structures it folds into allows this molecule to have these diverse functions within the cell. In my work, I will focus on the importance of the untranslated regions in mRNAs and how the differences in their sequence and content result in different protein expression, serving as a layer of regulation that cells use to achieve homeostasis and proper protein translation.

Chapter two of this thesis is on the study of alternative spliced transcripts of the mRNA that encodes human PTBP1. PTBP1 is an RNA-binding protein that complexes with heterogeneous nuclear RNA. It associates with pre-mRNA in the nucleus and participates heavily in the process of alternative splicing, mRNA metabolism and transport. It has four repeats of RNA-recognition motifs (RRM) that bind to poly-pyrimidine tracts of RNAs required to regulate pre-mRNA splicing. During my work, I have observed that the mRNA of PTBP1 directly binds to eukaryotic translation initiation factor 3 (eIF3) in mammalian cultures cells. *PTBP1* mRNA has 3 different main transcripts with unique 5'–UTRs and different lengths of 3'–UTRs. These differences influence the level and quality of binding between eIF3 and the mRNAs. In addition to this, eIF3 binding is also distinct in the different stages of the cell cycle and could be acting as a way to globally regulate alternative splicing of mRNAs related to cell division progression by increasing or decreasing translation of PTBP1, one of the main alternative splicing factors. This study was important to curate the main alternative spliced isoforms of *PTBP1*. Before this work, the literature had confounding sequences and defective annotation regarding the untranslated regions in *PTBP1* mRNA. I curated and validated these isoforms by sequencing them and verifying their presence in mammalian cells. This work also improved the understanding of transcript-specific translation and how eIF3 might be involved, in addition to differences in the UTRs of the transcripts.

Chapter three of this thesis focuses on assessing the 3'–UTR length differences of transcripts in developing brain cells. Recent technical advances have made the thorough study of the transcriptome, and therefore untranslated regions, possible. Lately, studies in the mammalian brain have showed that what were previously thought to be long intergenic noncoding (lincRNAs) from loci downstream from protein-coding genes, are instead strongly distal 3'–UTRs generated by alternative cleavage and poly-adenylation. Global studies show that such extensions are prevalent in mammalian brains and that they collectively contain hundreds to thousands of conserved miRNA binding sites. This phenomenon may be yet another expansion of the scope of post-transcriptional regulatory networks in mammals and have a great impact on

the development of the nervous system. Information on 3'–UTR elements show they can regulate transcript stability, translation efficiency and targeting. Investigations in different animals have examined the heterogeneity of 3'–UTRs and observed that many of these transcripts are differentially expressed in a developmentally regulated or tissue specific way. For example, it has been observed that there is an increase in the average 3'–UTR length in neuronal tissue. Due to the high volume of transcriptomics data with the potential to understand mRNA heterogeneity, I have analyzed a highly-cited dataset with the goal of determining if the preference of expression of different length of 3'–UTR extensions in mRNAs of genes related to translation control differ during brain cell development. Such differences could be due to gene regulation through increased or decreased translation, mRNA stability, or due to mRNA localization based on different micro-RNAs that would be able to bind different types of 3'–UTR.

CHAPTER 2: PTBP1 mRNA isoforms and regulation of their translation

2.1 Abstract

Polypyrimidine tract-binding proteins (PTBPs) are RNA binding proteins that regulate a number of posttranscriptional events. Human PTBP1 transits between the nucleus and cytoplasm and is thought to regulate RNA processes in both. However, information about PTBP1 mRNA isoforms and regulation of PTPB1 expression remains incomplete. Here we mapped the major PTBP1 mRNA isoforms in HEK293T cells and identified alternative 5′ and 3′ untranslated regions (5′-UTRs, 3′-UTRs), as well as alternative splicing patterns in the protein coding region. We also assessed how the observed PTBP1 mRNA isoforms contribute to PTBP1 expression in different phases of the cell cycle. Previously, PTBP1 mRNAs were shown to crosslink to eukaryotic translation initiation factor 3 (eIF3). We find that eIF3 binds differently to each PTBP1 mRNA isoform in a cell cycle dependent manner. We also observe a strong correlation between eIF3 binding to PTBP1 mRNAs and repression of PTBP1 levels during the S phase of the cell cycle. Our results provide evidence of translational regulation of PTBP1 protein levels during the cell cycle, which may affect downstream regulation of alternative splicing and translation mediated by PTBP1 protein isoforms.

2.2 Introduction

PTBP1 was first discovered as a purified protein that bound to polypyrimidine tract regions of introns (Garcia-Blanco et al. 1989). Initially, PTBP1 was thought to be part of the splicing machinery, until U2AF65 was discovered as the splicing factor responsible for recognizing the poly(U) tracts at the 3′ splice site during the assembly of the spliceosome (Gil et al. 1991). PTBP1 has since been shown to regulate alternative exon selection during mRNA processing by repressing exon inclusion (Xue et al. 2009). Although PTBP1 acts as an alternative splicing (AS) factor in the nucleus, it also shuttles between the nucleus and cytoplasm. When PTBP1 is present in the cytoplasm, it is thought to be involved in posttranscriptional regulation, processes that require cap- independent translational control, RNA localization or changes in mRNA stability (Kamath et al. 2001; Romanelli et al. 2013). In addition to its role in molecular processes including splicing, polyadenylation, translation initiation, and mRNA stability, PTBP1 has recently been linked to the regulation of the cell cycle (Monzón-Casanova et al. 2018).

PTBP1 is a 57 kDa protein comprised of four RNA recognition motifs (RRMs) with a bipartite nuclear localization domain (NLD) and a nuclear export signal (NES) at the amino terminus of the protein (Pérez et al. 1997; Wollerton et al. 2001; Li and Yen 2002). The expression of PTBP1 is tightly regulated through alternative splicing events (Wollerton et al. 2004). Its 15 exons have previously been shown to be alternatively spliced into three major mRNA isoforms, termed *PTBP1-1*, *PTBP1-2*, and *PTBP1-4*. The first described isoform, *PTBP1-1* encodes a protein of 521 amino acids containing all four RRMs. The alternatively spliced isoforms, *PTBP1-2* and *PTBP1-4*, encode an additional 19 or 26 amino acids, respectively, between the RRM2 and RRM3 domains derived from exon 9 inclusion (Garcia-Blanco et al. 1989; Valcárcel and Gebauer 1997; Sawicka et al. 2008; Romanelli et al. 2013). Despite being very similar, the different isoforms have distinct roles in splicing and internal ribosome entry site (IRES)-mediated initiation of translation. The absence or length of the unstructured region between RRM2 and RRM3 results in differential recognition of target RNAs. These functional differences coupled with differing PTBP1 isoform ratios in different cell lines suggests that changes in relative PTBP1 isoform expression levels may be a cellular determinant of alternative splicing events (Wollerton et al. 2001; Gueroussov et al. 2015). For example, in the case of tropomyosin alternative splicing, PTBP1-4 represses exon 3 inclusion more than PTBP1-1 both in vivo and in vitro, whereas PTBP1-2 harbors intermediate activity (Wollerton et al. 2001). Additionally, differences in exon 9 skipping in *PTBP1* mRNAs have been found to affect the levels of many additional alternative splicing (AS) events, likely modulating the timing of transitions in the production of neural progenitors and mature neurons so as to affect brain morphology and complexity (Gueroussov et al. 2015).

In eukaryotic mRNAs, the 5′ and 3′ untranslated regions (5′- and 3′-UTRs) serve as major *cis*-regulatory control elements. RNA sequences and structures in the 5′-UTR and 3′-UTR can act as binding sites for translation initiation factors and other RNA binding proteins to influence the translational output of an mRNA and its lifetime in the cell (Hinnebusch et al. 2016). To date, how the alternatively spliced isoforms of *PTBP1* are connected to different 5′-UTRs and 3′-UTRs in *PTBP1* mRNA has not been determined. Several annotation databases, such as ENSEMBL (Ensembl Release 94) (Zerbino et al. 2018), FANTOM5 (Riken Center for

Integrative Medical Sciences [IMS]) (Noguchi et al. 2017), and NCBI Gene (O'Leary et al. 2015), have information on *PTBP1* isoforms. However, the information on UTRs differs across these databases. In ENSEMBL, the three main isoforms have distinct 5′-UTRs and a common 3′-UTR. In the NCBI Gene (refseq) database, *PTBP1* has common 5′ and 3′-UTRs. The FANTOM5 database (The FANTOM Consortium and the RIKEN PMI and CLST [DGT] 2014) only accounts for two distinct 5′-UTRs for *PTBP1* and a common 3′-UTR. Finally, the APASdb database for polyadenylation signals (You et al. 2015) reports two major polyadenylation sites within the *PTBP1* 3′-UTR. These libraries need to be reconciled into a comprehensive model of *PTBP1* transcript isoforms allowing further biochemical analysis of the regulatory pathways that influence *PTBP1* mRNA isoform production and translation.

To better understand the regulation of *PTBP1* mRNA isoform levels in the cell, we mapped the major *PTBP1* mRNA variants present in mammalian HEK293T cells. We analyzed the 5′-UTR elements using 5′-RACE (RLM-RACE) and long-read sequencing (Oxford Nanopore). We also mapped the 3′-UTRs and open reading frames. Using western blots and mRNA reporters, we determined how the *PTBP1* mRNA isoforms are translated in different stages of the cell cycle. Previous evidence revealed that human translation initiation factor eIF3, the largest translation initiation factor, crosslinks to the 5′-UTR elements of several messenger RNAs, including *PTBP1*. While bound to mRNAs, eIF3 acts to either activate or repress their translation (Lee et al. 2015). For this reason, we also probed eIF3 interactions with *PTBP1* mRNAs to determine whether eIF3 may act as a *trans*-acting factor regulating *PTBP1* isoform translation.

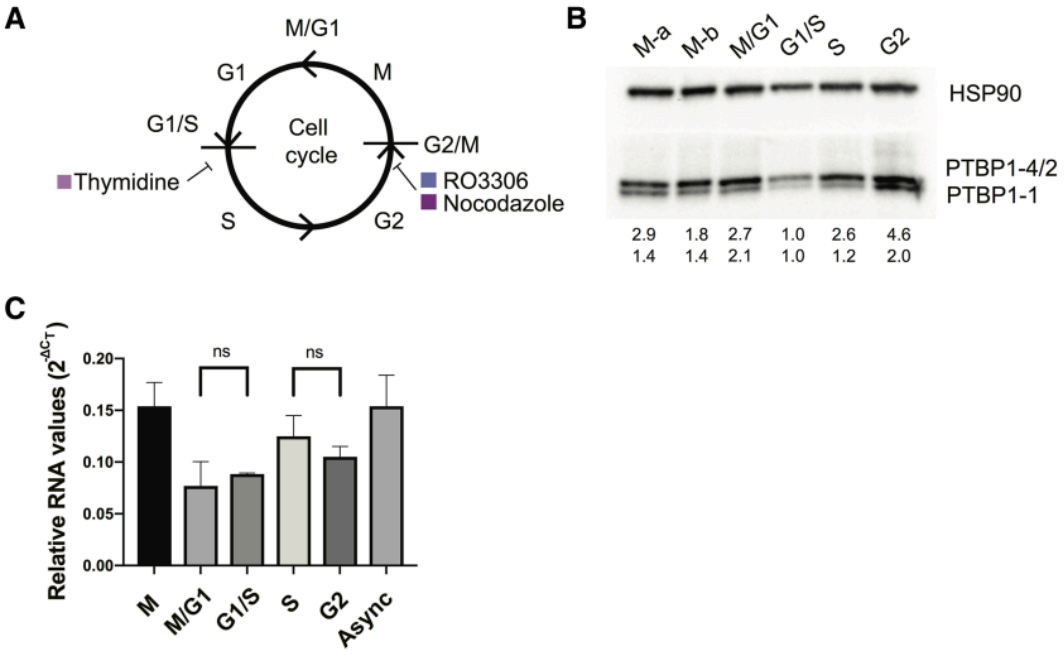

FIGURE 1. PTBP1 expression changes across the cell cycle. (A) Chemical inhibitors used to arrest cells at specific phases of the cell cycle: thymidine, arrest at G1/S, and RO3306 or Nocodazole, arrest at G2/M transitions. Cells were synchronized and collected at time points after release from the drugs. (B) Representative western blot of whole cell lysates of synchronized HEK293T cells prepared using synchronized samples. Separate methods to arrest G2/M were used. M-a and M-b samples were synchronized with the use of RO3306 and Nocodazole, respectively. PTBP1-2 and PTBP1-4 protein isoforms have similar sizes and comigrate in the gel. Below the gel are shown the amounts of the PTBP1 isoforms relative to that in G1/S phase, normalized to HSP90 levels. (C) Amounts of total PTBP1 mRNA were

Endogenous levels of PTBP1

Since PTBP1 has been implicated in regulating numerous processes including the cell cycle, we analyzed the endogenous levels of PTBP1 in HEK293T cells harvested in different stages of the cell cycle (Fig. 1A). We observed that PTBP1 isoforms vary dramatically during cell cycle progression. Cells harvested during the G2 or M phases had the highest levels of all three isoforms (PTBP1-1, PTBP1-2, PTBP1-4, Fig. 1B), with the upper band, comprising PTBP1-2 and PTBP1-4 (Wollerton et al. 2001), having a higher expression profile than PTBP1-1 regardless of cell cycle phase. All three isoforms exist at low levels during G1, and increase slightly during S, before a larger burst during G2/M occurs. Notably, *PTBP1* mRNA levels do not fluctuate as much as protein levels in the different stages of the cell cycle (Fig. 1C). Although we did not separate the contributions of translation and protein degradation, these results indicate that posttranscriptional regulation of PTBP1 expression occurs as a function of the cell cycle.

2.4 Mapping the 5'-UTR, CDS and 3'-UTR sequences in *PTBP1* mRNAs

To test whether *PTBP1* transcript isoform sequences in the ENSEMBL database are in agreement with the transcription start sites (TSS) in FANTOM5, we used RNA Ligase Mediated Rapid Amplification of cDNA Ends (RLM-RACE) and Nanopore sequencing of mRNAs extracted from HEK293T cells to map *PTBP1* transcripts (Fig. 2). Although both TSS in the FANTOM5 database were confirmed by RLM-RACE, we could not verify the presence of the 5′-UTR for ENSEMBL transcript ENST00000356948.10. Notably, our RLM-RACE data supports a different TSS for ENSEMBL transcript ENST00000349038.8, 7 nucleotides (nts) 5′ of the annotated TSS, in agreement with the TSS mapped in the FANTOM5 database (Fig. 2C,D). The longer TSS for this transcript is also in agreement with the fact that eIF3 crosslinks to nucleotides 5′ of the ENSEMBL-annotated TSS (Fig. 2B,D).

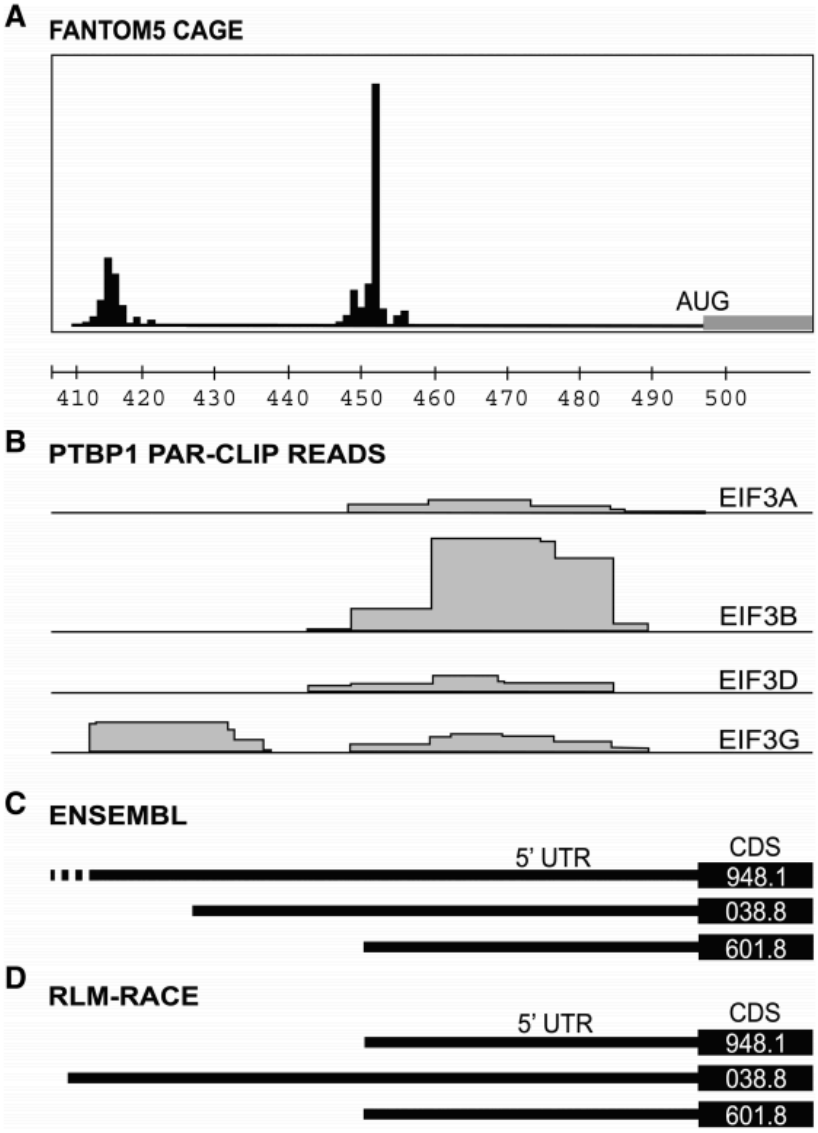

FIGURE 2. Database and experimental mapping of the three major transcript isoforms of PTBP1 mRNA. (A) Transcription start sites (TSS) determined by CAGE mapping in the FANTOM5 database (Riken Center for Integrative Medical Sciences [IMS]) (Noguchi et al. 2017), along with hg38 chromosome location, showing the last three digits of the chromosomal coordinates for the PTBP1 gene. (B) Sites of PTBP1 mRNA interaction with eIF3 mapped by photoactivatable RNA crosslinking and immunoprecipitation (PAR-CLIP) (Lee et al. 2015), by eIF3 subunit, indicated to the right. Coordinates of clusters are given in Table 2. (C) Annotated Ensembl transcripts for PTBP1, including the 5′ -UTR and the beginning of the CDS, as indicated by the last four digits of the Ensembl tag (i.e., ENST0000∗∗∗∗). (D) Experimentally determined 5′ -UTR elements in PTBP1 mRNAs determined by RLM-RACE. In all panels, the transcripts are vertically aligned with the chromosomal coordinates in panel A.

PTBP1 has three major protein isoforms that only differ with respect to exon 9 inclusion. *PTBP1-1* lacks exon 9 completely, *PTBP1-2* includes only part of exon 9 and *PTBP1-4* contains the full sequence coding for exon 9 (Fig. 3A). Although differences in exon properties have been implicated in the different biological roles of PTBP1, the connectivity between the different CDS variants and the mRNA 5′-UTR and 3′-UTR ends is not known. To map the 5′-UTRs for each predicted CDS in the *PTBP1* transcript isoforms, we used a variation of the RLM-RACE methodology (Fig. 3B). For each *PTBP1* exon 9 isoform, we observed a single species by RLM-RACE, indicating one major form of 5′-UTR for each CDS variant (Fig. 3C). This was confirmed by a second reaction in which we used a common inner primer to the 5′ adaptor and reverse primer to the common CDS region upstream of exon 9 (Fig. 3B, primers Fin and R4) to assess the amount of different *PTBP1* 5′-UTRs in the samples (Fig. 3D), which revealed two major 5′-UTR species. After sequencing the reactions in Figure 3C individually we were able to determine the exact sequence of each transcript up to the cap region. Isoform *PTBP1-1*, which lacks the exon 9 sequence, extends to the 5′ end of the long 5′-UTR, matching the upstream TSS mapped in FANTOM5 (Fig. 2A) and the RLM-RACE experiment described above (Fig. 2D). In contrast, isoforms *PTBP1-2* and *PTBP1-4*, which encode the truncated or full exon 9, respectively, each have the short 5′-UTR, with the downstream TSS mapped in FANTOM5 (Figs. 2A,D, 3A).

We also determined the 3′-UTR sequences of *PTBP1* transcript isoforms in HEK239T cells. The APASdb database (You et al. 2015), which contains precise maps and usage quantification of different polyadenylation sites, contains two major polyadenylation sites for *PTBP1* (Fig. 3E). We used this information to design specific primers to determine the presence of each poly(A) site in total RNA extracted from HEK293T cells. By using a forward primer that recognizes the splice junction specific to each transcript upstream of exon 9, we could determine the 3′-UTR length of each isoform by using a reverse primer on a poly(A) adapter (Fig. 3B). The resulting amplification pattern could be visualized by agarose gel (Fig. 3F) and then by sequencing. Using this amplification strategy, we observed all three *PTBP1* exon 9 isoforms predicted in the ENSEMBL database to have two different lengths of 3′-UTR resulting from the predicted poly(A) sites in the APASdb database (Fig. 3E,F), and possibly a third. Taken together, the present experiments define six *PTBP1* transcript isoforms in HEK293T cells (Fig. 3G).

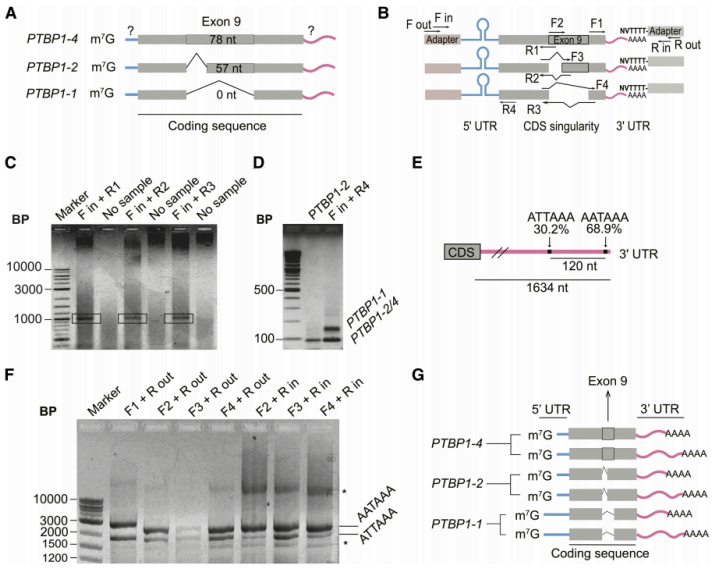

FIGURE 3. Mapping of the UTR elements of PTBP1 mRNAs. (A) Scheme of the composition of the 5′ -UTR of PTBP1 transcripts according to the coding sequence content, with question marks indicating the regions to be mapped. (B) Design of RLM-RACE experiments performed to determine the relationship between 3′ -UTR elements, exon 9 boundaries, and 5′ -UTR elements in PTBP1 mRNAs. (C) Agarose gel of final PCR reaction for the 5′ -UTR RLM-RACE. Bands in the black rectangles were extracted for sequencing. (D) Agarose gel showing the presence of two known and mapped 5′ -UTR lengths during RLM-RACE, using primers that anneal to all three PTBP1 isoforms. PTBP1-2 sequence was used as a control. (E) Representation showing high usage polyadenylation sites on PTBP1 mRNA 3′ -UTR (not to scale). Data from You et al. (2015). (F) Agarose gel of PCR reactions following RLM-RACE to identify the polyadenylation sites of PTBP1 transcript isoforms. (∗) Unidentified bands that only appear after second round of PCR. (G) Model for major PTBP1 transcript isoforms in HEK293T cells based on experimental observations. Blue bars, evidence for the existence of two lengths of the 5′ -UTR; pink bars, evidence that each transcript has at least two alternative polyadenylation sites, resulting in a long or short 3′ -UTR. Gray thick bar represents the alternatively spliced isoforms involving exon 9

2.5 *PTBP1* 5′-UTR and 3′-UTR contributions to translation regulation

In order to assess whether the differences in PTBP1 expression through the cell cycle are related to the 5′-UTRs and 3′-UTRs in *PTBP1* mRNAs, we used *Renilla* luciferase reporter mRNAs with the different *PTBP1* 5′-UTR and 3′-UTR elements in cell based assays. Using transfections of reporter mRNAs, we first assessed the relative translation levels of each reporter with respect to the cell cycle (Figs. 4, 5). We used 6 h transfections, as previous results have indicated that these early time points are in the linear range for mRNA transfections (Bert 2006). We determined that the mRNA was not degraded during the 6 h of the experiment (Fig. 5C). During the G2 and M phases of the cell cycle, the reporter transcript with the long *PTBP1* 5′-UTR and short *PTBP1* 3′-UTR (Fig. 4A) had the highest translation efficiency (Fig. 5). During the G1 and S phases of the cell cycle, the reporter transcript with the long *PTBP1* 5′-UTR and the long *PTBP1* 3′-UTR had higher translation efficiency (Fig. 5). Although these experiments are not normalized across cell cycle phases, due to the fact each experiment was carried out separately, we found the experiments synchronized in the G2 and M phases correlated well with unsynchronized cells (Fig. 5D). Furthermore, translation of the reporter mRNAs in the G2 and M phases also correlate well with each other (Fig. 5E). In contrast, translation in G1 and S synchronized cells did not correlate with the unsynchronized cells (Fig. 5D) but rather correlated with one another (Fig. 5F). These results indicate that translation in the G2 and M phases, even though relatively short time-wise (~2 h total) with respect to the entire cell cycle, dominate translation of the reporter mRNAs with *PTBP1* 5′-UTR and 3′-UTR elements. These results are consistent with endogenous PTBP1 levels observed by western blotting (Fig. 1B), suggesting that posttranscriptional regulation of PTBP1 levels occurs to a significant extent at the level of translation during the G2 and M phases of the cell cycle.

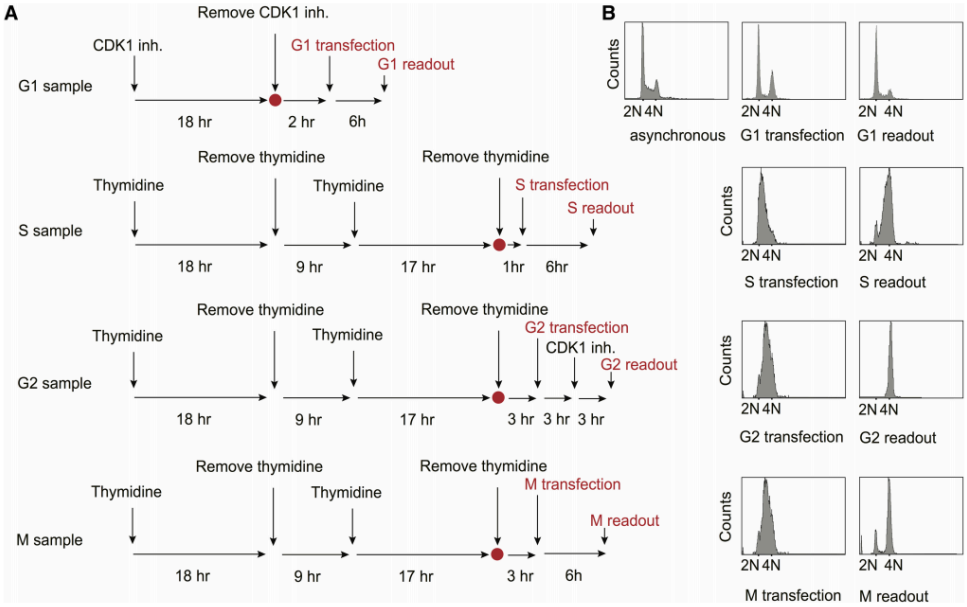

FIGURE 4. Timecourses for mRNA transfections, incubation, and luminescence readout. (A) Schematic overview of HEK293T synchronization protocol and of luciferase mRNA reporter transfections. RO-3306 (6 μM) was used as the CDK1 inhibitor. (B) G1, S, G2, andMsamples were transfected and assessed as outlined in A. FACS analysis is shown at the time of transfection and readout for G1, S, G2, andMsamples. Note: G2 andM samples were transfected at the same time. For G2, RO3306 was maintained in the media during the experiment to maintain the block in G2, due to the fast transition observed between G2 and M phases. two lengths of the 5′ -UTR; pink bars, evidence that each transcript has at least two alternative polyadenylation sites, resulting in a long or short 3′ -UTR. Gray thick bar represents the alternatively spliced isoforms involving exon 9.

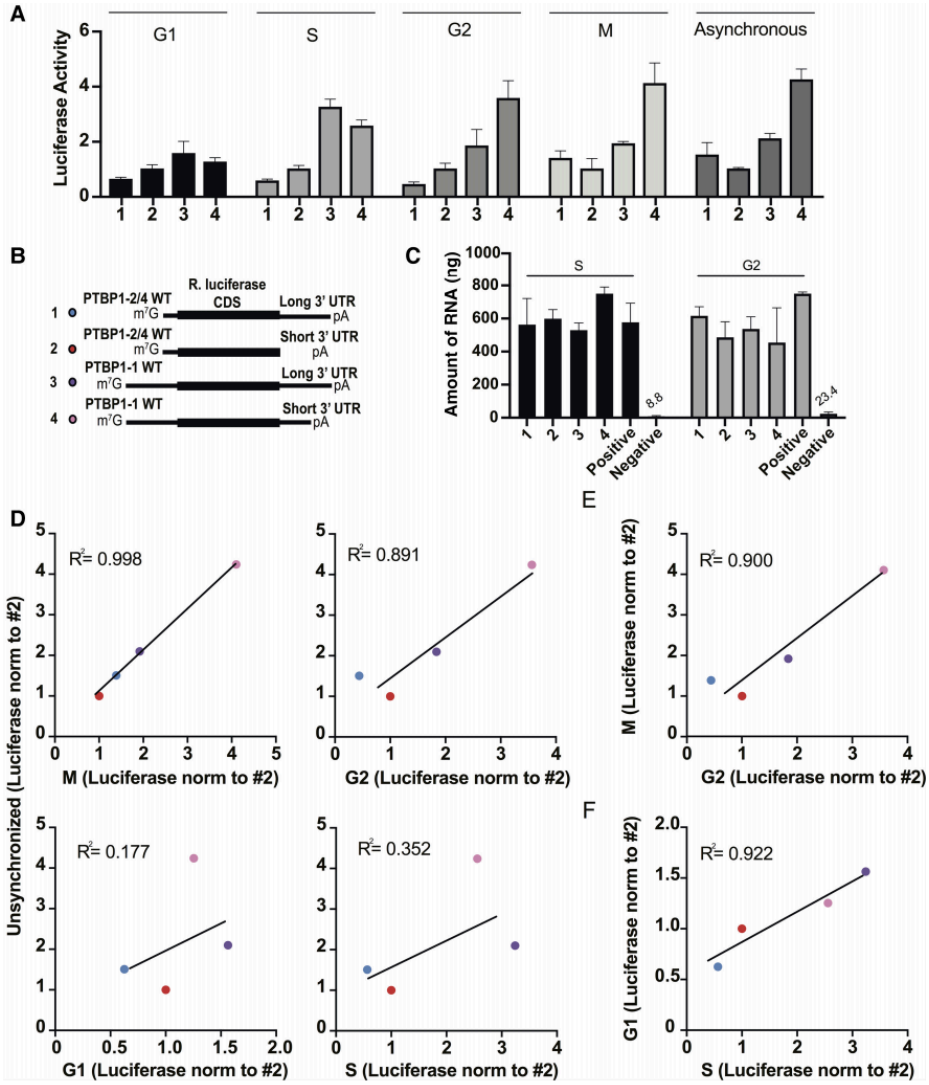

FIGURE 5. Translation profile of luciferase reporters with PTBP1 5′ -UTR and 3′ -UTR elements in different phases of the cell cycle. (A) Luciferase reporter readout for the experiments as diagrammed in Figure 4. All experiments were carried out in biological triplicate, with standard deviations shown. (B) Schematics of the luciferase reporters, with PTBP1 5′ -UTR and 3′ -UTR elements. (C) Determination of mRNA stability during the timecourse of the transfection experiment (6 h). We used the PSMB6 5′ -UTR and short random 3′ -UTR as positive control and water transfection as a negative control for background. (D) Luciferase activity from transcripts from the G1, S, G2, and

2.6 Implications of eIF3 binding to *PTBP1* mRNA on its translation

The results above suggest that translational regulation plays an important role in controlling PTBP1 isoform expression. Given the fact that eIF3 crosslinks to specific sequences in the 5′-UTR of *PTBP1* mRNA (Lee et al., 2015), we first confirmed that *PTBP1* mRNA binds eIF3 specifically in different cell types (Fig. 6A), by immunoprecipitating eIF3 from cell lysates using an antibody against EIF3B (Lee et al. 2015). In separate experiments, eIF3 immunoprecipitated from HEK293T cell lysates using an antibody against EIF3B (Lee et al. 2015) bound to all three endogenous *PTBP1* coding sequence isoforms in HEK293T cells (Fig. 6B). We next tested the importance of these eIF3–5′-UTR interactions in regulating *PTBP1* translation, also in HEK293T cells. We used luciferase reporter assays to measure differences in the translation output of mRNAs with the longer *PTBP1* 5′-UTR, which contains two sites of eIF3 crosslinking (Fig. 2B), or lacking regions known to bind eIF3 (Fig. 6C). In untreated HEK293T cells, individually deleting eIF3 crosslinking sites had a minimal impact on translation, whereas deleting both eIF3-interacting regions increased translation of these mRNAs (Fig. 6D).

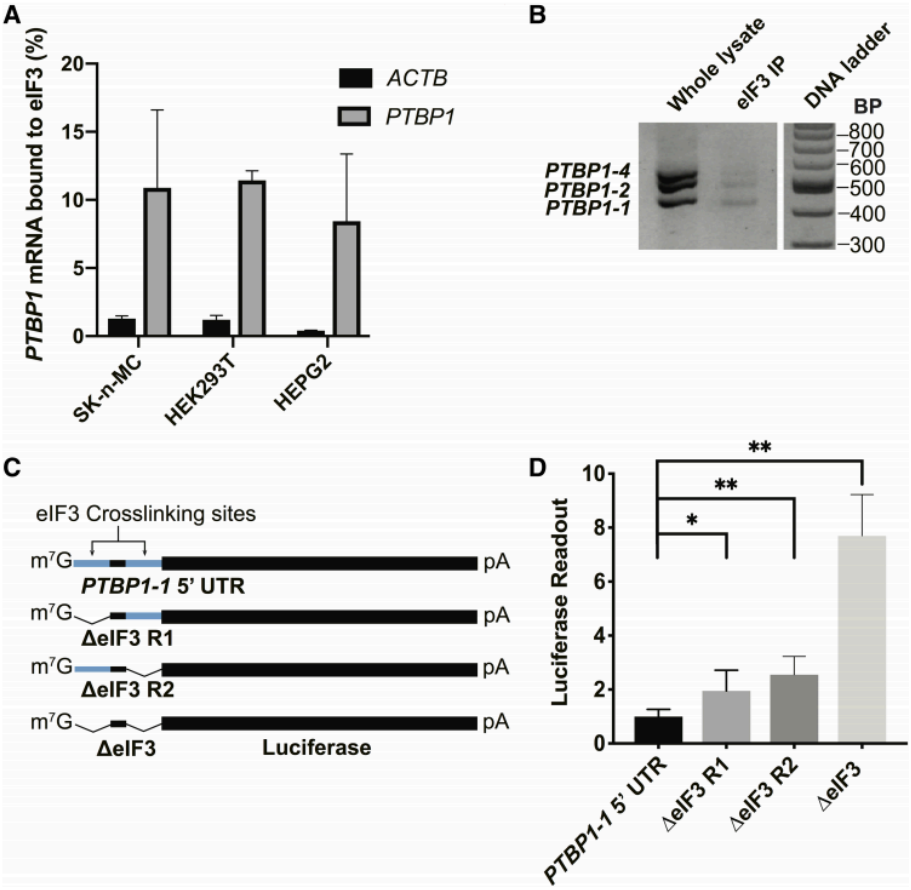

FIGURE 6. Binding of reporter mRNAs containing sites of interaction with eIF3. (A) qPCR quantification of PTBP1 mRNA bound to eIF3, with ACTB used as a negative control. Cell lines SK-n-MC, HEK293T, and HEPG2 were used. (B) PTBP1 mRNA exon 9 coding sequence isoforms that immunoprecipitate with eIF3, as determined by RT-PCR, and resolved on a 2% agarose. DNA ladder shown on the right. (C) Schematic of PTBP1 5′ -UTR–luciferase reporter mRNAs. WT, wild-type; ΔeIF3, deletion of eIF3 PAR-CLIP clusters, nucleotide positions 25– 49 (Region 1, R1), and/or 58–86 (Region 2, R2) for the PTBP1 transcript with the long 5′ -UTR (GenBank accession NM_002819). (D) Luciferase activity in HEK293T cells transfected with mRNAs containing PTBP1 5′ -UTR elements with or without deletions of the eIF3 crosslinking sites in Region 1 (R1) and/or Region 2 (R2). Experiments were carried out in biological triplicate, with standard deviation shown, significant with (∗) $P > 0.01$, (∗∗) $P > 0.001$.

To check if the 5′-UTR of *PTBP1* is sufficient for eIF3 binding, and whether both lengths of *PTBP1* 5′-UTR bind similarly to eIF3 across the cell cycle, we designed mRNAs with either the long or short *PTBP1* 5′-UTR sequences upstream of a luciferase open reading frame. We also tested a reporter with a mutated 5′-UTR in which the sequences that crosslink to eIF3 were deleted (Figs. 2B, 7A). These mRNAs were transfected into HEK293T cells, and the cells were collected in different stages of the cell cycle (Fig. 7C). We then immunoprecipitated eIF3 from cell lysates as above (Lee et al. 2015), followed by RNA extraction and quantitative PCR using primers for the luciferase CDS (Fig. 7C). Upon deletion of the eIF3 crosslinking sites, eIF3 no longer bound to the reporter mRNAs (Fig. 7D). Notably, although the longer *PTBP1* 5′-UTR interacts with eIF3 more efficiently than the short 5′-UTR, both species of 5′-UTR bind to eIF3 more efficiently during the S phase and less so during G2, and even less during G1 (Fig. 7D). In the above immunoprecipitation experiments, we used a random 3′-UTR instead of the 3′-UTR elements derived from *PTBP1* transcript isoforms to assess the influence of the *PTBP1* 5′-UTR. To test whether eIF3 binding might also be influenced by the *PTBP1* 3′-UTR, we designed chimeric mRNAs with different combinations of *PTBP1* 5′-UTR and *PTBP1* 3′-UTR, using the same reporter system (Fig. 7B). Similarly to the 5′-UTR experiment (Fig. 7D), binding of the mRNAs containing the *PTBP1* 3′-UTR to eIF3 is more prevalent during the S phase compared to the other cell phases. Interestingly, the length of the 3′-UTR interacting with eIF3 changes as cell phases progress, with a switch happening during the mitotic phase (Fig. 7E). Altogether, these results indicate that eIF3 binds to *PTBP1* mRNAs likely by interacting with both 5′-UTR and 3′-UTR elements in a cell cycle dependent manner (Fig. 7).

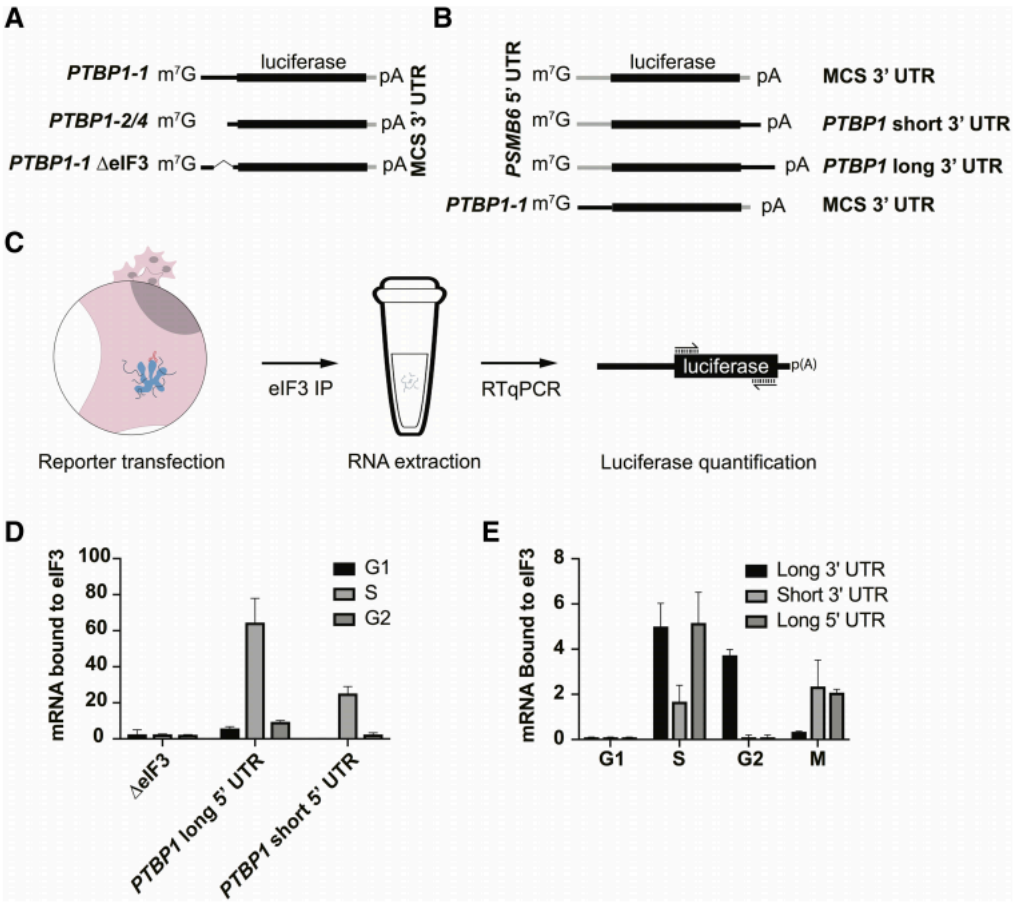

FIGURE 7. Differential binding of PTBP1 UTR elements to eIF3 across the cell cycle. (A,B) Schematics of the luciferase reporters used in the experiments. (C) Schematic of the transfection, immunoprecipitation, and quantification method used to determine luciferase reporter mRNA binding to eIF3. (D) Distribution of binding to eIF3 across the cell cycle for the different PTBP1 5′ -UTR elements as well as the deletion mutant. (E) Distribution of binding to eIF3 across the cell cycle for the different PTBP1 3′ -UTR elements, as well as the long form of the PTBP1 5′ -UTR. Binding experiments were carried out in biological triplicate, with standard deviation shown. Luciferase arbitrary units were normalized to WT for graphing.

2.7 Discussion

A transcript set is the collection of mRNA isoforms that originate from a given genomic sequence. Transcripts are defined by introns, exons, UTRs, and their positions. Human transcript set information is stored in large databases and browsers such as ENSEMBL, REFSEQ, and UCSC (Zhao and Zhang 2015). However, cases in which the annotations of isoforms are inconsistent across databases are not uncommon (Brenner 1999; Schnoes et al. 2009; Promponas et al. 2015). Given the existence of overlapping, variable transcript isoforms, determining the functional impact of the transcriptome requires identification of full-length transcripts, rather than just the genomic regions that are transcribed (Pelechano et al. 2013). While working with *PTBP1* mRNAs we noticed that sequences available in the ENSEMBL and FANTOM5 databases had discrepancies with respect to the TSSs of the major mRNA transcripts (Fig. 2). We therefore decided to validate the major mRNA isoforms for *PTBP1* as the basis for future functional analysis of posttranscriptional regulation of PTBP1 expression. We were able to confirm at least six mRNA forms (Fig. 3G). These mRNA isoforms had differences in the 5′-UTR, coding sequence and 3′-UTR, suggesting that PTBP1 protein isoform expression may be regulated in multiple ways. PTBP1 is a pleiotropic protein, functioning in a variety of cellular processes. It is still unclear if the multiple activities of PTBP1 share a mechanistic pathway and more importantly how PTBP1 could act in posttranscriptional regulation in a tissue-specific way that is singular to the physiology of a certain set of cells. Although PTBP1 has been extensively studied, the multiple *PTBP1* transcript isoforms we have identified will now enable biochemical analysis of *PTBP1* mRNA regulation and function in different stages of the cell cycle.

Identifying RNA exon–exon connectivity remains a challenge when dealing with unknown mRNA isoforms. By combining long-read sequencing, and biochemical validation, we were able to fully characterize *PTPB1* transcript isoforms. We used nanopore long-read sequencing with the goal to resolve connectivity between 5′-UTR, CDS and 3′-UTR elements of *PTBP1* mRNAs. However, due to the inability of long-read sequencing to accurately reach the 5′ end of mRNA transcripts (Workman et al. 2018), we complemented nanopore sequencing with RNA Ligase Mediated Rapid Amplification of cDNA Ends (RLM-RACE) in order to determine the full length of *PTBP1* mRNA isoforms present in HEK293T cells. This approach should be useful to identify the collection of PTBP1 variants in different cell types or culture conditions (Lundberg et al. 2010). *PTBP1* mRNA has three major isoforms in the coding sequence that differ from each other at exon 9. *PTBP1-1* lacks exon 9, *PTBP1-2* has a partial sequence of exon 9 and *PTBP1-4* has full-length exon 9. Because there are three different coding sequences (CDS) (Fig. 3A), resulting in three different proteins, and two distinct lengths of 5′-UTR, we aimed at determining the exact full-length sequence of each transcript. We found that *PTBP1-1* bears the longer 5′-UTR and *PTBP1-2* and *PTBP1-4* both bear the shorter 5′-UTR. There is only one visible band in the agarose gel for each transcript, meaning that there is only one major form of the 5′-UTR for each transcript (Fig. 3C). Consistent with the APASdb database for alternative polyadenylation sites (You et al. 2015), we identified two alternative polyadenylation sites with significant usage, resulting in each of the three major *PTBP1* transcripts having two distinct 3′-UTR lengths (Fig. 3E,F).

The mapping of all major *PTBP1* transcripts in HEK293T cells (Fig. 3G) generated the information necessary for the biochemical analysis of *PTBP1* translational regulation. Translational control elements can be located within the 5′-UTR and the 3′-UTR, with overall translation being affected by characteristics such as length, start-site consensus sequences as well as the presence of secondary structure, upstream AUGs, upstream open reading frames (uORFs) and internal ribosome entry sites (IRESs), and binding sites for *trans*-acting factors (Wilkie et al. 2003; Ma and Mayr 2018). UTR elements have been found to be involved in regulating cell cycle dependent translation. For example, histone translational control in Leishmania requires both 5′ and 3′-UTRs to properly restrict H2A translation to the S phase (Abanades et al. 2009). Differences in 3′-UTR length due to alternative polyadenylation have also been shown to result in acceleration of the cell cycle in cancer cells (Wang et al. 2018). We found that differences in the length of *PTBP1* UTRs result in altered translational efficiency as the cell cycle progresses (Figs. 4–7), which may reflect the need to regulate PTBP1 protein isoform translation quickly depending on cellular demands (Sonenberg 1994; Pesole et al. 2001; Mayr 2017).

We previously found that eIF3 binds to *PTBP1* through its 5′-UTR (Fig. 2B; Lee et al. 2015). Here we found both lengths of 5′-UTR are able to bind to eIF3 through two different sequence regions (Fig. 7). We also found that eIF3 can bind to the *PTBP1* 3′-UTRs (Fig. 7E). Interestingly, the different UTR lengths have differing impacts on translation (Figs. 4–6) and eIF3 binding (Fig. 7) in a cell cycle dependent manner. However, there is no obvious correlation between eIF3 binding and translational output of the mRNAs, indicating the role of eIF3 in PTBP1 translational regulation is more complex than simple binding of eIF3 to the transcripts. Although eIF3 binding to the 5′-UTR is likely direct (Lee et al. 2015), eIF3 binding to 3′-UTRs may be more common than previously appreciated, and may have been missed in (Lee et al. 2015) due to the sequencing depth and/or the types of contact to eIF3 involved. For example, eIF3 binding to the *PTBP1* 3′-UTR may require *trans*-acting factors.

Binding of eIF3 to *PTBP1* mRNA isoforms is most abundant during the S and G2 phases, with the length of the 3′-UTR seeming to influence the extent of eIF3 binding. During S phase, binding is mediated predominantly through the long 3′-UTR and through the 5′-UTR, which correlates with overall repression of translation (Figs. 1, 5, 6). During G2, eIF3 interacts only with mRNAs bearing the long 3′-UTR, which correlates with repression of translation of these transcripts (Fig. 5). Consistent with eIF3 acting as a repressor, previous observations indicate that long 3′-UTRs often repress translation (Szostak and Gebauer 2013; Yamashita and Takeuchi 2017). Future experiments will be required to establish a mechanistic basis for isoform specific eIF3 repression of *PTBP1* mRNA translation in the S and G2 phases of the cell cycle. Although we did not measure eIF3 levels in different stages of the cell cycle in this study, several groups have shown that some eIF3 subunits have different expression patterns throughout the cell cycle. Subunit EIF3F expression peaks in S and M phases in A569 cells (Higareda-Mendoza and Pardo-Galván 2010), and subunit EIF3A is also translated more during the S-phase (Dong et al. 2009). Depletion of EIF3B has been shown to decrease the levels of S-phase and G2/M phase cyclins in a bladder cancer cell line (Wang et al. 2013) and EIF3B/C depletion studies showed a profound cell size increase in G1 followed by a decrease in size during S-phase (Schipany et al. 2015).

Recently, PTBP1 has been found to be important for cell cycle progression. For example, PTBP1 enables germinal center B cells to progress through the late S phase of the cell cycle

rapidly (Monzón-Casanova et al. 2018). In addition, knockout of Ptbp1 in mice results in embryonic lethality due to prolonged G2 to M progression (Shibayama et al. 2009). Notably, we find that PTBP1 expression is highest during the G2 and M cell cycle phases (Fig. 1B), which could be explained by the increase in the cell's demand for PTBP1 in late S phase for proper cell progression. Although several studies have shown global protein synthesis is repressed during mitosis (Fan and Penman 1970), a number of transcripts escape translational repression during M phase (Wilker et al. 2007; Marash et al. 2008; Ramírez-Valle et al. 2010; Stumpf et al. 2013; Tanenbaum et al. 2015; Park et al. 2016). Notably, some of these studies used Nocodazole as a synchronizing agent, which has been shown to disrupt translation (Coldwell et al. 2013) and proper cell cycling (Cooper et al. 2006), limiting the utility of these experiments for comparing effects on specific transcripts. However, alternative approaches using RO3306 (Tanenbaum et al. 2015) still show a modest global reduction in translation during mitosis, with more pronounced effects on a small subset of transcripts. Since the levels of *PTBP1* mRNA remain relatively unchanged even as protein abundance increases substantially (Fig. 1), posttranscriptional regulation seems to be central to PTBP1 expression in the S to G2/M transitions, at least in HEK293T cells. With curated information on *PTBP1* mRNA isoforms present in HEK293T mammalian cells (Fig. 3G), it will now be possible to dissect the posttranscriptional regulatory mechanisms involved in cell cycle dependent expression of PTBP1 isoforms, and the downstream physiological consequences.

2.8 Material and methods

Cells and transfections

Cells and transfections Human HEK293T cells were cultured in DMEM (Invitrogen) supplemented with 10% of Fetal Bovine Serum (FBS) (Seradigm) and 1% Pen/Strep (Gibco, cat. # 15140122). RNA transfections were performed using Mirus TransIT-mRNA Transfection Kit (cat. # MIR 2250), with the following modifications to the manufacturer's protocol. Sixteen hours before transfection, HEK293T cells were seeded into opaque 96-well plates to reach ~80% confluence at the time of transfection. For each well, 9 μL of prewarmed OptiMEM (Invitrogen) was mixed with 250 ng of RNA, 0.27 μL of Boost reagent and 0.27 μL of TransIT mRNA reagent. Reactions were incubated for 3 min at room temperature, added drop-wise to the well, and luciferase activity was measured 6–8 h (as indicated) after transfection, using the Renilla Luciferase assay kit (Promega, cat. # E2820) and a Microplate 309 Luminometer (Veritas). Transfections were done in triplicate and on two different occasions using HEK293T cells.

For G1 transfections and luminescence readouts, cells were grown to 30% confluence and compound RO3306 (Vassilev et al. 2006; Tanenbaum et al. 2015) was added to a final concentration of 6 μM. Cells were incubated for 18 h. After 18 h, cells were released and incubated for 2 h with fresh media, to allow the cells to recover before mRNA transfection. After this time, we examined the cells by bright field microscopy to ensure they were well attached to the dish. Cells were transfected with the desired mRNA and luminescence was then measured after 6 h of incubation. For S-phase transfections and readouts, cells were grown to 20% confluence in standard media and thymidine was added to a final concentration of 2 mM. Cells were incubated for 18 h in a tissue culture incubator, followed by two washes of HBSS media (Invitrogen) to remove the thymidine. Fresh media was added and cells were incubated for 9 h, at which point thymidine was added again to a final concentration of 2 mM. Cells were incubated in a tissue culture incubator for 15 h, then washed with HBSS media (Invitrogen) and released into fresh media. After 1 h of incubation to allow the cells to recover, cells were transfected with desired mRNA and luminescence was measured after 6 h. For G2 transfection and luminescence readouts, cells were synchronized with the same protocol as for S phase. After release, however, they were incubated for 4 h before mRNA transfection. After mRNA transfection, cells were incubated for 3 h and RO3306 was added at a final concentration of 6 μM. This guaranteed the cells would not progress into M phase before luminescence was measured. Cells were then incubated for 3 h prior to assessing luminescence. For M transfection and readouts, cells were treated exactly as G2, except RO3306 was not added, allowing them to progress into M phase after 4 h. Luminescence was measured after 6 h of incubation. Two batches of synchronization were done and transfections were performed in triplicate in each of the batches.

Cell cycles analysis

Cells were harvested and washed twice with phosphate-buffered saline (PBS) followed by fixation with 80% ethanol for 30 min at room temperature. Cells were then collected by centrifugation and stained with 50 μg/μL propidium iodide. The cells were then treated with 100 μg/μL RNAse for 15 min at 37°C followed by analysis using a BD Fortessa Flow Cytometer. Cell cycle distribution was analyzed using BD FACSDiva 7.0 software. We used two-parameter flow cytometry with forward (FSC) and side scatter (SSC) information, along with PE-TexasRed signal on an untreated (not synchronized) sample to determine the size distribution and locations of G1 and G2 phases on the plot (SSC vs. FSC; FSC vs. FSC and PE-TexasRed vs. FSC). Next, we analyzed the samples collected at different times after synchronization to assess the cell cycle distribution at each time point. We used this information to set the timing of the luciferase reporter experiments, as shown in Figure 4.

Analysis of luciferase reporters during different phases of the cell cycle

In experiments to analyze the influence of *PTBP1* 5′ -UTR and 3′ - UTR elements, luciferase readings after very short times after transfection were too noisy to be interpretable. We therefore timed the experiments such that luminescence readout was conducted in the desired phase of the cell cycle. We relied on normalizing luciferase expression from the transcript isoforms internally to each experiment in Figure 5. When comparing the relative translation of each transcript isoform to that observed in unsynchronized cells, the relative translation of each isoform in the experiments spanning G2 and M closely matched that in unsynchronized cells. The high correlation between these experiments and the unsynchronized cells indicates that translation of the PTBP1 mRNAs is highest in G2 and M phases, and relatively low in G1 and S phases.

To analyze the mRNA levels after 6 h of incubation, we transfected the desired reporters (750 ng of RNA per well) following the synchronized transfection protocol in 24 well plates in triplicate. After 6 h of incubation, cells were harvested and lysed with NP40 lysis buffer and 10 μL was removed for western blot control. The remaining 50 μL was extracted with the use of an RNeasy Mini Kit (QIAGEN) and RNA concentrations were assessed on nano drop and normalized to 50 ng/μL. Two hundred and fifty nanograms of RNA were used to perform RT-PCR with the use of a Superscript III Reverse Transcriptase kit (Thermo Fisher scientific, cat. #18080044). After reverse transcription, samples were treated with RNAse H enzyme for 30 min at 37°C. qPCR was done using 500 ng of cDNA and Sybr Green master mix with run conditions as follows: 95°C for 15 sec, followed by 40 cycles at 95°C for 15 sec, 60°C for 60 sec, and 95°C for 1 sec. Standard curves for assessing primer annealing and amplification were calculated for the ACTB primers and for the luciferase primers and the absolute amount of RNA was then calculated based on the equation given for each curve. Final RNA amounts were normalized to ACTB amounts to control for cell number differences across samples. Although we performed this control for mRNA stability, determination of cytoplasmic mRNA levels may be complicated by the route of mRNA entry into the cell due to the transfection protocol (Kirschman et al. 2017).

Plasmids

To generate the luciferase plasmids used on this work, sections of either the PTBP1 5′ -UTR (GenBank accession NM_002819) or the PTBP1 3′ -UTR were first amplified from human cDNA extracted from HEK293T cells. These were then placed downstream from a T7 RNA polymerase promoter using overlap extension PCR and InFusion cloning. The 5′ -UTRs were then inserted together with Renilla luciferase into plasmid pcDNA4 V102020 (Invitrogen). The eIF3 binding mutants and PSMB6- PTBP1 chimeras were made by insertional mutagenesis with primers annealed to the pcDNA4 plasmid digested at the desired insertion site. Primers and sequences are included in Table 1.

Western Blot

Western Blot analysis was carried out using the following antibodies: anti- EIF3B (Bethyl A301-761A), anti-HSP90 (BD 610418), and anti-PTBP1 (MABE986, clone BB7); all antibodies were used with a 1:10000 dilution.

In vitro transcription

RNAs to be used for transfections were made by in vitro transcription with T7 RNA polymerase (NEB). For luciferase mRNAs, transcription was performed in the presence of 3′ -O-Me-m7 G (5′) ppp(5′)G RNA Cap Structure Analogue (NEB), using linearized plasmid as template, then polyadenylated using poly(A) polymerase (Invitrogen). RNAs were purified by phenol–chloroform extraction and ethanol precipitation or using the RNA clean and concentrator kit (Zymo Research). RNA quality was verified using 2% agarose gels, to ensure mRNAs were

intact before transfection. RNAs were quantified using nanodrop and agarose gels to account for

TABLE 1. Primers used in this study

Primer	Sequence 5’ - 3’
Ensembl 5' - UTR F	GCCACGTACCCACTCTCAAGAT
Ensembl 5' - UTR R	GGGACCCAGAGAAATCGCAG
Ensembl 5' - UTR 2 F	TTCTGGCCAGTGGGAGGTGC
RefSeq 5' - UTR F	TGCGGGCGTCTCCGCC
PTBP1-1 extended 5' - UTR F	GTGAGTCTATAACTCGGAGCCGT
PTBP1-1 5' - UTR F	TGGGTCGGTTCCTGCTATTCCG
PTBP1-2 5' - UTR F	ATTCCGGCGCCTCCACTCCG
PTBP1 ATG F	TCTGCTCTGTGTGCCATGGAC
PTBP1 5' - UTR End F	GCGGGTCTGCTCTGTGTGCC
PTBP1 General R (R4)	AGATCCCCGCTTTGTACCAACG
PTBP1 Exon 3/4 junction R	CATTTCCGTTTGCTGCAGAAGC
CDS F (Exon 6)	CCTCTTCTACCCTGTGACCC
CDS F (F1)	AAGTCCACCATCTAGGGGCA
Unique PTBP1-4 F (F2)	GTGCACCTGGTATAATCTCAGCCTCTCC
Unique PTBP1-2 F (F3)	CGGCCTTCGCCTCTCCGTAT
Unique PTBP1-1 F (F4)	GCCTTCGGCCTTTCCGTTCC
PTBP1-4 Exon junction R (R1)	TACCAGGTGCACCGAAGGCC
PTBP1-2 Exon junction R (R2)	ATACGGAGAGGCGAAGGCCG
PTBP1-1 Exon junction R (R3)	GGAACGGAAAGGCCGAAGGC
PTBP1 Exon 11 R	AGAGGCTTTGGGGTGTGACT
PTBP1 Exon 11 R2	ACTTGCCTGTCGCTCTATCTTCACCGTAGACGCCGAAAAGAA
PTBP1 3UTR2	ACTTGCCTGTCGCTCTATCTTCACACAGGGCTAGACAAGGGA
PTBP1 3UTR1	ACTTGCCTGTCGCTCTATCTTCGTAAGGCAACGGAATGTGCG

Primers used in this study (continued)

Primer	Sequence 5’ - 3’
UMI	ACTTGCCTGTCGCTCTATCTTCN$_{12}$TTTTTTTTTTTT
Renilla luciferase F	GGAATTATAATGCTTATCTACGTGC
Renilla luciferase R	CTTGCGAAAAATGAAGACCTTTTAC
ACTB F	CTCTTCCAGCCTTCCTTCCT
ACTB R	AGCACTGTGTTGGCGTACAG

free NTPs. The amounts of each mRNA isoform were normalized prior to transfections.

RNA immunoprecipitation and RT-PCR

HEK293T cells grown on 10 cm plates were lysed as needed in three volumes of NP40 lysis buffer (50 mM HEPES-KOH pH = 7.5, 500 mM KCl, 2 mM EDTA, 1% Nonidet P-40 alternative, 0.5 mM DTT). Dynabeads were prepared with rabbit IgG (Cell Signaling 2729) and rabbit anti-EIF3B antibody (Bethyl A301- 761A). The lysate was split into three parts, the Dynabeadsantibody mixture was added, and the suspensions incubated for 2 h at 4°C. The beads were washed four times with NP40 buffer, and bound RNAs were isolated by phenol–chloroform extraction and ethanol precipitation. The resulting cDNA was reverse transcribed using random hexamers and Superscript III (Thermo Fisher scientific), and PCR was performed using DNA polymerase Q5 (NEB). qPCR was always performed in duplicates. Primers used to quantify PTBP1 RNA levels: PTBP1_Forward: GTACAAAGCGGGGATCTGAC PTBP1_Reverse: CGGCTGTCACCTTTGAACTT qPCR run conditions are as follows: 95°C for 15 sec, followed by 40 cycles at 95°C for 15 sec, 60°C for 60 sec, and 95°C for 1 sec.

Oxford nanopore sequencing

Nanopore sequencing was carried out using the manufacturer protocol for 1D Strand switching cDNA by ligation (SQKLSK108). The user defined primer was specific for exon 11 in PTBP1 mRNA: 5′ -ACTTGCCTGTCGCTCTATCTTCAGAGGCTTTGGGGTGTGA CT-3

Rapid Amplification of cDNA ends (RACE)

RACE analysis followed the protocol described for the FirstChoice RLM-RACE kit (Ambion), using the thermostable Vent DNA polymerase (NEB) and the adapter primers provided by the kit. The user-defined primers were: For the 5′ -UTR RACE: PTBP1-2 Exon junction reverse
(R2): 5′ -ATA CGG AGA GGC GAA GGC CG-3′ PTBP1-1 Exon junction reverse (R3): 5′ -GGA ACG GAA AGG CCG AAG GC-3′ PTBP1-4 Exon junction reverse (R1): 5′ -TAC CAG GTG CAC CGA AGG CC-3′ PTBP1 general reverse (R4): 5′ -AGA TCC CCG CTT TGT ACC AAC G-3′ For the 3′ -UTR RACE: Unique PTBP1-4 (F2): 5′ -GTGCACCTGGTATAATCTCAGCCT CTCC-3′ Unique PTBP1-2 (F3): 5′ -CGGCCTTCGCCTCTCCGTAT-3′ Unique PTBP1-1 (F4): 5′ -GCCTTCGGCCTTTCCGTTCC-3′ CDS F (F1): 5′ -AAGTCCACCATCTAGGGGCA-3

Table 2. PAR-CLIP crosslinking sites in hg38 coordinates.

EIF3 Subunit	Cluster Start	Cluster End	Replicate number	Number of reads
EIF3A	chr19 797,450	chr19 797,498	1	5
	-	-	2	0
	chr19 797,450	chr19 797,485	3	3
	chr19 797,379	chr19 797,404	3	1
EIF3B	chr19 797,444	chr19 797,485	1	27
	chr19 797,461	chr19 797,485	2	32
	chr19 797,423	chr19 797,435	2	1
	chr19 797,461	chr19 797,490	3	5
EIF3D	chr19 797,444	chr19 797,485	1	6
	-	-	2	0
	chr19 797,461	chr19 797,485	3	5
	chr19 797,379	chr19 797,404	3	9
EIF3G	chr19 797,450	chr19 797,485	1	9
	chr19 797,418	chr19 797,442	2	7
	chr19 797,418	chr19 797,441	3	13
	chr19 797,464	chr19 797,490	3	3

CHAPTER 3: Single Cell RNA-seq analysis of public datasets on 3'–UTR length during neuron development.

3.1 Abstract

The availability of computational data has increased logarithmically in the last decade. Multiple genome-wide studies acquire very similar data and mine the information needed for their specific hypothesis. However, the data contained in these large datasets are still very useful for answering other questions, but remain nearly unused. It takes knowledge and expertise to be able to extract the proper information in an unbiased way. This chapter focuses on the effort to use publicly available datasets of scRNA-seq to assess 3'–UTR sequence and length information in brain cells. mRNAs with long 3'–UTRs have been observed in mature brain cells and not anywhere else in the body. These lengthened 3'–UTRs might be related to mRNA transit, stability and processing due to microRNA binding sites but little is known about the essentiality of these to the genes that carry them or the function they exert. I analyzed the 3'–UTR content of eIF3 subunits in maturing brain cells. The mRNAs encoding three subunits, EIF3H, EIFM and EIF3F; have a switch in 3'–UTR length as cells mature, suggesting a role in the regulation of eIF3 during brain development. The rest of the mRNAs encoding eIF3 subunits do not vary too much regardless of the cell's development leading us to believe these would be stable in the main core of the complex. Although these observations require validation at the bench, this chapter shines a light into a possible specialized translational machinery working in a concerted way during brain development.

3.2 Introduction

Genetic information is passed from DNA to mRNA. This message will then be translated into a protein based on the cell's needs. For a long time, the scientific community believed that the transfer between DNA into protein depended solely on the coding sequence of a messenger RNA (mRNA). Counterintuitively, the number of genes encoding proteins is similar between species occupying different places in the evolutionary tree. Genomes of humans and simple eukaryotic organisms have roughly the same amount of protein coding genes (Mayr 2017). This pointed to protein sequence conservation and led researchers to think that sophistication in regulation might actually be taking place based on alternatively spliced isoforms and within the untranslated regions of mRNAs. The length of 3'–UTR sequences is directly related with the evolution of higher organisms and correlates well with their cellular complexity (Chen et al. 2010).

The 3'–UTR elements of mRNAs often contain *cis*-elements that can be recognized and bound by *trans*-acting factors and mediate mRNA expression and localization. The most common set of elements are alternative polyadenylation signals, micro-RNA/Argonaute binding sites and AU-rich elements. Different polyadenylation cleavage sites (PCS) depend on their polyadenylation signals (PAS) and *trans*-factors bound in their vicinity. The interplay between sequence context and polyadenylation signal location regulates the usage of PCSs, mediating either the lengthening or shortening of the 3'UTR, which leads to the formation of 3'UTR isoforms that contain different *cis*-regulatory elements (Kim et al. 2015). Alternative cleavage and polyadenylation (APA) enables cell-type and condition specific expression of 3'–UTR isoforms. The most dominant class of APA events happens within the 3' end of a transcript coding region, leading to 3'–UTR lengthening or shortening. The shortening of 3'–UTRs is a characteristic of proliferating cells and cancer cells (Sandberg et al. 2008), whereas 3'–UTR lengthening was observed during embryonic development and differentiation (Ji et al. 2009).

Sandberg and collaborators (Sandberg et al. 2008) showed that mammalian brains have a higher expression of longer 3'–UTRs and sequencing of 3' ends of polyadenylated transcripts uncovered several distal APA sites in cultured neurons in comparison to ES cells (Shepard et al. 2011). Studies in mice cerebellum (Pal et al. 2011) and *Drosophila* brains (Hilgers et al. 2012) revealed even more 3'–UTR extensions across hundreds of transcripts, pointing towards a conserved phenotype of 3'–UTR lengthening events in the nervous system.

According to NIH's National Human Genome Research Institute, with the substantial decrease in sequencing cost, the amount of genomic data exponentially increases with time (Figure 8); providing a large analytical space that has not yet been fully explored. Large datasets are often used for specific analysis and forgotten. It is important that computational analysis takes advantage of unused publicly available data to answer questions without the use of another "unnecessary" sequencing project.

Untranslated regions located on the 3' end of mRNAs have the advantage of location, with respect to sequencing technologies. Recently, several techniques (MACE-seq, Tag-Seq,

PAS-seq, MAPS, 10X (Boneva et al., 2020, Rosenberg et al., 2016, Ye et al., 2018) have been developed that aim to sequence the mRNA from the 3' end, which then guarantees that the untranslated region located there would have the best resolution in the final sequence space. With the purpose of studying lengthening events during neuronal development, I chose a set of publicly available datasets and analyzed these with the purpose of answering questions related to regulation of translation initiation factors.

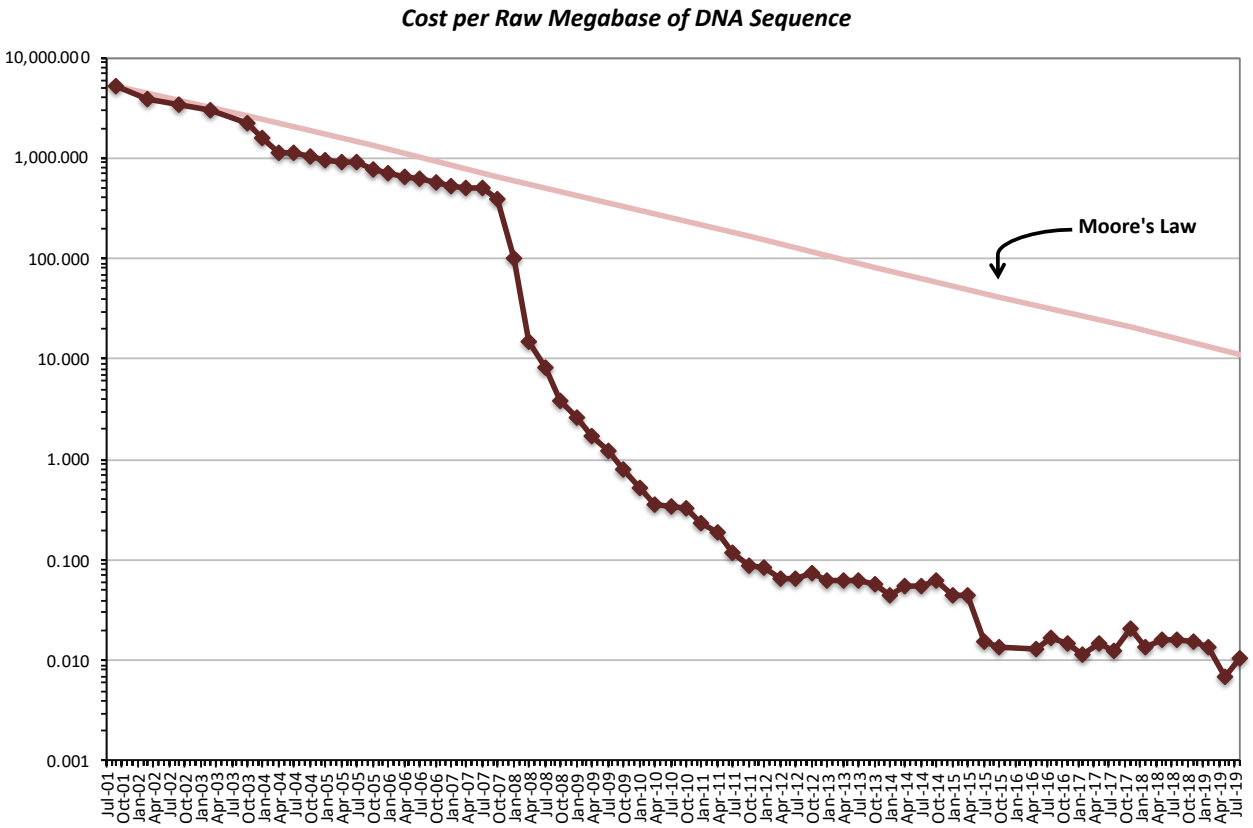

FIGURE 8: Cost/megabase sequencing from 2001 to 2019. Figure taken from NIH's National Human Genome Research Institute (https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data).

3.3 Datasets

Recently, the NIH through the Allen Institute for Brain Science (Jones 2010) have funded a large-scale effort to catalogue and analyze cells from mammalian brains. These data are shared and are a great resource to any researcher who wants to answer questions regarding the transcriptome. The datasets go through deposition policies and are entered into open repositories such as Gene Expression Omnibus and Sequence Read Archive (SRA) to drive further analyses by other groups and enable across-group comparisons (Keil et al. 2018). The drive for a more transparent transcriptomic space is leading groups to also share their raw data and deposit them into the above mentioned repositories. Slowly, user-friendly databases such as the one from the Allen Institute for Brain Science will incorporate the new datasets to allow a more complete analysis.

For the purpose of my study, I was interested in how the human brain changes the composition of 3' untranslated regions throughout development. It is well known that 3'–UTRs in mature brains are very long. Some of these long isoforms are not present anywhere else in the body (2019). It is still a standing question in the field if these extensions are tissue specific (different parts of the brain), or if they are gene-specific. The other hypothesis is that these extensions are coordinating brain development at the translational level by slowing expression of a subset of genes and allowing the brain to mature (Wehrspaun et al. 2014; Hilgers et al., 2011; Tushev et al. 2018). To be able to de-convolute these possibilities, the best way to look at this question was using a dataset obtained from single cells (scRNA-seq) (Figure 9). By pooling RNA-seq data by cell-type, I could potentially analyze the connection between 3'–UTR extensions, the location of these cells in the brain and the connection with the developmental stage of the cells. The transcription profiling data I analyzed is from a study published by the Alex Pollen group in 2014 (Pollen et al. 2014). In this paper, the authors carried out a large scale survey of single cell expression to determine the sequencing depth requirements to successfully determine lineage relationships. This dataset was especially interesting for my study, because they assayed 301 cells from 11 different populations and performed cell-type classification and biomarker identification. These cells are derived from multiple progenitor and neuronal subtypes which would allow for my investigations.

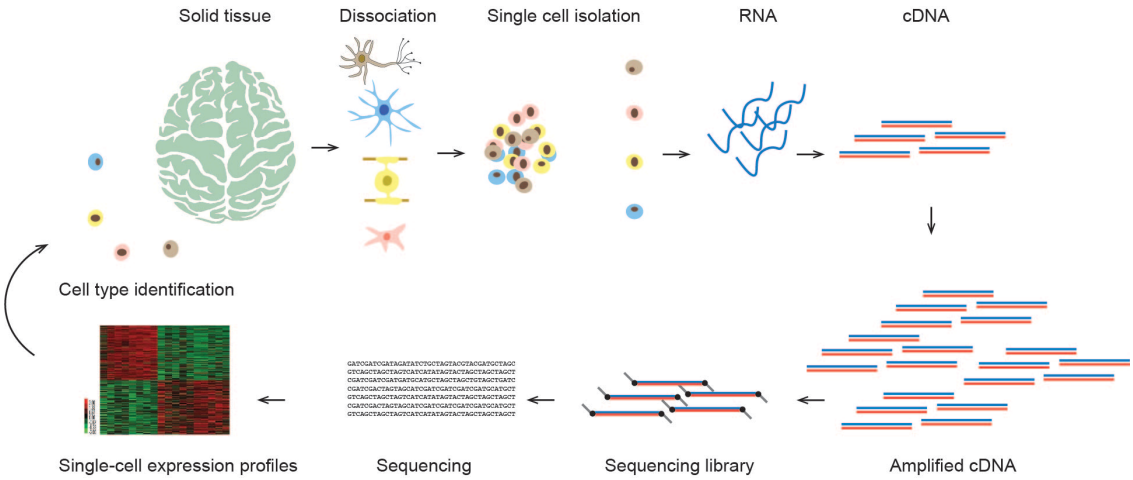

FIGURE 9. Single cell sequencing workflow, modified from Wikipedia; https://en.wikipedia.org/wiki/File:RNA-Seq_workflow-5.pdf

3.4 Pre-processing steps

Downloading raw data

So far, there is no written protocol of how to submit raw RNA-seq data into databases. For this reason, it is sometimes challenging to find the complete data package needed for analysis. In the case of the dataset from Pollen, 2014; I was able to download it from NCBI. The Run Selector is a friendly user platform from SRA-NCBI that can be searched by study. The Pollen study number is provided in the paper: SRP041736

https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP041736&ff=on

While in the Run Selector, I selected the tissue on the left side of the screen. I selected "brain." This made the total size of the data decrease from 191.61Gb to 54.67Gb. The website organizes the results into a table with the categories number, acc, AvgSpotLen, Bases, Biomaterial_provider, BioSample, Bytes, Experiment, Library Name, MBases, Mbytes, population, sample_acc, Sample_Name, Cell_Line, Age, tissue, disease, sex, disease_stage and health_state.

Since this is a single-cell RNA-seq experiment, each cell is a different SRA file. To be able to download all brain files I was interested in, I made a list with the accession numbers of SRAs pertaining to brain tissue. While in the spotlight search of a MacBook (Command + space), type terminal and hit enter. Using the terminal window:

#cat > SRR_list

And paste all of the numbers of the desired SRAs, in this case, the brain tissue SRAs.

#for i in $(cat SRR_list);do wget -r -nd -nH ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR127/$i/*; done

This will give the address and file names as a list to to download the SRA files.

Converting the files into fastq format

To be able the use the SRA files, they must first be converted to fast format. This can be done using the SRA toolkit. To speed the process, I used the UC Berkeley Savio Server using a simple batch script to convert all files from my list into fastq files:

1. Copy the .sh script into a template file

```
#!/bin/sh
############################## BATCH SCRIPT
###################################
### FILENAME:fastqdump.sh
###
### PURPOSE:loop to convert SRA into fastq
###
###
### USAGE:
###
### AUTHOR: Luisa Tacca
### DATE: Sept 2018
###
###
#################################################################################
#
#SBATCH --job-name=fastqdump
#SBATCH --partition=savio2
#SBATCH --account=fc_ribosome
#SBATCH --cpus-per-task=1
#SBATCH --time=10:00:00
cd /global/scratch/latacca/Pollen/sra
## RUN!!
fastq-dump --skip-technical --readids --dumpbase --split-files YYYZ
```

2. Then in the terminal window:

```
#$ for i in *sra; do echo $i; done
#$ for i in *sra; do sed -e "s/YYYZ/$i/g" fastqc_template.sh > fastqc_${i}.sh; done
#$for i in *sra.sh; do sbatch $i; done
```

This process generates two fastq files for each SRA entry, since the sequencing data is paired-end reads.

Another option of downloading data

Github (*https://github.com/*) and stack overflow (*https://stackoverflow.com/)* are great resources for learning and answering computational questions one might have. During my research I discovered a tool developed by Phil Ewels, called SRA Explorer (*https://github.com/ewels/sra-explorer)*. This tool (*https://sra-explorer.info/*) allows one to look for the desired data by inputting the SRA number as noted above, in my case SRP041736 for the Pollen study.

After this one can select the desired runs by accession number and add to the collection of saved datasets. While in the tab of saved datasets, one can choose to download the fastq file instead of the SRA. One can also click the option of batch script download and the tool will write a script to use in the terminal and download the desired data.

The following is an example script:

```
#!/bin/bash
mkdir data
cd data
for i in `seq 25 40`;
do
  mkdir DRR0161${i};
  cd DRR0161${i};
  wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/DRR016/DRR0161${i}/
DRR0161${i}_1.fastq.gz;
  wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/DRR016/DRR0161${i}/
DRR0161${i}_2.fastq.gz;
  cd ..;
done
cd ..
```

Read quality check

After downloading the data and converting it to fastq format, it is essential to first check the quality of the reads (Sheng et al., 2016). For this purpose, I used a tool called FastQC (Babraham Institut). This is a quality control tool for sequencing data that is compatible with either bulk RNA-seq or single-cell RNA-seq. One of the challenges of single-cell RNA-seq is the large amount of data that must be processed. In my case, I used more than 300 cells;,which means that I had more than 300 FastQC reports to analyze. I therefore used a tool called MultiQC (Ewels et al., 2016) which will search a working folder for all of the files with

extension .html that were generated by fastQC and then transform all reports into one summary report.

Separating Accessions into folders

To calculate the usage of 3'–UTR length, I used a tool developed by the Morris lab (Ha et al., 2018). In this paper, the authors describe 'Quantification of APA' (qAPA), a method that calculates Alternative Polyadenylation (APA) from conventional RNA-seq data. In order to use this method, each RNA-seq experiment (In the case of this study, each cell), has to be in its own folder. Each folder has two read files, which are the paired reads for each accession.

```
#!/bin/bash
for file in *.fastq.gz; do
  if [[ -f "$file" ]]; then
    mkdir "${file%_Cell_Diversity_Hiseq_RNA-Seq_*.fastq.gz}"
    mv "$file" "${file%_Cell_Diversity_Hiseq_RNA-
Seq_*.fastq.gz}"
  fi
done
```

Modified Salmon for 3' - UTR read quantification

To quantify 3'–UTR usage, I only needed to use the 3'–UTR index and not the whole genome index.

```
#!/bin/bash
for fn in 12/SRR29677{34..43};
do
samp=`basename ${fn}fi`
echo "Processing sample ${samp}"
salmon quant -i utr_library -l A \
         -1 ${fn}/${samp}_Homo_sapiens_RNA-Seq_1.fastq.gz \
         -2 ${fn}/${samp}_Homo_sapiens_RNA-Seq_2.fastq.gz \
         -p 8 --validateMappings -o quants/${samp}_quant
done
```

The program Salmon (Patro et al., 2017) creates a folder 'quants' that has all quantification files. This file is named 'SRR1275*_ quant' and qAPA does not recognize this name pattern. For this reason, I changed the name of the files to remove the _quant.

```
#rename "s/_quant//" *
```

Now, the qAPA command can be run:

```
qapa quant --db ensembl.identifiers.txt project/sample*/quant.sf
> pau_results_total.txt
```

qAPA quantifies APA levels using RNA-seq reads that uniquely map to 3'–UTR sequences demarcated by annotated poly(A) sites in last exons (Ha et al., 2018). As a workflow for qAPA, I used the protocol developed by Ha and collaborators:

1. 3'–UTR extraction from GENCODE
2. Incorporate poly(A) site annotations from other databases
3. Update 3' ends based on supplemental databases

This creates the qAPA 3'–UTR reference library that will be used to calculate the poly(A) usage.

After the library is made, Salmon quantifies the levels of the transcripts in the RNA-seq data. Then these levels are mapped to the library by the qAPA tool to output the usage. Poly(A)

usage (PAU) for a given 3'–UTR is the ratio of its expression to the sum of the expression of all detected 3'–UTR isoforms from its gene (Ha et al., 2018).

3.5 Results and Discussion

Based on prior work (Wang et al., 2014; Kuklin et al, 2017 and Gruner et al., 2019) I first looked at the differences in the specific case of *Calmodulin 1 (CALM1)* mRNA. *CALM1* is part of the calcium signaling pathway in the brain and it is important for normal neural development. Through alternative poly-adenylation, *CALM1* exists in short and long isoforms. The long isoforms is largely restricted to neuronal tissues (Gruner et al., 2019). Thus this mRNA seemed like a good test case to explore the data and start looking into long 3' - UTR forms in brain cells (Figure 10).

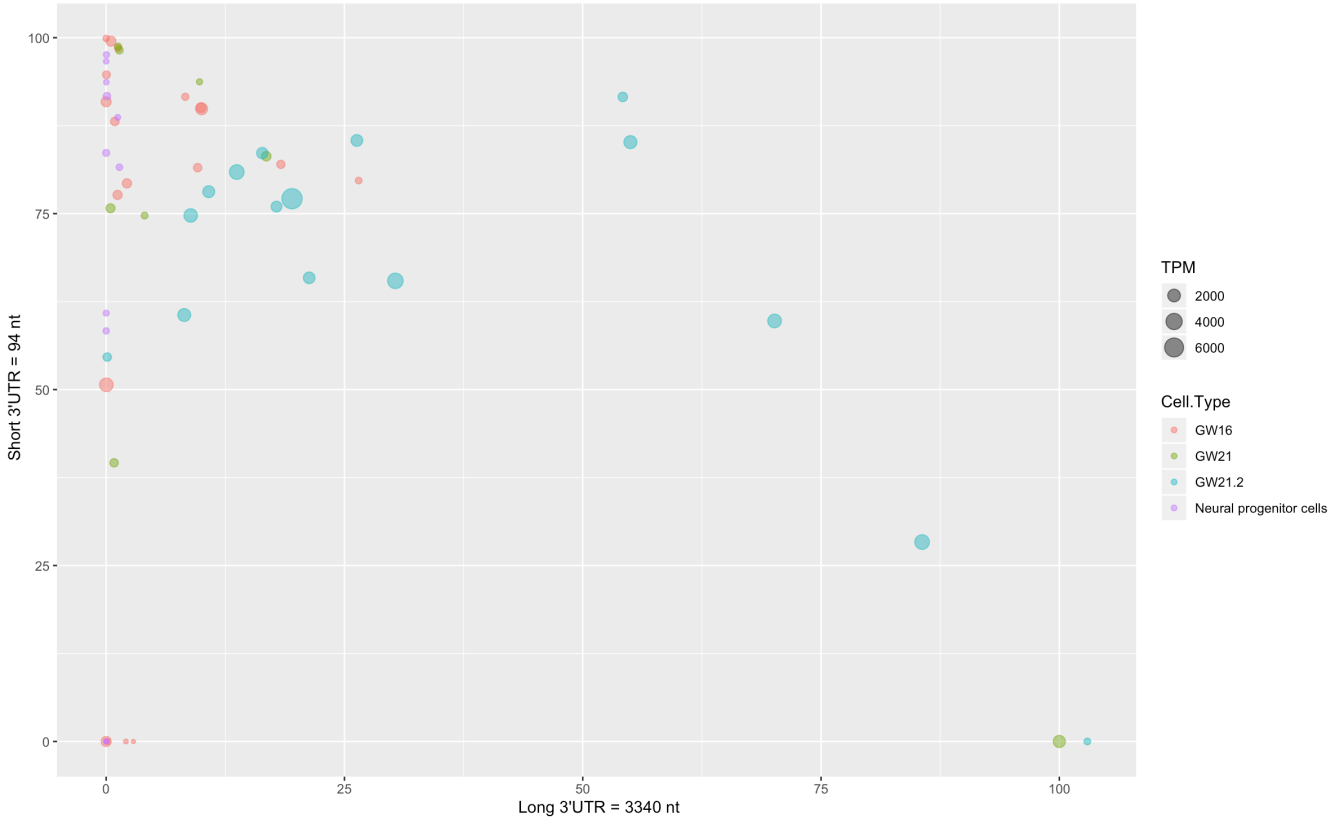

FIGURE 10. *CALM1* expression in the human brain. Data are from Pollen et. al, 2014. Maximum and minimum length of 3' UTR were plotted according to cell type. GW refers to gestational age and the numbers refer to weeks. Size of the data points refer to their transcript expression measured in transcripts/million.

As shown by Hilgers and collaborators (Hilgers et al., 2011) and others, *CALM1* in Pollen et al, 2014 dataset follows the trend of having shorter 3'–UTR isoforms expressed in younger cells. Neural progenitor cells (NPC) have a higher expression of *CALM1* with a short 3' - UTR and older gestational age neurons, which are therefore more mature, have a higher presence of longer form of 3'–UTR isoforms.

After demonstrating that *CALM1* 3'–UTR length changed with neuronal development in the Pollen data, as previously described, I was interested in exploring the differences in 3'–UTR content in mRNAs encoding translation factors and proteins involved with translational control. There is little information in the literature regarding the state of isoform specific translation of genes deeply connected to the process of translation in the brain.

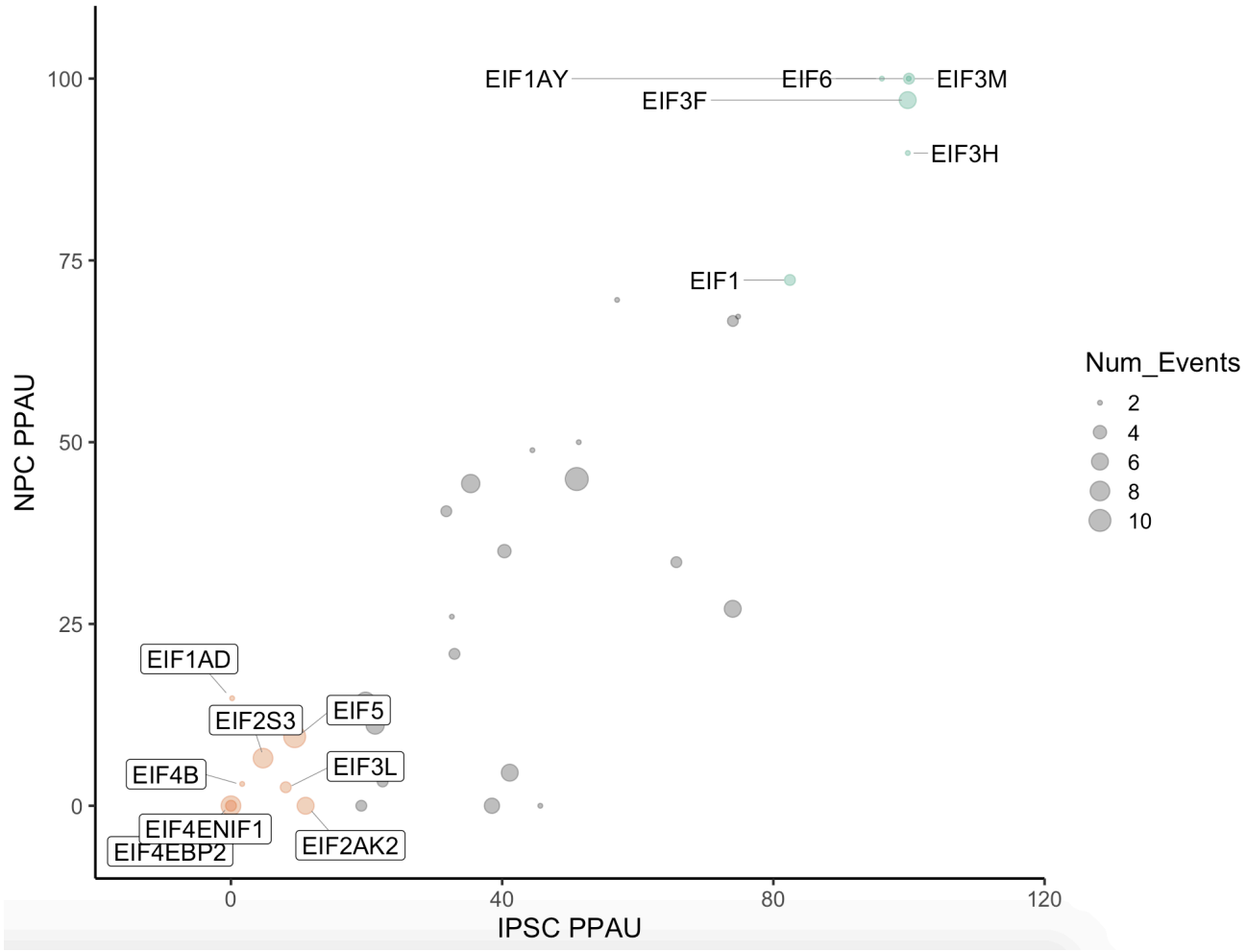

FIGURE 11. Usage of short 3'–UTR isoform of translation factors in neural progenitor cells and in IPSc cells. PPAU refers to Proximal poly-adenylation usage and is measured in percent. Num_Events refers to number of poly-adenylation events for each transcript.

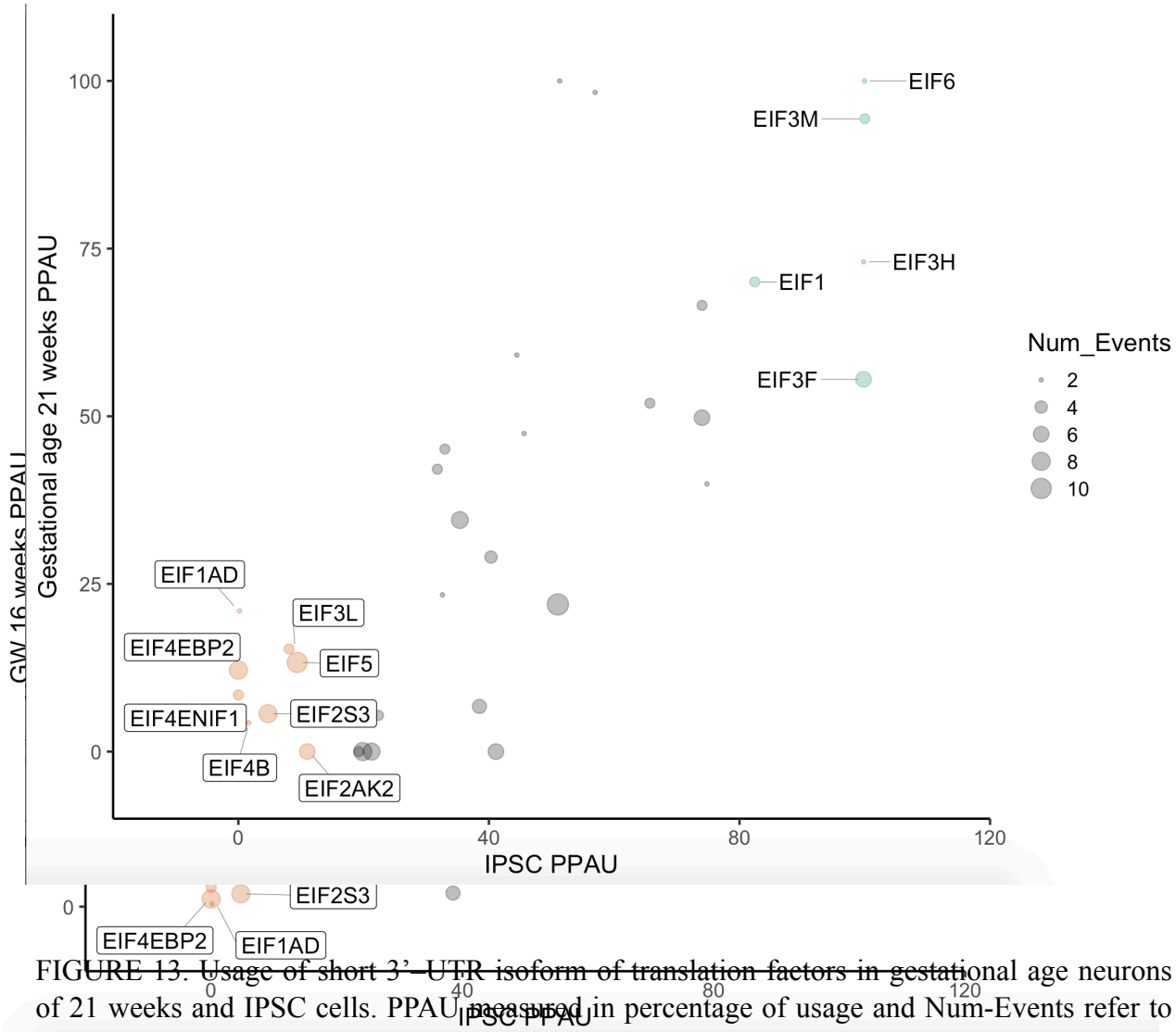

FIGURE 13. Usage of short 3'–UTR isoform of translation factors in gestational age neurons of 21 weeks and IPSC cells. PPAU measured in percentage of usage and Num-Events refer to amount of poly-adenylation events for each transcript.

FIGURE 12. Usage of short 3'–UTR isoform of translation factors in gestational age neurons at 16 weeks and IPSs cells. PPAU measured in percentage of usage and Num-Events refer to number of poly-adenylation events.

Using the data from Pollen, 2014, I determined the percentage of usage of the short 3'–UTR isoform of a list of translation factors. A high percentage in the graph means that the mRNA in question has a low usage of long 3'–UTRs and a low percentage means that the mRNA has a high usage of the long 3'–UTR isoform. Using this approach, I find that mRNAs encoding subunits EIF3M, F and H have predominantly a short 3'–UTR early in cell differentiation (Figures 11-14). However, in the case of *EIF3F* and *EIF3H* mRNAs, the usage of the long 3'–UTR isoforms increases as the cell moves into cells from fetuses in week 16 of gestation (radial glia).

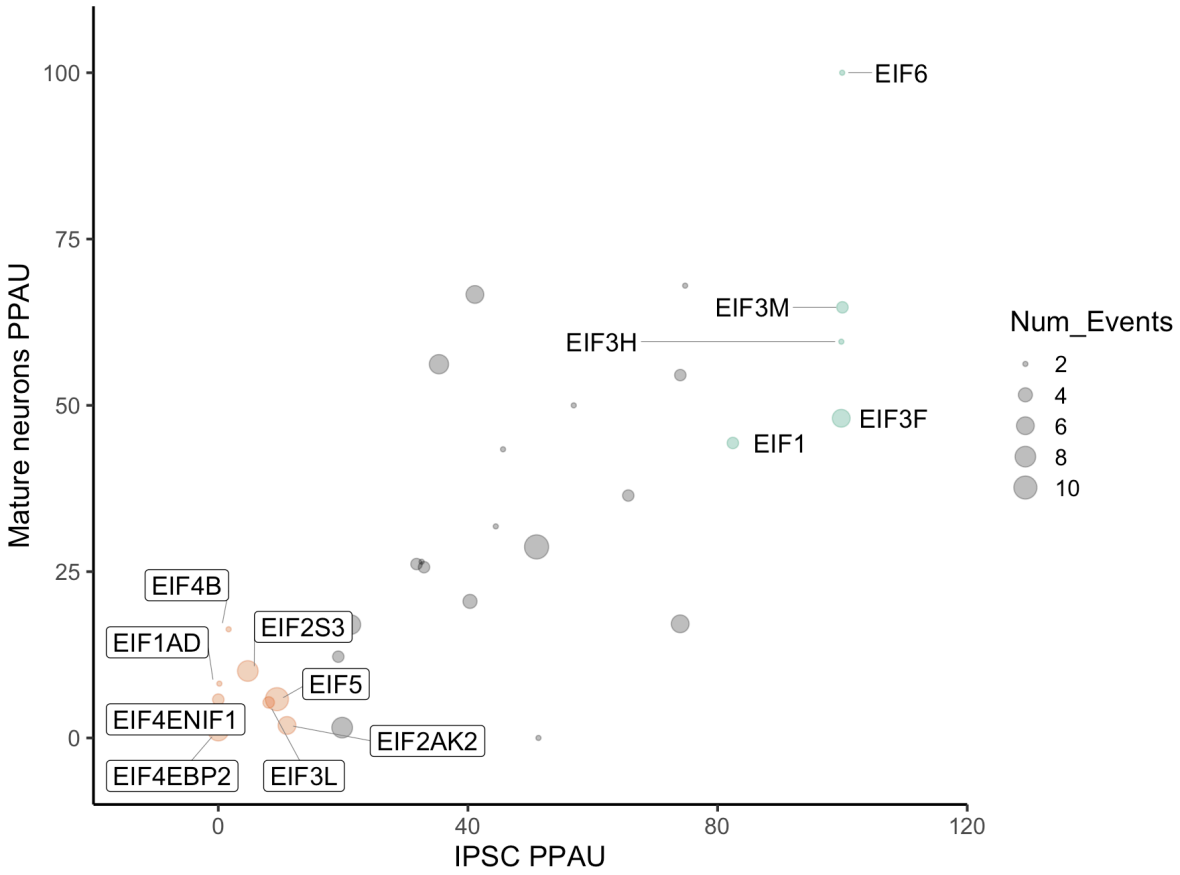

FIGURE 14. Usage of short 3'–UTR isoform of translation factors in mature neurons and IPSC cells. PPAU measured in percentage of usage.

As cells progress into 21 weeks of gestation, mRNAs for eIF3 subunits H, M and F maintain the proximal usage pattern (i.e. have short 3'–UTRs), with *EIF3M* still mostly using the shorter 3'–UTR iso
. *EIF3F* has a slight increase in longer 3'–UTR usage, as its proximal usage goes down to 50%. *EIF3H* maintains 75% proximal 3'–UTR usage (Figure 13).

As cells mature further into gestational week 21+2 days (Figure 14) and become mature neurons, these eIF3 subunits continue to exhibit changes in 3'–UTR isoform usage. The long 3'–UTR isoform of *EIF3F* goes up to 58% usage,while that for *EIF3H* remains at around 40%. For *EIF3M,* which during gestational week 16 was mostly expressed with the short 3'-UTR form, is now expressed with 32% usage of the long 3'–UTR isoform.

Once neurons become mature, *EIF3M* shows a drastic change in 3'–UTR usage. It joins *EIF3H* and *EIF3F* with around 50% usage of the proximal poly-adenylation site (Figure 14). Interestingly, subunits EIF3H:M:F are a sub-complex within the larger 13-subunit eIF3 initiation factor. Through gas phase mass spectrometry, Zhou and collaborators (Zhou et al., 2008) observed that eIF3 dissociation occurs as a function of ionic strength to form three stable modules EIF3(c:d:e:l:k); eIF3(f:h:m), and eIF3(a:b:i:g). It is possible that we are seeing the EIF3H:M:F module being regulated through differential usage of 3'–UTR isoforms to ensure correct assembly of the larger eIF3 complex. Future experiments will be needed to validate these transcript-level observations. Gruner and collaborators showed that when *CALM1* long 3'UTR was removed with the use of Cas9, it resulted in impairment of mice root ganglion development. Following the same principle, it would be interesting to remove the extensions on the 3' - UTR of EIF3 H:M:F and assess differences in brain function and EIF3 complex formation.

The observation of changes in 3'–UTR usage in mRNAs encoding subunits of eIF3 resulted from the analysis of previously available data. This highlights the amount of information buried in published data, and the need for continued computational analysis to extract such information. An important part of computational analysis is the validation of the findings. This project opens doors for further testing the importance of 3'–UTR extensions of eIF3 components and the impact that might have in translation regulation and proper assembly of the complex. However, one needs to be careful with the amount of information contained in 3'–UTR extensions as it is a hub for transcript regulation in other already curated mechanisms such as RNA stability and degradation and microRNA binding. Once again, computational analysis would be an excellent aid to de-convoluting different mechanisms, highlighting the importance of the two fields of computational and experimental biology complementing each other.

REFERENCES

Abanades DR, Ramírez L, Iborra S, Soteriadou K, González VM, Bonay P, Alonso C, Soto M. 2009. Key role of the 3′ untranslated region in the cell cycle regulated expression of the Leishmania infantum histone H2A genes: minor synergistic effect of the 5′ untranslated region. BMC Mol Biol 10: 48. doi:10.1186/1471-2199-10-48

Bert AG. 2006. Assessing IRES activity in the HIF-1 and other cellular 5′ UTRs. RNA 12: 1074–1083. doi:10.1261/rna.2320506

Brenner SE. 1999. Errors in genome annotation. Trends Genet 15:132–133. doi:10.1016/S0168-9525(99)01706-0

Coldwell MJ, Cowan JL, Vlasak M, Mead A, Willett M, Perry LS,

Morley SJ. 2013. Phosphorylation of eIF4GII and 4E-BP1 in response to nocodazole treatment: a reappraisal of translation initiation during mitosis. Cell Cycle 12: 3615–3628. doi:10.4161/cc.26588

Chen C-Y, Chen S-T, Juan H-F, Huang H-C. 2010. Lengthening of 3'UTR Increases Morphological Complexity in Animal Evolution. Nature Precedings. http://dx.doi.org/10.1038/npre.2010.4915.1.

Cooper S, Iyer G, Tarquini M, Bissett P. 2006. Nocodazole does not synchronize cells: implications for cell-cycle control and whole-culture synchronization. Cell Tissue Res 324: 237–242. doi:10.1007/s00441-005-0118-8

Dong Z, Liu Z, Cui P, Pincheira R, Yang Y, Liu J, Zhang J-T. 2009. Role of eIF3a in regulating cell cycle progression. Exp Cell Res 315:1889–1894. doi:10.1016/j.yexcr.2009.03.009

Fan H, Penman S. 1970. Regulation of protein synthesis in mammalian cells. II. Inhibition of protein synthesis at the level of initiation during mitosis. J Mol Biol 50: 655–670. doi:10.1016/0022-2836(70)90091-4

The FANTOMConsortium and the RIKEN PMI and CLST (DGT). 2014. A promoter-level mammalian expression atlas. Nature 507: 462–470. doi:10.1038/nature13182

Garcia-Blanco MA, Jamison SF, Sharp PA. 1989. Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. Genes Dev 3: 1874–1886. doi:10.1101/gad.3.12a.1874

Gil A, Sharp PA, Jamison SF, Garcia-Blanco MA. 1991. Characterization of cDNAs encoding the polypyrimidine tract-binding protein. Genes Dev 5: 1224–1236. doi:10.1101/gad.5.7.1224

Gueroussov S, Gonatopoulos-Pournatzis T, Irimia M, Raj B, Lin Z-Y,

Gingras A-C, Blencowe BJ. 2015. An alternative splicing event amplifies evolutionary differences between vertebrates. Science 349:868–873. doi:10.1126/science.aaa8381

Gruner HN, Bae B, Lynch M, Oliver D, So K, Mastick GS, Yan W, Miura P. Precise removal of Calm1 long 3′ UTR isoform by CRISPR-Cas9 genome editing impairs dorsal root ganglion development in mice. http://dx.doi.org/10.1101/553990.

Higareda-Mendoza AE, Pardo-Galván MA. 2010. Expression of human eukaryotic initiation factor 3f oscillates with cell cycle in A549 cells and is essential for cell viability. Cell Div 5: 10. doi:10.1186/1747-1028-5-10

Hilgers V, Lemke SB, Levine M. 2012. ELAV mediates 3' - UTR extension in the Drosophila nervous system. Genes Dev 26: 2259–2264.

Hinnebusch AG, Ivanov IP, Sonenberg N. 2016. Translational control by 5′-untranslated regions of eukaryotic mRNAs. Science 352:1413–1416. doi:10.1126/science.aad9868

Ji Z, Lee JY, Pan Z, Jiang B, Tian B. 2009. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci U S A 106: 7028–7033.

Jones AR. 2010. Mapping gene expression in the CNS: Tools and data from the Allen Institute for Brain Science. Neuroscience Research 68: e4. http://dx.doi.org/10.1016/j.neures.2010.07.244.

Kamath RV, Leary DJ, Huang S. 2001. Nucleocytoplasmic shuttling of polypyrimidine tract-binding protein is uncoupled from RNA export. Mol Biol Cell 12: 3808–3820. doi:10.1091/mbc.12.12.3808

Keil JM, Qalieh A, Kwan KY. 2018. Brain Transcriptome Databases: A User's Guide. The Journal of Neuroscience 38: 2399–2412. http://dx.doi.org/10.1523/jneurosci.1930-17.2018.

Kim HH, Lee SJ, Gardiner AS, Perrone-Bizzozero NI, Yoo S. 2015. Different motif requirements for the localization zipcode element of β-actin mRNA binding by HuD and ZBP1. Nucleic Acids Res 43: 7432–7446.

Kirschman JL, Bhosle S, Vanover D, Blanchard EL, Loomis KH, Zurla C,

Murray K, Lam BC, Santangelo PJ. 2017. Characterizing exogenous mRNA delivery, trafficking, cytoplasmic release and RNAprotein correlations at the level of single cells. Nucleic Acids Res 45: e113. doi:10.1093/nar/gkx290

Lee ASY, Kranzusch PJ, Cate JHD. 2015. eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. Nature 522: 111–114. doi:10.1038/nature14267

Li B, Yen TSB. 2002. Characterization of the nuclear export signal of polypyrimidine tract-binding protein. J Biol Chem 277: 10306–10314. doi:10.1074/jbc.M109686200

Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, Älgenäs C, Lundeberg J, Mann M, Uhlen M. 2010. Defining the transcriptome and proteome in three functionally different human cell lines. Mol Syst Biol 6: 450. doi:10.1038/msb.2010.106

Ma W, Mayr C. 2018. A membraneless organelle associated with the endoplasmic reticulum enables 3′UTR-mediated protein-protein interactions. Cell 175: 1492–1506. doi:10.1016/j.cell.2018.10.007

Marash L, Liberman N, Henis-Korenblit S, Sivan G, Reem E, Elroy-Stein O, Kimchi A. 2008. DAP5 promotes cap-independent translation of Bcl-2 and CDK1 to facilitate cell survival during mitosis. Mol Cell 30: 447–459. doi:10.1016/j.molcel.2008.03.018

Mayr C. 2017. Regulation by 3′-untranslated regions. Annu Rev Genet 51: 171–194. doi:10.1146/annurev-genet-120116-024704

Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. 2013. Widespread and extensive lengthening of 3' - UTRs in the mammalian brain. Genome Res 23: 812–825.

Monzón-Casanova E, Screen M, Díaz-Muñoz MD, Coulson RMR, Bell SE, Lamers G, Solimena M, Smith CWJ, Turner M. 2018. The RNA-binding protein PTBP1 is necessary for B cell selection in germinal centers. Nat Immunol 19: 267–278. doi:10.1038/s41590-017-0035-5

Noguchi S, Arakawa T, Fukuda S, Furuno M, Hasegawa A, Hori F, Ishikawa-Kato S, Kaida K, Kaiho A, Kanamori-Katayama M, et al. 2017. FANTOM5 CAGE profiles of human and mouse samples.Sci Data 4: 170112. doi:10.1038/sdata.2017.112

O'Leary NA, Wright MW, Rodney Brister J, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako Adjei D, et al. 2015. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44: D733–D745. doi:10.1093/nar/gkv1189

Pal S, Gupta R, Kim H, Wickramasinghe P, Baubet V, Showe LC, Dahmane N, Davuluri RV. 2011. Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. Genome Res 21: 1260–1272.

Park J-E, Yi H, Kim Y, Chang H, Kim VN. 2016. Regulation of poly(A)tail and translation during the somatic cell cycle. Mol Cell 62:462–471. doi:10.1016/j.molcel.2016.04.007

Pelechano V, Wei W, Steinmetz LM. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. Nature 497: 127–131. doi:10.1038/nature12121

Pérez I, McAfee JG, Patton JG. 1997. Multiple RRMs contribute toRNA binding specificity and affinity for polypyrimidine tract binding protein. Biochemistry 36: 11881–11890. doi:10.1021/bi9711745

Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S. 2001. Structural and functional features of eukaryoticmRNAuntranslated regions. Gene 276: 73–81. doi:10.1016/S0378-1119(01)00674-6

Promponas VJ, Iliopoulos I, Ouzounis CA. 2015. Annotation inconsistencies beyond sequence similarity-based function prediction phylogeny and genome structure. Stand Genomic Sci 10: 108.doi:10.1186/s40793-015-0101-2

Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, Li N, Szpankowski L, Fowler B, Chen P, et al. 2014. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol 32: 1053–1058.

Ramírez-Valle F, Badura ML, Braunstein S, Narasimhan M, Schneider RJ. 2010. Mitotic raptor promotes mTORC1 activity, G2/M cell cycle progression, and internal ribosome entry site-mediated mRNA translation. Mol Cell Biol 30: 3151–3164. doi:10.1128/ MCB.00322-09

Romanelli MG, Diani E, Lievens PM-J. 2013. New insights into functional roles of the polypyrimidine tract-binding protein. Int J Mol Sci 14: 22906–22932. doi:10.3390/ijms141122906

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science 320: 1643–1647.

Sawicka K, Bushell M, Spriggs KA, Willis AE. 2008. Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. Biochem Soc Trans 36: 641–647. doi:10.1042/BST0360641

Schipany K, Rosner M, Ionce L, Hengstschläger M, Kovacic B. 2015. eIF3 controls cell size independently of S6K1-activity. Oncotarget 6: 24361–24375. doi:10.18632/oncotarget.4458

Schnoes AM, Brown SD, Dodevski I, Babbitt PC. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5: e1000605. doi:10.1371/journal.pcbi.1000605

Shepard PJ, -A. Choi E, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA 17: 761–772. http://dx.doi.org/10.1261/rna.2581711.

Shibayama M, Ohno S, Osaka T, Sakamoto R, Tokunaga A, Nakatake Y, Sato M, Yoshida N. 2009. Polypyrimidine tract-binding protein is essential for early mouse development and embryonic stem cell proliferation. FEBS J 276: 6658–6668. doi:10
.1111/j.1742-4658.2009.07380.

Sonenberg N. 1994. mRNA translation: influence of the 5′ and 3′ untranslatedregions. Curr Opin Genet Dev 4: 310–315. doi:10.1016/S0959-437X(05)80059-0

Stumpf CR, Moreno MV, Olshen AB, Taylor BS, Ruggero D. 2013. The translational landscape of the mammalian cell cycle. Mol Cell 52: 574–582. doi:10.1016/j.molcel.2013.09.018

Szostak E, Gebauer F. 2013. Translational control by 3′-UTR-binding proteins. BriefFunctGenomics 12: 58–65. doi:10.1093/bfgp/els056 Tanenbaum ME, Stern-Ginossar N, Weissman JS, Vale RD. 2015.

Regulation of mRNA translation during mitosis. Elife 4: e07957.doi:10.7554/elife.07957

Tushev G, Glock C, Heumüller M, Biever A, Jovanovic M, Schuman EM. 2018. Alternative 3' - UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. Neuron 98: 495–511.e6.

Valcárcel J, Gebauer F. 1997. Post-transcriptional regulation: the dawn of PTB. Curr Biol 7: R705–R708. doi:10.1016/S0960-9822 (06)00361-7

Vassilev LT, Tovar C, Chen S, Knezevic D, Zhao X, Sun H, Heimbrook DC, Chen L. 2006. Selective small-molecule inhibitor reveals critical mitotic functions of human CDK1. Proc Natl Acad Sci 103: 10660–10665. doi:10.1073/pnas.0600447103

Wang H, Ru Y, Sanchez-CarbayoM, Wang X, Kieft JS, Theodorescu D. 2013. Translation initiation factor eIF3b expression in human cancer and its role in tumor growth and lung colonization. Clin Cancer Res 19: 2850–2860. doi:10.1158/1078-0432.CCR-12-3084

Wang Q, He G, Hou M, Chen L, Chen S, Xu A, Fu Y. 2018. Cell cycle regulation by alternative polyadenylation of CCND1. Sci Rep 8:6824. doi:10.1038/s41598-018-25141-0

Wehrspaun CC, Ponting CP, Marques AC. 2014. Brain-expressed 3'UTR extensions strengthen miRNA cross-talk between ion channel/transporter encoding mRNAs. Front Genet 5: 41.

Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed [01/13/2020].

Wilker EW, van Vugt MA, Artim SA, Huang PH, Petersen CP, Reinhardt HC, Feng Y, Sharp PA, Sonenberg N, White FM, et al.2007. 14-3-3σ controls mitotic translation to facilitate cytokinesis. Nature 446: 329–332. doi:10.1038/nature05584

Wilkie GS, Dickson KS, Gray NK. 2003. Regulation of mRNA translation by 5′- and 3′-UTR-binding factors. Trends Biochem Sci 28: 182–188. doi:10.1016/S0968-0004(03)00051-3

Wollerton MC, Gooding C, Robinson F, Brown EC, Jackson RJ, Smith CW. 2001. Differential alternative splicing activity of isoforms of polypyrimidine tract binding protein (PTB). RNA 7: 819–832. doi:10.1017/S1355838201010214

Wollerton MC, Gooding C, Wagner EJ, Garcia-Blanco MA, Smith CWJ. 2004. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. Mol Cell 13: 91–100. doi:10.1016/S1097-2765(03)00502-1

Workman RE, Tang A, Tang PS, Jain M, Tyson JR, Zuzarte PC, Gilpatrick T, Razaghi R, Quick J, Sadowski N, et al. 2018. Nanopore native RNA sequencing of a human poly(A) transcriptome. bioRxiv doi:10.1101/459529

Xue Y, Zhou Y, Wu T, Zhu T, Ji X, Kwon Y-S, Zhang C, Yeo G, Black DL,Sun H, et al. 2009. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. Mol Cell 36: 996–1006. doi:10 .1016/j.molcel.2009.12.003

Yamashita A, Takeuchi O. 2017. Translational control of mRNAs by 3′- untranslated region binding proteins. BMB Rep 50: 194–200. doi:10.5483/BMBRep.2017.50.4.040

You L, Wu J, Feng Y, Fu Y, Guo Y, Long L, Zhang H, Luan Y, Tian P, Chen L, et al. 2015. APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. Nucleic Acids Res 43: D59–D67. doi:10.1093/nar/gku1076

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. Nucleic Acids Res 46: D754–D761. doi:10.1093/nar/gkx1098

Zhao S, Zhang B. 2015. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics 16: 97. doi:10.1186/s12864-015-1308-8