Fakultät für Informatik
Technische Universität München

Deciphering the genetic code of post-transcriptional gene expression regulation via multi-omics data integration

Başak Eraslan

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Burkhard Rost

Prüfende der Dissertation:
1. Prof. Dr. Julien Gagneur
2. Prof. Dr. Mathias Wilhelm

Die Dissertation wurde am 25.05.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 06.08.2021 angenommen.

Acknowledgments

First of all, I would like to thank all members of the Quantitative Biosciences Munich Graduate School for giving me the opportunity to become a member of the multi-cultural and inter-disciplinary academic environment they have created. Their generous support and guidance enabled me to develop myself to become a researcher.

I also want to thank my lab friends; Juri, Vicente, Veronika, Elaine, Daniel, Leo, Jun, Daniela, Ziga, Chris, Felix, Agne and Mathilde for the joyful moments and semi-scientific discussions we shared. My special thanks go to Inga Weise who helped me on various official issues.

I am thankful to our collaborators; Dongxue Wang, Hannes Hahne and Prof. Bernhard Küster for their exceptional cooperation and understanding throughout our study. I especially thank Mirjana Gusic not only for putting lots of effort to improve the quality of our work, but also for her honest friendship. Finally, thanks to Prof. Julien Gagneur for editing the manuscript.

My lifelong companion Gökçen, my family, my little gazelle Ayça and my ever best friend Pınar; I am very lucky to have you and always happy to be with you. Thank you very much for your love, endless support and encouragement.

Summary

The advances in high throughput 'omics' technologies have paved the way for understanding the inner workings of the cell on an unprecedented scale. Upon the technological improvements, several experimental techniques have been developed to identify, monitor and quantify various processes in the cell. We are now able to learn the code of "life" in a totally unbiased, data-driven way.

Even though all the cells in an organism have the same DNA, there is a huge diversity of cell types and states due to the complex mechanisms of the gene expression regulation. This regulation is achieved through multiple steps, starting from transciption, followed by splicing, translation and protein degradation. Therefore, deciphering the regulatory mechanisms within and across each of these steps is critical for understanding how the cells work.

Even though, transcriptional regulation has long been assumed to be the major regulatory step among all, we are now realizing the importance of the post-transcriptional events and the diseases that occur in the case of their miscoordination. New experimental techniques such as ribosome profiling, e-CLIP and mass-spec proteomics shed lights on identification and quantification of the regulatory events happening during translation and protein degradation.

In this thesis I present two studies in which I designed, implemented and executed several statistical analyses to gain better insight about the post-transcriptional regulation events. I display that through integrated analyses of high-throughput omics data sets, we are able to identify new regulatory motifs and generate novel hypotheses about the gene regulation mechanisms. My analyses provide a data driven system overview by connecting multiple components of the post-transcriptional regulation. Performance of our predictive models, which were developed based on the known and identified regulatory elements, help us to have an idea about where we stand in the way of understanding genotype-phenotype relationships.

The first study I present aims to monitor the translational regulation in maturating human dendritic cells by utilizing RNA sequencing and ribosome profiling data sets generated at three different time points in the first 24h interval of the maturation process. Our major finding was the novel observation of ribosome accumulation at 5' and 3' untranslated regions of various genes during this 24h time-scale, possibly due to ABCE1 gene downregulation.

The second study aims to discover new mRNA and protein sequence motifs that are effective in post-transcriptional regulation, and to get a system overview between the different components of the post-transcriptional regulation via omics data integration. To this end, I developed an interpretable model for predicting the amount of proteins produced per mRNA (PTR ratio), which at the same time provides insights about the

contributions of various regulators effective in translation initiation, elongation, termination as well as protein degradation. While PTR ratios span more than 2 orders of magnitude, my integrative model predicts PTR ratios at a median precision of 3.2-fold. Moreover, our integrative model led to a new metric of codon optimality that captures the effects of codon frequency on protein synthesis and degradation. Altogether, this study showed that a large fraction of PTR ratio variation in human tissues can be predicted from sequence, and it identified many new candidate post-transcriptional regulatory elements.

Publications

B cell genomics behind cross-neutralization of SARS-CoV-2 variants and SARS-CoV

Johannes F. Scheid*, Christopher O. Barnes*, **Basak Eraslan***, Andrew Hudak, Jennifer R. Keeffe, Lisa A. Cosimi, Eric M. Brown, Frauke Muecksch, Yiska Weisblum, Shuting Zhang, Toni Delorey, Ann E. Woolley, Fadi Ghantous, Sung-Moo Park, Devan Phillips, Betsabeh Tusi, Kathryn E. Huey-Tubman, Alexander A. Cohen, Priyanthi N.P. Gnanapragasam, Kara Rzasa, Theodora Hatziioanno, Michael A. Durney, Xiebin Gu, Takuya Tada, Nathaniel R. Landau, Anthony P. West Jr., Orit Rozenblatt-Rosen, Michael S. Seaman, Lindsey R. Baden, Daniel B. Graham, Jacques Deguine, Paul D. Bieniasz, Aviv Regev, Deborah Hung, Pamela J. Bjorkman, Ramnik J. Xavier DOI: https://doi.org/10.1016/j.cell.2021.04.032

Cell (2021) 1-17, 184

Quantification and discovery of sequence determinants of protein per mRNA amount in 29 human tissues

Basak Eraslan*, Dongxue Wang*, Mirjana Gusic, Holger Prokisch, Bjorn M Hallstrom, Mathias Uhlen, Anna Asplund, Fredrik Ponten, Thomas Wieland, Thomas Hopf, Hannes Hahne, Bernhard Kuster, Julien Gagneur
DOI: 10.15252/msb.20188513

Molecular Systems Biology (2019) 15, e8513

A deep proteome and transcriptome abundance atlas of 29 healthy human tissues

Dongxue Wang*, **Basak Eraslan***, Thomas Wieland, Bjorn M Hallstrom, Thomas Hopf, Daniel Paul Zolg, Jana Zecha, Anna Asplund, Li-hua Li, Chen Meng, Martin Frejno, Tobias Schmidt, Karsten Schnatbaum, Mathias Wilhelm, Fredrik Ponten, Mathias Uhlen, Julien Gagneur, Hannes Hahne, Bernhard Kuster (2018) Biorxiv,
DOI: 10.15252/msb.20188503

Molecular Systems Biology (2019) 15, e8503

Throughput satisfaction-based scheduling for centralized cognitive radio networks

D.Gozupek*, **B.Eraslan***, F.Alagoz
DOI: 10.15252/msb.20188503

IEEE Transactions on Vehicular Technology (2012), vol.61, no.9, pp.4079-4094

An auction theory based algorithm for throughput maximizing scheduling in centralized cognitive radio networks

B.Eraslan, D.Gozupek, F.Alagoz
DOI: 10.15252/msb.20188503

IEEE Communications Letters (2011), vol.15, no.7, pp.734-736

Contents

1 Introduction

1.1 Biological background

1.1.1 Multiple layers of gene expression regulation

The central dogma of biology describes that the information in the genome flows into proteins by mainly a two-step process: transcription of DNA into RNA molecules and translation of messenger RNA (mRNA) into protein molecules (Figure 1.1). This information flow is regulated in multiple, diverse ways acting on transcription, translation and degradation of the molecules which is collectively called 'gene expression regulation'.

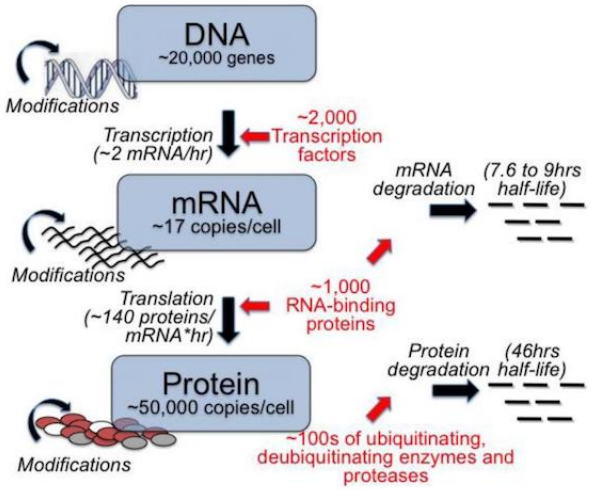


Figure 1.1: Depiction of Central Dogma of Biology. Taken from Ref. [3], Figure 1. Information encoded in the genome is decoded into proteins through transcription and translation. The abundance levels of mRNA and protein molecules are regulated by several factors that determine transcription, translation, mRNA degradation and protein degradation rates. Numbers are for illustration purposes and represent overall estimates for mammalian cells [3, 4].

Orchestration of transcriptional gene regulation depends on chromatin accessibility [5] and thousands of DNA-binding regulators, which directly or indirectly bind to specific sites in promoters or enhancers of regulated genes [6]. In eukaryotes transcribed precursor mRNAs are transformed into mature mRNAs through splicing, during which introns are removed and exons are joined together [7]. Alternative splicing generates an average of four transcript variants per human gene [8].

For a long time the general understanding was that the main gene expression regulation occurs during transcription, and mature mRNA levels are good proxies for the resulting protein levels. However, with the advances in high-throughput biological data acquisition techniques, transcripts and proteins quantified at genomic scales revealed that even though the sequence of an mRNA determines the amino acid sequence of the resulting polypeptide, there is no trivial relationship between the concentration of a transcript and the concentration(s) of the protein(s) derived from that particular transcript [3].

The major processes that influence mRNA-protein concentration dynamics are translation initiation rates, translation elongation rates, ribosome recycling rates and protein degradation rates. Each of these mechanisms are investigated for decades on a single-gene-basis and various factors modulating these processes have been revealed. Technological advances in generating genomewide measurements of various biological data modalities now enable us to study these processes in the global scale and understand the effects of the underlying regulatory mechanisms on the cellular protein throughputs.

In the following sections of this chapter first I describe the four main high-throughput omics data types that have been integrated throughout the studies explained in the rest of this thesis. After that, I provide a background on our current understanding of the mRNA - protein level dynamics in various scenarios and point out the known mRNA and protein sequence elements that play important roles in post-transcriptional gene expression regulation. Finally, I outline the scope and contributions of this thesis.

1.1.2 Major high-throughput data modalities utilized in understanding translational and post-translational regulation mechanisms

Several high throughput data acquisition protocols have been developed to measure the molecular abundance and the functional activities of various molecules in the cell. With the advance of these techniques now we are able to investigate the contribution of transcription, translation and post-translational events to the gene expression regulation at a genome-wide scale. Next generation RNA-sequencing [9], and label-free mass-spectrometry based peptide quantification [10] are two of the most widely used techniques to quantify the complete set of transcripts and proteins in a cell, respectively. Complementing to these methods, ribosome profiling [11] and enhanced UV crosslinking and immunoprecipitation (eCLIP) [12] are examples of experimental protocols that we utilize to discover the occupancy patterns of ribosomes and RNA-binding proteins, which in turn help disentangling the translation dynamics.

1.1.2.1 Next generation RNA sequencing

The transcriptome is the whole set of transcripts in a cell in a particular condition. Therefore, accurate quantification of the transcriptome is essential for understanding the first level of gene expression regulation, that is transcriptional regulation. The

key objectives of transcriptomics are to catalogue all isoforms of genes, as well as non-coding RNAs, and to quantify the changing expression levels of each transcript in different conditions. Before high-throughput RNA sequencing, several hybridization or sequence-based technologies were developed to infer and quantify the transcriptome. Hybridization-based methods were based on incubating fluorescently labelled cDNA with custom-made microarrays [13, 14, 15, 16]. However, these approaches had to rely on the existing knowledge about genome sequence and had a limited dynamic range of detection due to background and saturation of signals. Moreover, comparing expression levels across different experiments was often difficult and required complicated normalization methods [9].

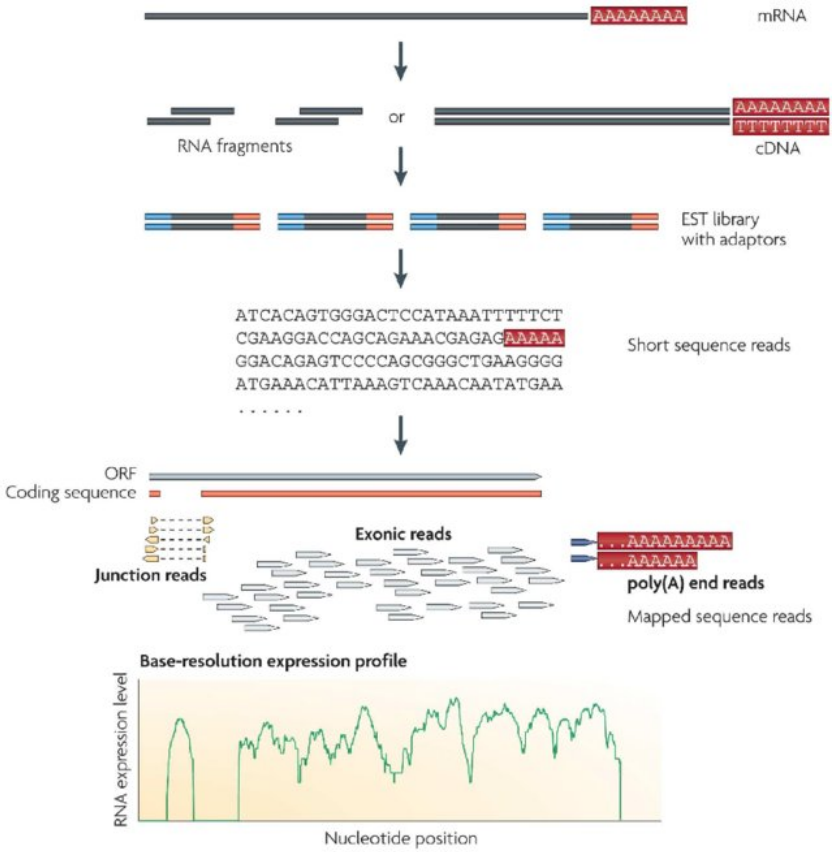


Figure 1.2: Typical next generation RNA-seq workflow. Taken from Ref. [9], Figure 1. RNAs are converted into a library of cDNA fragments through with or without RNA fragmentation. Then sequencing adaptors are added to each cDNA fragment which is later on sequenced at single or both ends. The resulting sequence reads are mapped to the reference genome to get a base-resolution gene expression profile.

Development of high-throughput DNA sequencing methods also gave rise to a new technique, namely RNA-seq, which enabled better mapping and quantifying of the transcriptomes. In RNA-seq the population of RNA molecules is converted to a cDNA

library which has adaptor sequences attached to one or both ends. This is generally followed by an amplification step which increases the overall signal strength, but can also introduce bias based on the base composition of the cDNA sequences [17]. Each molecule is then sequenced either from one end (single-end sequencing) or both ends (paired-end sequencing) to obtain short sequences which are then either mapped to a reference genome or assembled de novo to produce a transcription map (Figure 1.2).

1.1.2.2 Mass-spectrometry based shotgun proteomics

Proteins are the functional molecules that regulate various processes in the cells and the 'proteome' describes the total set of proteins encoded by the genome [18]. Even though the structures and functions of selected proteins have been studied for decades by the use of biochemical and biophysical methods such as quantitative western blots or ELISA assays, the study of the proteome has become available with the development of the mass-spectrometry (MS) based methods [19]. MS based methods are also able to identify and localize the modified amino acids in the polypeptide chain, and to determine the composition of the subunits of the protein complexes [19].

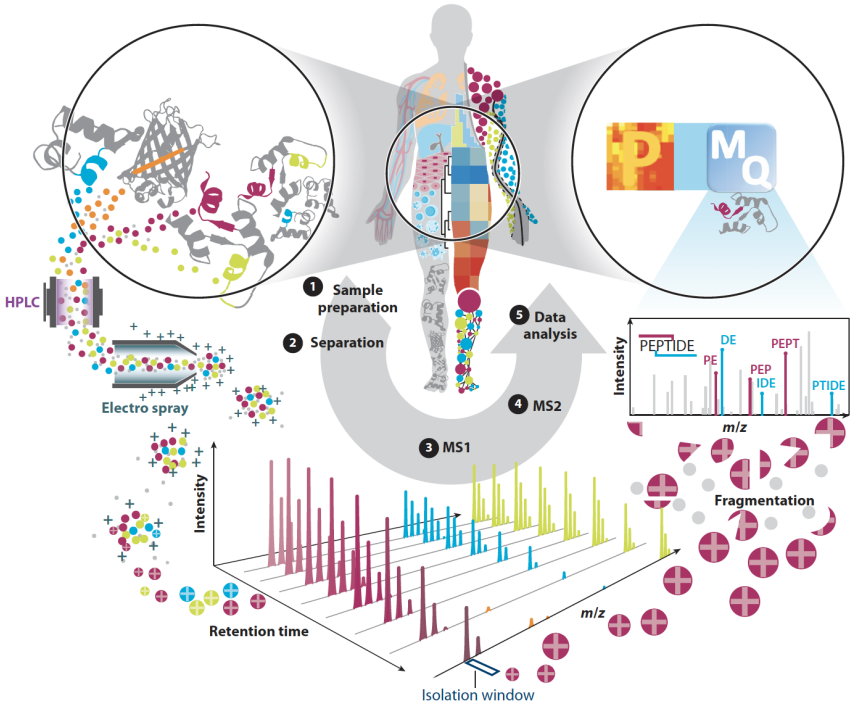


Figure 1.3: Typical shotgun proteomics workflow. Taken from Ref. [20], Figure 2. Proteins are extracted from a sample and are digested to peptides with a digestion enzyme. After high-performance liquid chromatography (HPLC) separation, peptides are ionized and fed into a mass-spectrometer. First and second stage mass spectra (MS1 and MS2) are recorded and analyzed by the computational proteomics software.

In top-down mass-spec proteomics, the proteins are studied as intact molecules by the mass spectrometry. Even though this approach enables to detect all modifications on the same molecule, bottom-up approach, in which proteins are first digested into peptides, is more common in use. In bottom-up approcah first the proteins are digested into peptides, by a protease enzyme such as trypsin. Then the resulting mixture of peptides are separated by liquid chromatography coupled with electrospray ionization. Then the peptides are fed into the mass spectrometer, where they are fragmented to generate the MS/MS spectra. The peaks in a mass spectrum correspond to molecular features of the peptides. In data-dependent acquisition (DDA) methods, no prior knowledge about the proteome is integrated to the pipeline and a full spectrum of the peptides is acquired through MS1 and MS2 levels [19]. Thereon, the information in the spectra is computatinally extracted by softwares such as MaxQuant to identify and quantify specific peptides (Figure 1.3). The identification of the proteins with the sequenced peptides also requires an additional computational step. Even though current approaches have made it realistic to distinguish and quantify proteoforms, the different molecular forms of a protein translated from the same gene, the identification of the complete set of proteoforms still remains a challenge due to the huge combinatorial space defined by the possible combinations of protein post-translational modifications [19, 20].

1.1.2.3 Ribosome profiling

Ribosome profiling is an emerging high-throughput technique that enables monitoring in vivo translation and helps us to get a better understanding of the translation process and its role in modulating protein levels [11]. The position of a ribosome on an mRNA transcript can be determined by its 28-30 nucleotides footprint which is protected from nuclease digestion [21]. In this method, the ribosome protected fragments are selected, high-throughput sequencing is utilized to determine the pool of footprints of the ribosomes that are translating in vivo (Figure 1.4). Finally, these sequences are mapped back to the reference genome to extablish the ribosome occupancy profiles per gene.

Ribosome profiling data provides various information about the ongoing translation processes. First of all, since these protected fragments indicate the exact location of the ribosomes, fragment density profiles on the transcripts indicate the translation initiation and end sites (Figure 1.5) and helps us identify the alternative reading frames. Second, ribosome profiling data provides the ribosome distribution along the transcripts in subcodon resolution and therefore can be utilized in identification of the ribosome stalling sites, and delineating the differences in the elongation speed at different regions of the transcripts, as well as the variability in codon decoding times [22, 23]. Finally, apart from these spatial information about mRNA translation, ribosome profiling data provides a proxy for the translation efficiency per gene when coupled with the corresponding mRNA levels. Ribosome footprint density of a gene, which is the total number of ribosome protected fragments mapped to a gene divided by the number of mapped mRNA fragments, provides an estimate of the synthesis level of that protein from its corresponding mRNA transcripts (Figure 1.5).

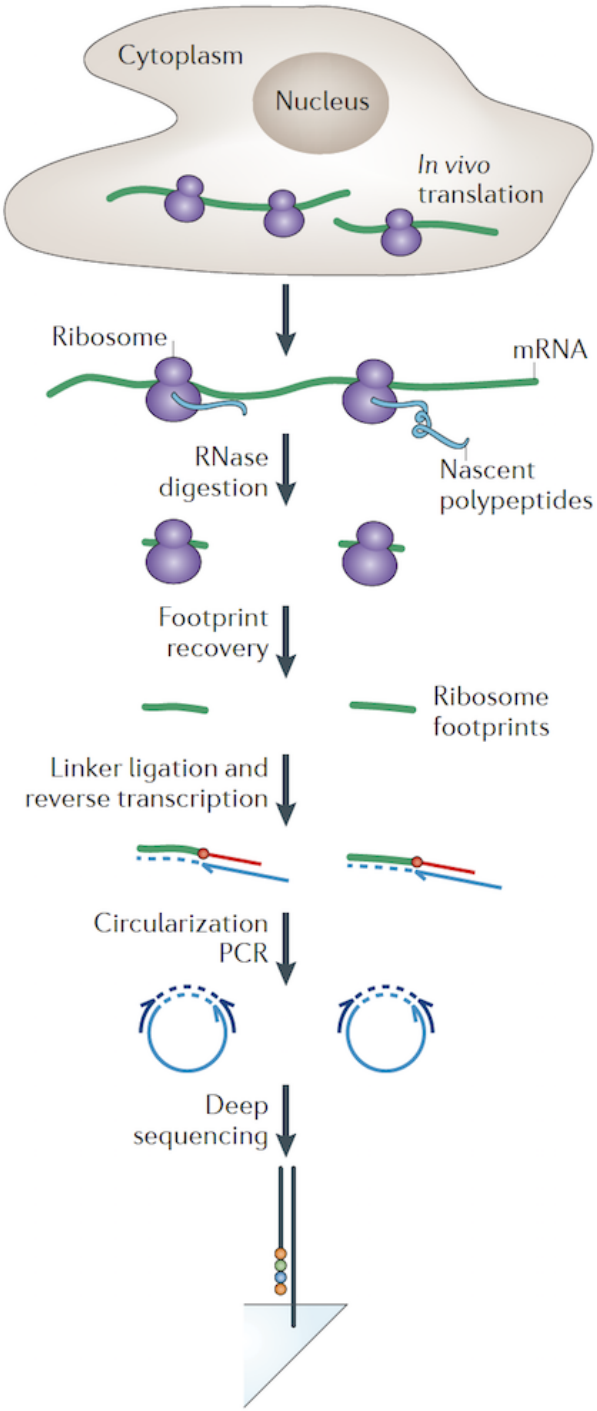


Figure 1.4: Typical ribosome profiling workflow. Taken from Ref. [22], Figure 1. Ribosomes physically enclose 28-30 nucleotides of the transcript, protecting this region from nuclease digestion. These ribosome footprints are recovered and converted into a DNA library which is deep sequenced by high-throughput sequencing. Mapping these sequences back to the reference genome provides high-precision measurements of in-vivo translation [11].

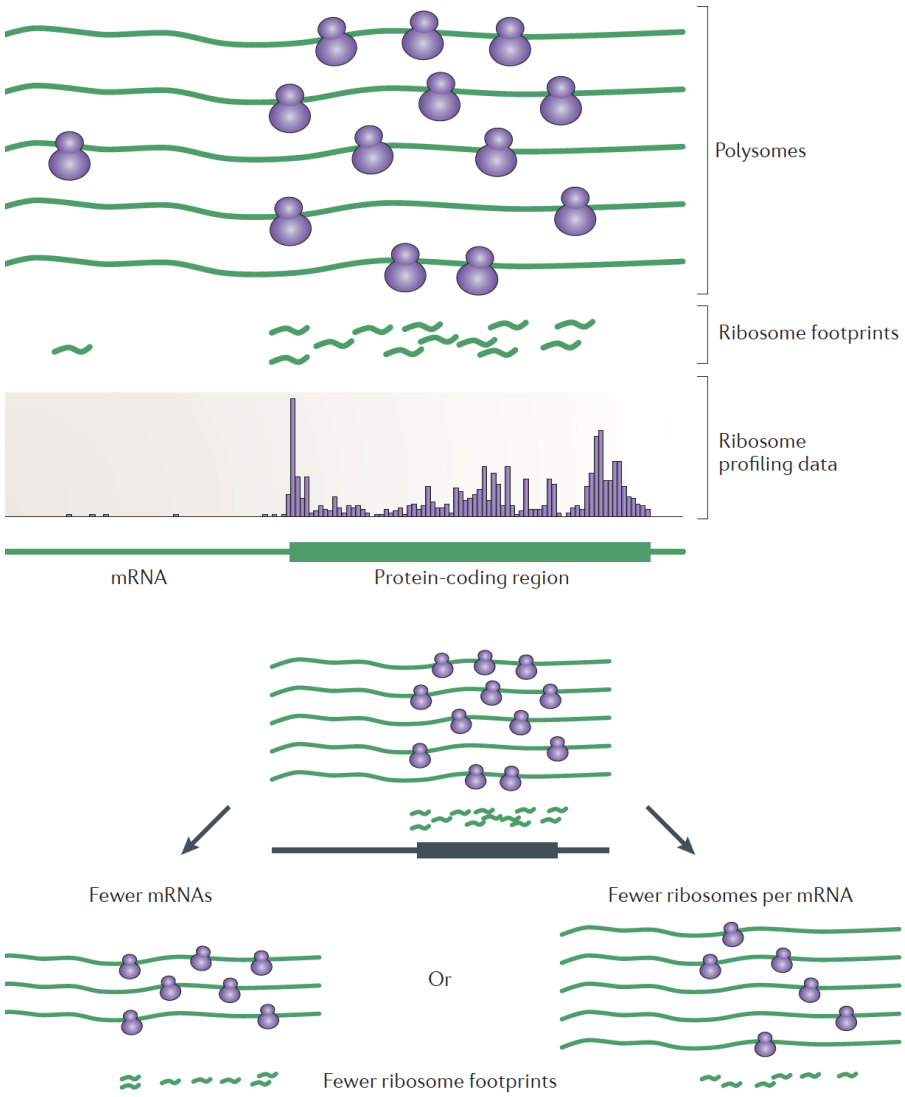


Figure 1.5: Multiple information obtained from ribosome profiling data. Taken from Ref. [22], Figure 2. Ribosome profiling data provides us various information about the translation process. The density of the ribosomes at each position of the transcipt enables us to recognize the alternative start and stop sites, as well as ribosome stalling locations and variance in codon decoding efficiency. Ribosome footprints incidate the total number of ribosomes engaged in translation of a transcript. Accordingly, coupled with the mRNA level information, ribosome profiling also provides ribosome density per gene which is a proxy of the translational efficiency.

1.1.2.4 Enhanced UV crosslinking and immunoprecipitation (eCLIP)

RNA binding protein (RBP) interacts with an RNA to regulate its translation, localization, stability, modification and processing [24]. Accordingly, understanding the in-

teractions between RBPs and RNA is an important step to improve our knowledge about translational gene expression regulation. In recent years several high throughput sequencing methods have been developed to identify the targets of the RBPs, such as RNP immunoprecipitation (RIP), chemically induced covalent crosslinks (CLIP), photoactivitable-ribonucleotise-enhanced CLIP (PAR-CLIP), individual-nucleotide-resolution CLIP (iCLIP) and enhanced CLIP (eCLIP) [24]. eCLIP is a variant of of iCLIP with upgraded sensitiviy [12]. The pipeline of the eCLIP-seq protocol is displayed in Figure 1.6

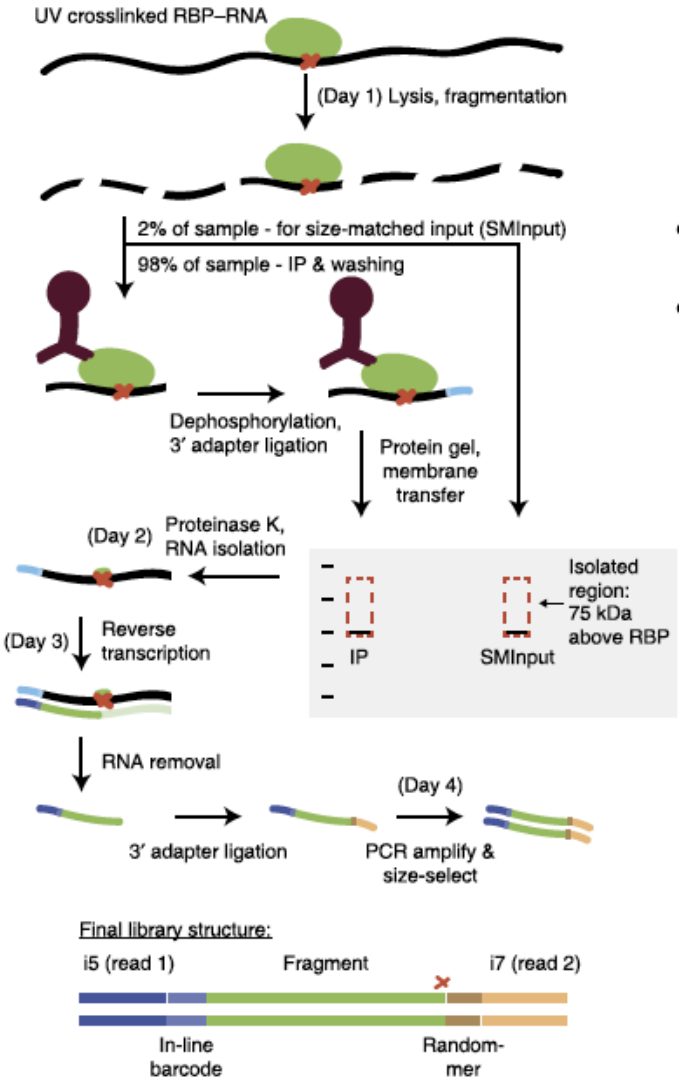


Figure 1.6: eCLIP-seq pipeline. Taken from Ref. [12], Figure 1. RBP-RNA binding is stabilized using UV crosslinking which is followed by cell lysis, RNA fragmentation, and RNA-protein immunoprecipitation with an RBP specific antibody. The immunoprecipitated RNA-RBP complex is resolved, the RNA is recovered and subjected to high-throghput sequencing after reverse transcription into cDNA.

1.1.3 Measurement of mRNA and protein turnover rates

1.1.4 Background about the relationship between mRNA and protein levels under steady-state conditions

The relationship between mRNA and protein levels has been investigated under various scenarios, such as steady state, long-term state changes or short term adaptations [25, 3, 26, 27]. Even though it is challenging to define the 'steady state' for the cells, large amounts of cells have been assumed as being at steady state if the average protein and mRNA levels remain relatively stable over several hours.

In these studies two main approaches are followed to explore the relationship between mRNA and protein levels: *i)* observing mRNA and protein measurements of the same gene across different individuals, conditions, tissues or time points *ii)* inspecting the mRNA and protein levels of different genes in a certain tissue, condition or time point. First approach examines to what extend the variation in the mRNA levels are propogated to the protein levels. On the contrary, the second approach aims to learn the variation of protein versus mRNA amounts of different genes and to find out to what extend differences in mRNA levels are reflected to the respective protein levels.

Some of the studies that have followed the first approach have analysed gene-specific correlation of mRNA and protein measurements across human tissues [28, 29, 30]. These studies have suggested that the comparative analysis of the mRNA and protein levels shows high correlation and a constant protein–mRNA ratio is preserved across human cell lines. However, following studies [31, 32] have claimed these suggestions to be inaccurate due to insufficient statistical analysis and have advocated the high impact of differential post-transcriptional regulation on shaping tissue-type-specific proteomes.

Studies that analysed the protein and mRNA correlations across various number of genes have also reported conflicting results; even though initial studies stated poor overall correlation between mRNA and protein concentrations (Spearman rank correlation between 0.45 and 0.61) [33, 34, 35, 36, 4, 37], more recent studies concluded that mRNA levels can explain most of the variance in steady-state protein levels [38, 39, 26] (%84, %85, %52 respectively). Nevertheless, all these studies have observed a high dynamic-range difference between transcriptome and proteome levels, which reflects the importance of post-transcriptional regulation on determining the protein abundances.

It is important to note that these studies have displayed different factors which may lead to having different conclusions about mRNA-protein level dynamics. First of all, the quality of the input data generated by different experimental techniques has a big role in the significance of the mRNA-protein correlations. As an example, Jovanovic et al. [26] displayed that correcting for mRNA and protein reproducibility increased the mRNA - protein levels correlations from 42% to 52% in mouse dendritic cells. Second, different data anlysis strategies can lead to drastically different conclusions. As an illustration, Schwanhäusser et al. [4] had concluded that gene-to-gene differences in protein synthesis rates contributes most to final protein levels (55%) while mRNA abundance explains only 40% of the variation and degradation of mRNA and protein plays minor roles. However, reanalysis of this data by Li et al. [38] came up with completely different

conclusions, arguing that mRNA levels may explain the variance in protein levels by 84% while protein synthesis contributes by only 8%.

1.1.5 Sequence features that are important for the post-transcriptional gene expression regulation

Parts of the introduction presented in this section are part of the manuscript "Quantification and discovery of sequence determinants of protein per mRNA amount in 29 human tissues" from Eraslan and Wang et al. 2019 [1].

Decades of single-gene studies have revealed numerous sequence elements affecting initiation, elongation, and termination of translation as well as protein degradation. Eukaryotic translation is canonically initiated after the ribosome, which is scanning the 5' UTR from the 5' cap, recognizes a start codon. Start codons and secondary structures in 5' UTR can interfere with ribosome scanning [40, 41]. Also, the sequence context of the start codon plays a major role in start codon recognition [42]. The translation elongation rate is determined by the rate of decoding each codon of the coding sequence [43, 44, 45]. It is understood that the low abundance of some tRNAs leads to longer decoding time of their cognate codons [46], which in turn can lead to repressed translation initiation consistent with a ribosome traffic jam model (reviewed in [45]). However, estimates of codon decoding times in human cells and their overall importance for determining human protein levels are highly debated [47, 48, 45]. Secondary structure of the coding sequence and chemical properties of the nascent peptide chain can further modulate elongation rates [49, 50, 51, 52]. Translation termination is triggered by the recognition of the stop codon. The sequence context of the stop codon can modulate its recognition, whereby non-favorable sequences can lead to translational read-through [53, 54, 55, 56]. Furthermore, numerous RNA binding proteins (RBPs) and microRNAs (miRNAs) can be recruited to mRNAs by binding to sequence-specific binding sites and can further regulate various steps of translation [57, 58, 59, 60, 61, 62]. However, not only predicting the binding of miRNAs and RBPs from sequence is still difficult, but the role of few of these binding events in translation is well understood.

Complementary to translation, protein degradation also plays an important role in determining protein abundance. Degrons are protein degradation signals which can be acquired or are inherent to protein sequences [63]. The first discovered degron inherent to protein sequence was the N-terminal amino acid [64]. However, the exact mechanism and its importance are still debated, with recent data in yeast indicating a more general role of hydrophobicity of the N-terminal region on protein stability [65]. Further protein-encoded degrons include several linear and structural protein motifs [66, 63, 67], or phosphorylated motifs that are recognized by ubiquitin ligases [68]. Altogether, numerous mRNA and protein-encoded sequence features contribute to determining how many protein molecules per mRNA molecule cells produce. However, it is known neither how comprehensive the catalogue of these sequence features is nor how they quantitatively contribute to protein-per-mRNA abundances.

1.2 Aims and scope of this thesis

This thesis presents an in-depth investigation of the mechanisms and effects of post-transcriptional gene expression regulation in i) maturating human dendritic cells (DC) ii) samples of 29 healthy human tissues. Furthermore, it introduces a machine learning model which uses sequence features effective in post-transcriptional regulation to predict the amount of protein produced per mRNA molecule under steady-state conditions.

In order to get a better understanding of the links between the different components of the post-transcriptional gene expression regulation, I have designed, implemented and executed sets of integrated statistical analyses on various types of omics data sets. Moreover, I have planned follow-up experiments which were executed by our collaborators to validate some of our findings. The outcomes of these experiments are also included in the respective chapters. The major contributions of my statistical analyses can be summarized as follows:

- We spotted a novel observation of ribosome accumulation at 5' UTR and 3' UTR sites during DC maturation, possibly as a result of ABCE1 gene downregulation.
- We provided a better understanding of the functional relationship between the mRNA and protein levels in 29 human tissues.
- We generated a comprehensive catalogue of known sequence features controlling protein-to-mRNA (PTR) ratios and quantification of their effects.
- We identified and validated novel sequence features that are predictive of PTR ratios.
- We dissected the effects of the sequence elements that are predictive of PTR ratios into their effects in regulating distinct post-transcriptional steps: mRNA degradation, protein translation and protein degradation by the integration of independent mRNA half-life, ribosome profiling and protein half-life data sets.
- We proposed a new metric of codon optimality that captures the effects of codon frequency on protein synthesis and degradation.

Despite the progress in recent years, high-quality proteomic data is still expensive to obtain compared to acquiring the corresponding RNA sequencing profiles. Therefore, good machine learning models that are capable of predicting the protein levels is invaluable for the field. For this purpose, I also developed a novel and interpretable machine learning model which predicts the amount of protein produced per mRNA molecule by utilizing the sequence features that are important for the translation efficiency and the protein degradation rates. While PTR ratios span more than 2 orders of magnitude, my integrative model predicts PTR ratios at a median precision of 3.2-fold while providing insights about the features that are more informative for this prediction task.

2 Translational gene expression regulation in maturating human dendritic cells

2.1 Maturation of dendritic cells

Dendritic cells (DCs) are vital for the orchestration of immune responses. First, they prevent the immune system from reacting against "self" and thereby help to avoid autoimmune diseases [69]. Secondly, in response to danger signals like microbial components or malignant cell signatures, they are responsible for the induction of adequate immune responses leading to elimination of infection or prevention of tumorigenesis, respectively [70, 71]. DCs are the professional antigen-presenting cells that play a central role in the initiation andregulation of immune responses. DCs regulate both the innate (e.g. macrophages, granulocytes and natural killer (NK) cells) and the adaptive (e.g. T and B cells) immunity [71].

Upon activation, DCs phenotypically mature. This involves the upregulation of several surface molecules, most importantly MHC II and the costimulatory molecules CD40, CD80 and CD86 [72]. Depending on the type of danger signal leading to their activation, DCs start to express a specific cytokine profile, thereby polarizing naive T cells to differentiate into the various T cell subsets [73].

To adapt to changes in the environment, immune cells need to rapidly modulate their protein abundance. During this dynamic fast-evolving process, in order to avoid slow de novo transcription, gene expression is predominantly regulated by translational control [4, 74, 26]. Accordingly, one of the most translationally regulated pathways in DCs is protein synthesis itself [75].

To reveal regulatory processes during DC maturation, lots of studies have either focused on mRNA or protein production. However, these are suboptimal proxies as they are highly regulated processes themselves. Therefore, the best representation of the actual protein repertoire of a cell is given by translatome studies.

In this study we elucidated the effects of a cytokine cocktail containing IFNG, TNFA, IL-1B, PGE2 and the TLR7/8 agonist R84 on the translatome of primary human monocyte-derived DCs. By ribosomal profiling (RPF-seq), a technique based on the isolation of RNAs contained within ribosome-protected fragments (RPFs) followed by next-generation sequencing and RNA abundance quantification, we obtained snapshots of actively translated RNA. Collecting matched RNA-seq and RPF-seq data at 0 (immature DCs), 4h and 24h time points after the LPS maturation stimulus enabled us

to monitor the changes in the translation efficiency over time during the maturation process.

2.2 Ribosome density change in untranslated regions during dendritic cell maturation

In order to get insights abouts the translation dynamics during dendritic cell differentiation, we inspected the change in ribosome densities in gene coding and 5' and 3' untranslated regions. Here, ribosome densities are approximated by the ratio of the number of ribosome profiling reads over the RNA sequencing reads mapped to a certain region because of the reasons explained in the introduction section (see Figure 1.5).

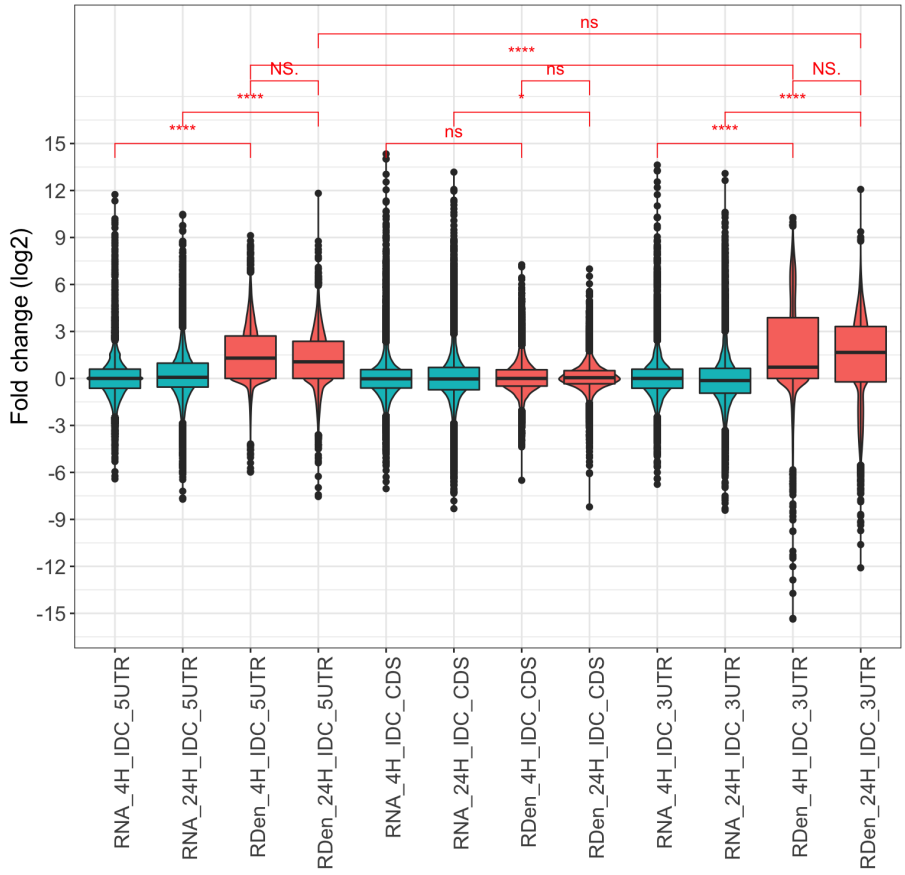


Figure 2.1: Distribution of 4h/iDC, 24h/iDC RNA and ribosome density fold change values for 5' UTR, coding and 3' UTR regions. Distributions of the 4h/iDC and 24h/iDC log2 fold changes in number of mapped RNA reads (RNA) and ribosome densities (RDen) for coding and untranslated regions (denoted as 5UTR, CDS, and 3UTR in the x-axis text). We observe an overall increase in the ribosome densities in 5' UTR and 3' UTR regions at 4h and 24h time points in the maturation process.

It's expected that sequencing reads originating from ribosome protected fragments predominantly map to coding regions. Therefore we were surprised by the observation that an increased number of RPF reads mapped to 5' and 3' UTR UTR regions at 4h as well as at 24h. Since the ribosome profiling protocol that was used in the experiments included the isolation of 80S monosomes after nuclease digestion via a sucrose gradient, the sequencing reads had to be originated from true ribosome protected fragments. A comparison of the ribosome density on the features 5' UTR, CDS and 3' UTR for all samples showed that the increase is not a consequence of elevated expression levels and is apparently a global effect rather than specific for certain genes (Figure 2.1, $p-value < 0.001$). The ribosome density on 5' and 3' UTRs was increased at 4h levels stayed on a comparable level at 24h (Figure 2.1, $p - value > 0.05$).

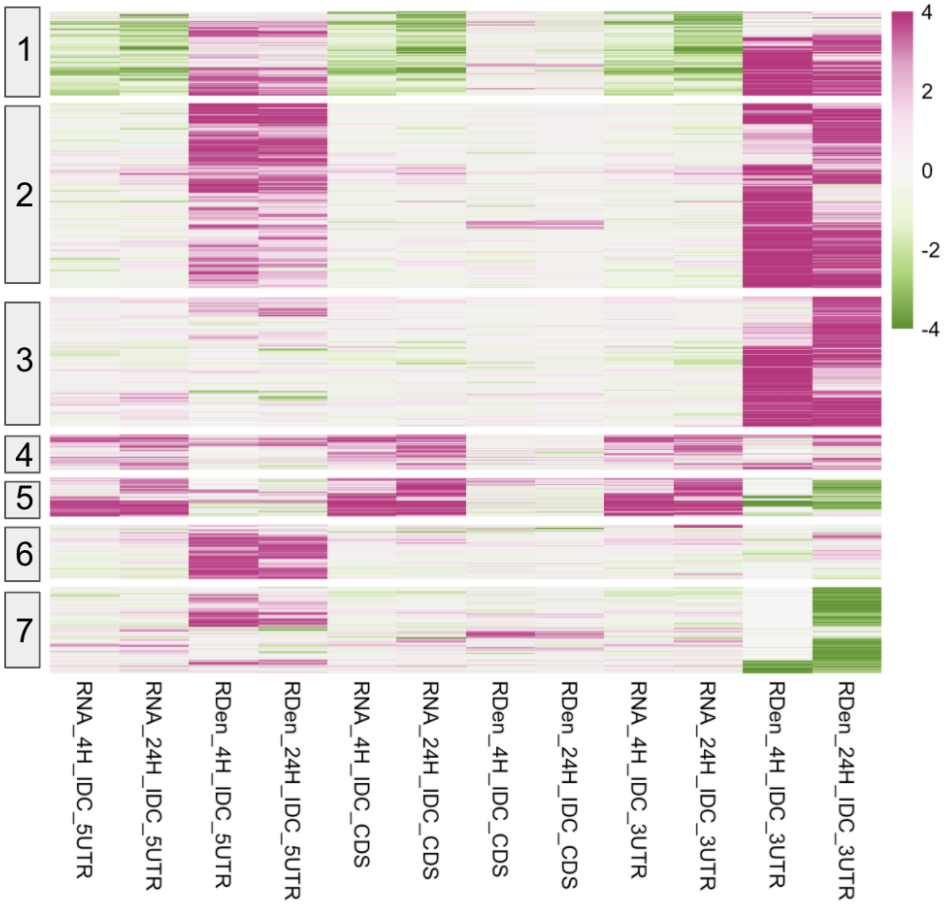


Figure 2.2: 4h/iDC, 24h/iDC mapped mRNA reads and ribosome density fold change values for 5' UTR, coding and 3' UTR regions. Heatmap displaying the 4h/iDC and 24h/iDC log2 fold changes of the number of mapped mRNA reads and ribosome densities for coding and untranslated regions.

When we had a closer look at the genes which has at least 2 fold ribosome density change in any of the 3 inspected regions at 4h and/or 24h time points, seven distinct

fold-change patterns across the genes (Figure 2.2) were detected. Group 1 which displayed a significant down-regulation in RNA levels at 4h and 24h time points (Figure 2.2) were enriched in genes that are involved in regulation of multicellular organismal process. Conversely, groups 4 and 5 were transcriptionally upregulated during the maturation process (Figure 2.2). Gene set enrichment analysis reported that the genes in group 4 and 5 are enriched in interferon-gamma-mediated signaling pathway and regulation of vitamin D biosynthetic process respectively (Appendix Figure A.1). Remaining groups 2, 3, 6 and 7 did not show overall big change in the RNA expression levels but displayed significant change patterns in the 5' and 3' ribosome density approximations. The gene set enrichemnt analysis of the genes belonging to each of these groups were not particularly enriched in any specific biological or metabolic process.

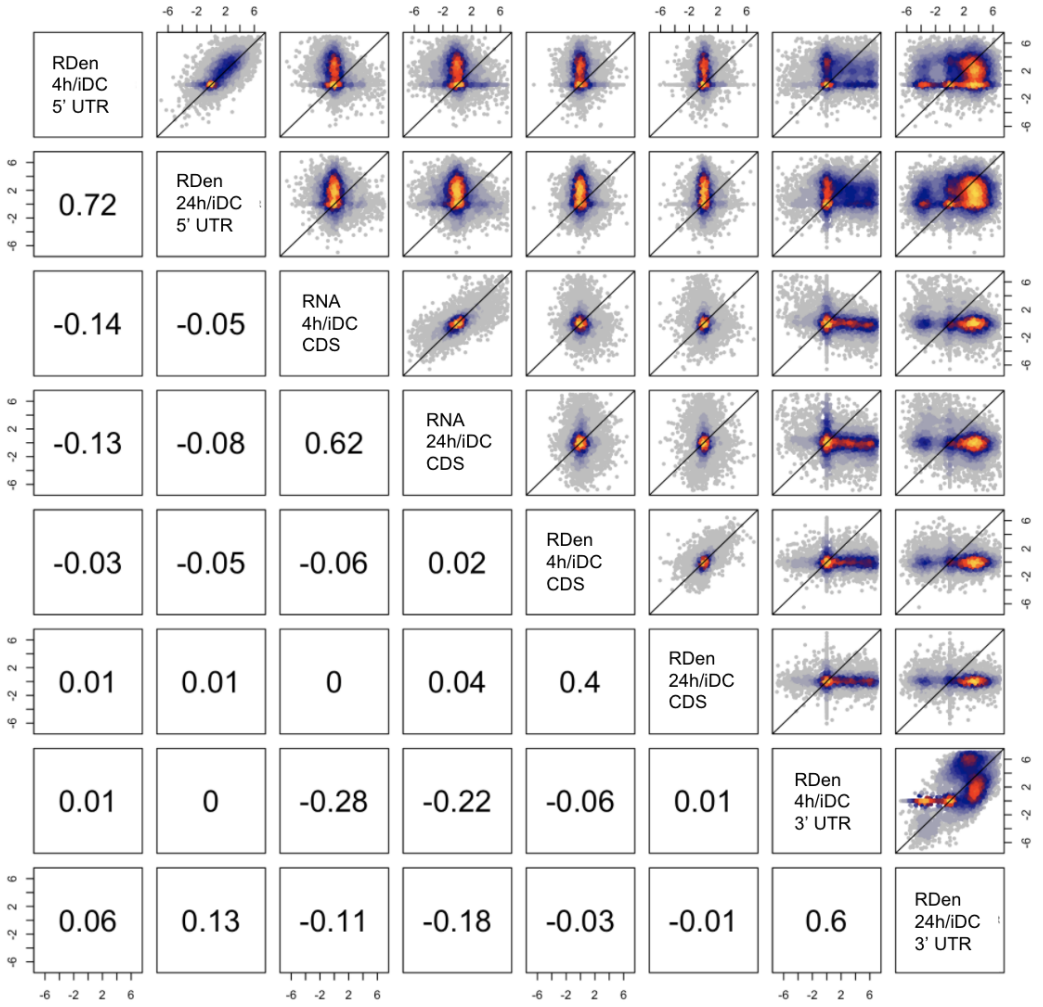


Figure 2.3: Pairwise correlations between 4h/iDC and 24h/iDC fold change values at 5' UTR, CDS and 3' UTR regions. Upper triangle of the figure displays the scatter plots showing the pairwise relationships between the 4h/iDC and 24h/iDC fold-change values. Lower triangle of the figure displays the corresponding Spearman's correlation coefficients.

Plotting the ribosome density fold change at 4h/iDC and 24h/iDC time intervals in 5' UTR, CDS and 3' UTR regions revealed no significant correlation between the ribosome density changes at these three regions (Figure 2.3) This implied that accumulation of ribosomes in 5' UTRs has no impact on translation of the coding sequence. Therefore, based on the analysis of the data at hand, the biological relevance of ribosomes located in 5' UTRs during maturation of DCs remains enigmatic. A growing number of publications report on so called small open reading frames (defined as being smaller than or equal to 300 nucleotides) found in several different organisms, which are located in regions annotated as non-coding [76, 77, 78] However, validation of their coding potential is often limiting. Since identification of small open reading frames by ribosome profiling requires the treatment of cells with a translation initiation inhibitor, e.g. harringtonine or lactimidomycin, our initial observation should be investigated in further studies.

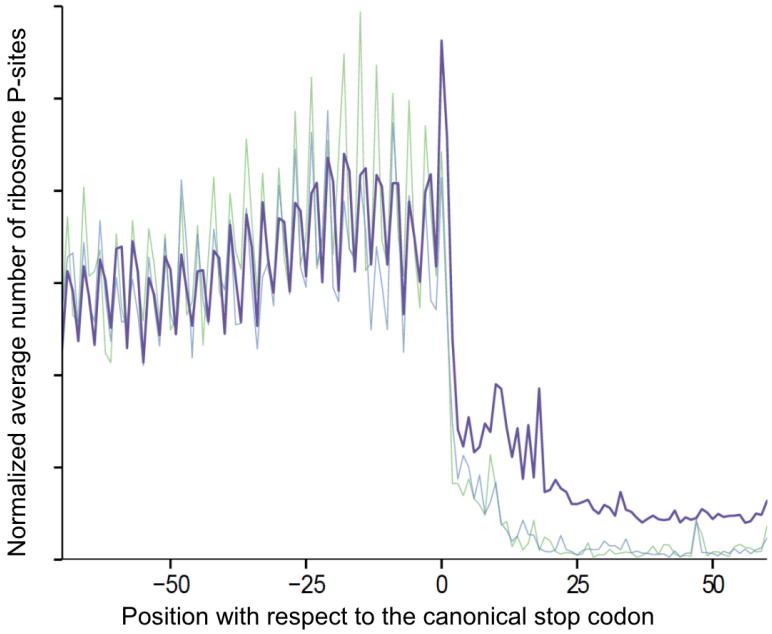


Figure 2.4: Metagene plot displaying loss of the three nucleotide periodicity of the ribosome P-sites after the stop codon. Metagene plot displaying the positions of the ribosome P-sites estimated based on the mapped RPF reads. Three different colors display the position based on three possible reading frames, purple color stands for the canonical reading frame. The periodicity is lost at the 3' UTR region.

Similar to the increase of ribosomes in 5' UTRs, the elevated ribosome density in 3' UTRs was unrelated to the ribosome density in the corresponding CDS (Figure 2.3). The accumulation of ribosomes in 3' UTRs at 24h could be a result of an impaired translation termination causing stop codon readthrough. A metagene plot of genes aligned at their stop codon showed that ribosomes entering the 3' UTR are losing the three nucleotide periodicity, a typical feature of RPF reads originating from active translating ribosomes,

rendering this explanation unlikely (Figure 2.4). However, in order to completely refute this possibility, further studies should be designed where mass-spectrometry data is generated for the search of the corresponding peptides.

In order to get some insights about the mechanisms that might lead to the observed ribosome recycling defect, we inspected of the genes that are known to be responsible for the ribosome release at the stop codon. Eukaryotic peptide chain release factor subunit 1 (eRF1, encoded by ETF1) and eukaryotic peptide chain release factor GTP-binding subunit ERF3A (eRF3, encoded by GSPT1) or the known rescue factor HBS1L showed no differences in their expression levels (Figure 2.5). Utilizing a defined in vitro system from rabbit reticulocyte lysate Skabkin et al. showed that post-termination complexes, i.e. after peptide release and in the absence of the ribosome recycling factor ABCE1, can start to diffuse along the mRNA and are able to rebind to codons cognate to the P-site tRNA, which they still carry [79]. In a more recent publication, the Green group showed by ribosome profiling of yeast samples that depletion of Rli1 (homolog of mammalian ABCE1), leads to an increase of ribosomes in 3' UTRs, which apparently reinitiate translation by a frame independent non-canonical mechanism [80]. In fact, inspection of our sequencing data revealed a slight decrease of ABCE1 by one third at 4h and a three-fold downregulation at 24h (Figure 2.5).

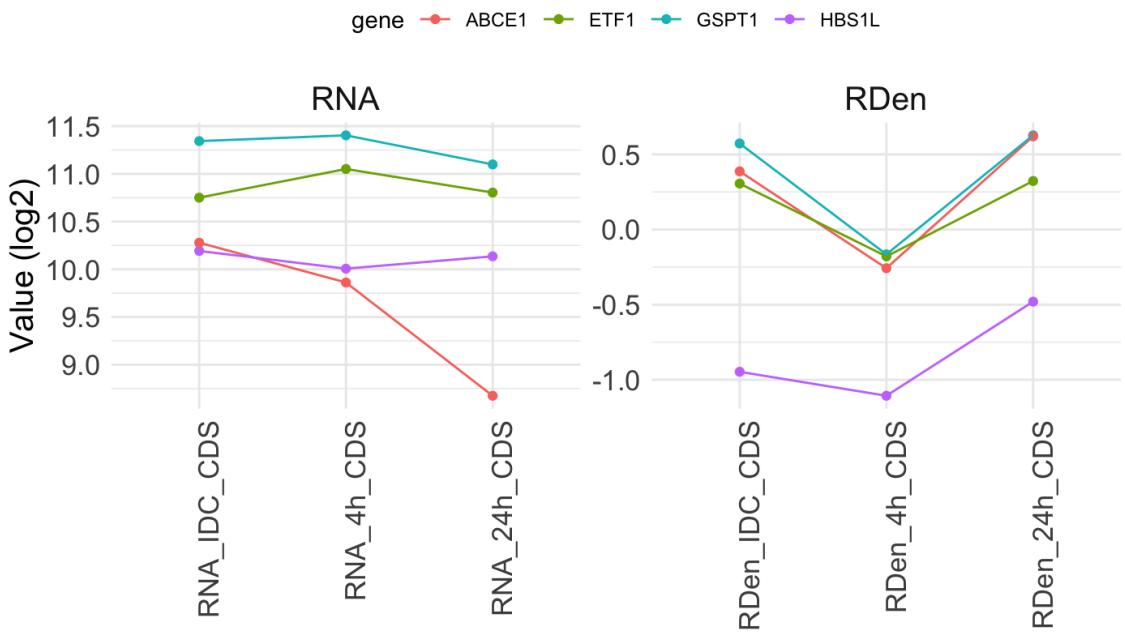


Figure 2.5: RNA expression and ribosome density values of ribosome recycling factors at iDC, 4h, and 24h. RNA expression of ribosome recycling factors ETF1, GSPT1 and HBS1L does not change over the 24h time frame during DC maturation. However, ABCE1 displays a significant ($FDR < 0.1$) down-regulation between iDC and 24h.

3 The relationship between the human transcriptome and the proteome

The methodology, results and figures presented in this section are part of the manuscript "Quantification and discovery of sequence determinants of protein per mRNA amount in 29 human tissues" from Eraslan and Wang et al. 2019 [1] and "A deep proteome and transcriptome abundance atlas of 29 healthy human tissues" from Wang and Eraslan et al. 2019 [2]

The mRNA and protein concantration dynamics of gene i is commonly modelled [81, 82] as:

$$\frac{dM_{it}}{dt} = k_{si} - k_{mi}M_{it} \tag{3.1}$$

$$\frac{dP_{it}}{dt} = k_{ri}M_{it} - k_{pi}P_{it} \tag{3.2}$$

where M_{it} and P_{it} denote the cellular mRNA and protein concentrations at time t, k_{si} is the mRNA transcription rate [mRNA/min], k_{mi} is the mRNA degradation rate [mRNA/min], k_{ri} is the protein translation rate [protein/(mRNA*min)] and k_{pi} is the protein removal rate of gene i. If the translation and protein removal rates did not vary by gene and by condition/tissue, we would be observing perfect correlation between the mRNA and the protein levels. However mRNA-protein correlations reported by many studies [33, 34, 35, 36, 4, 37, 31, 32] are far away from being perfect, which attracts the attention to the role of post-transcriptional events in gene expression regulation.

In order to gain a better understading of the steady-state mRNA-protein level dynamics in and across-tissues, we profiled the proteomes and transcriptomes of adjacent cryo-sections of 29 histologically healthy tissue specimens collected by the Human Protein Atlas project [83] by utilizing label-free quantitative proteomics and RNA-Seq [2]. We modeled every gene with a single transcript isoform because there was little evidence for widespread expression of multiple isoforms and to avoid practical difficulties of calling and quantifying isoform abundance consistently at mRNA and protein levels. The number of genes with multiple quantified isoforms on protein level was small (10% of the 13,664 genes with a protein detected in at least in one tissue). Also, for 5,636 (43%) genes the same isoform was the most abundant one across all tissues at mRNA level (out of 12,978 genes with at least one mRNA transcript isoform expressed [$FPKM > 1$] in at least in one tissue) (Figure 3.1).

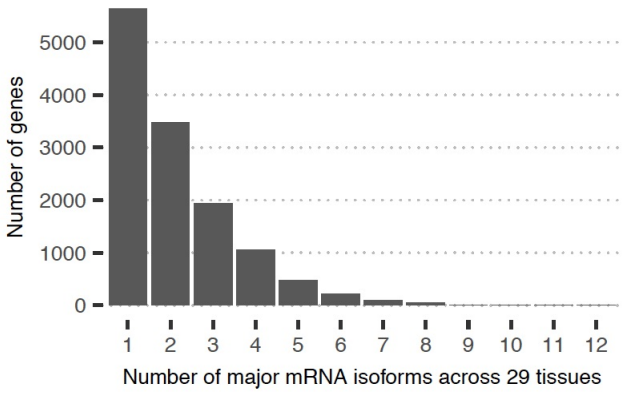


Figure 3.1: mRNA isoform distribution Distribution of the number of different major mRNA isoforms each gene has across 29 tissues.

Moreover, 4,303 (34%) genes had a perfect match between the RNA-Seq-defined and the proteomics-defined major isoform in all the tissues they were detected (out of 12,920 genes with matched protein and mRNA measurements). For the remaining genes, there were some mismatches between the RNA-Seq-defined and the proteomics-defined major isoforms in a varying number of tissues, yet the number of matched RNA-Seq-defined and proteomics-defined major isoforms were larger than the unmatched ones in almost all tissues (Figure 3.2).

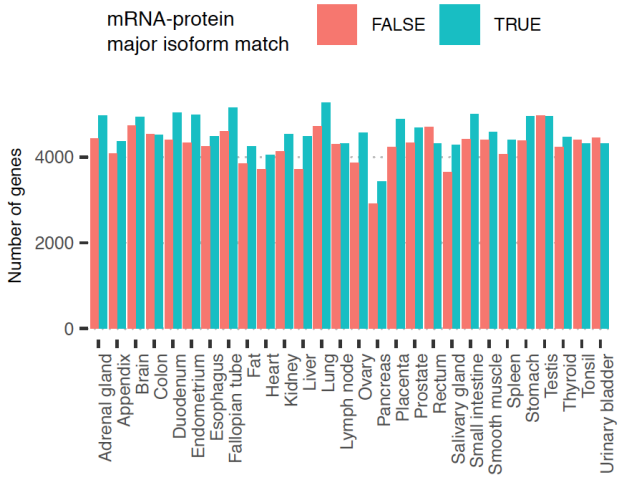


Figure 3.2: mRNA - protein isoform match Number of genes in each tissue with matched and unmatched mRNA and protein major isoforms.

Since we were restricted by the small number of isoform counts on proteome level, we defined the transcript isoform with the largest average protein abundance across tissues as its major transcript isoform. The mRNA levels were estimated from RNA-Seq data by subtracting length and sequencing-depth-normalized intronic from exonic coverages. RNA-Seq technical replicates were summarized using the median value. In our

downstream analysis, for each gene we required at least 10 sequencing-depth-normalized mRNA reads per kilobase pair, because low expression values on transcript and protein levels were associated with a larger measurement error. Altogether, this analysis led to matched quantifications of protein and mRNA abundances for 11,575 genes across 29 tissues, where an average of 7,972 (69%, minimum 7,300 and maximum 8,869) protein-to-mRNA ratios were quantified per tissue.

3.1 mRNA - protein level variations across genes

The proportion of variance in protein levels explained by mRNA levels of the same tissue (R^2) ranged from 20% (ovary) to 39% (liver) when an orinary least squares model was fit independently for each tissue. These relationships corresponded to Pearson correlation coefficients ranged from 0.45 (ovary) to 0.62 (liver) (Figure 3.3). These numbers are in line with previous studies [35, 4, 37]

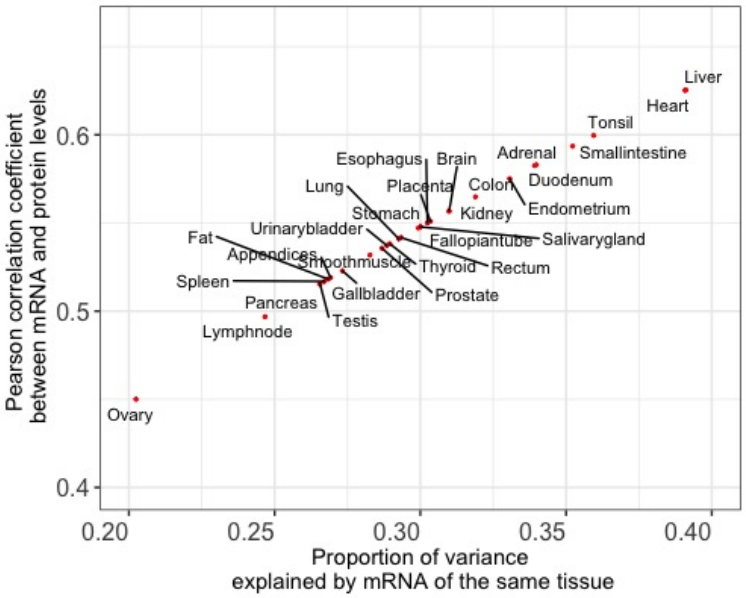


Figure 3.3: Per tissue mRNA - protein relationships Proportion of the variance in protein levels explained by mRNA levels of the same tissue (x-axis) versus the pearson correlation coeffcients between tissue specific mRNA and protein levels (y-axis) for 29 human tissues.

We observed that for certain tissues the mRNA levels of other tissues were also predictive of their mRNA levels (Figure 3.4). Protein levels of tissues which express reletively higher tissue-specific genes, such as Brain, Lymphnode, Testis and Thyroid [2], were only significantly predictive by their tissue specific mRNA measurements. On the contrary, the protein levels of other tissues, such as Colon, Duodenum, Endometrium, Esophagus, Lung, Rectum and Urinary bladder were also predicted by the mRNA levels of other

tissues to a certain extend. This observation may be attributed to the common cell types and active cellular programs in these tissues.

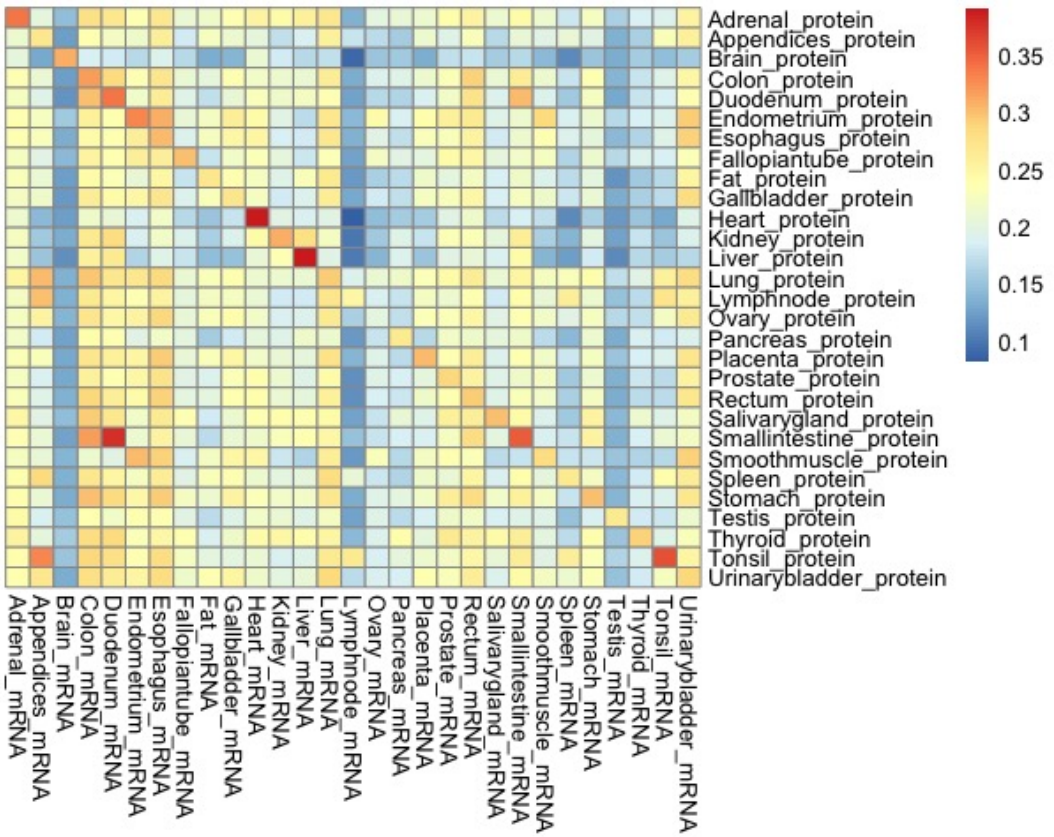


Figure 3.4: Explained variance in protein levels by the mRNA levels of other tissues. Heatmap displaying the explained variance (R^2) in protein levels (rows) by the mRNA levels (columns) of the 29 tissues.

In line with this observation, we saw that much larger proportions of the variance in protein levels could be explained by using mRNA profiles across all tissues (between 41% for pancreas and 56% for liver, $P < 10^{-132}$ for each tissue) (Figure 3.5). The reasons for this increase in explained variance are at least two fold. Biologically, as mentioned above, the mRNA levels of common cell types and cellular programs active in different tissues are predictive of the protein levels in other tissues. Technically, this increase may also be driven by the more robust nature of mRNA profiles across all tissues compared to the mRNA level measures in a single tissue. This is consistent with observations by Csárdi et al [39] that de-noising of mRNA measurements of budding yeast can enhance the explained variance of protein levels. In that regard, we emphasize that this analysis would benefit significantly if we had more replicates of matched tissue specific transcriptome and proteome measurements.

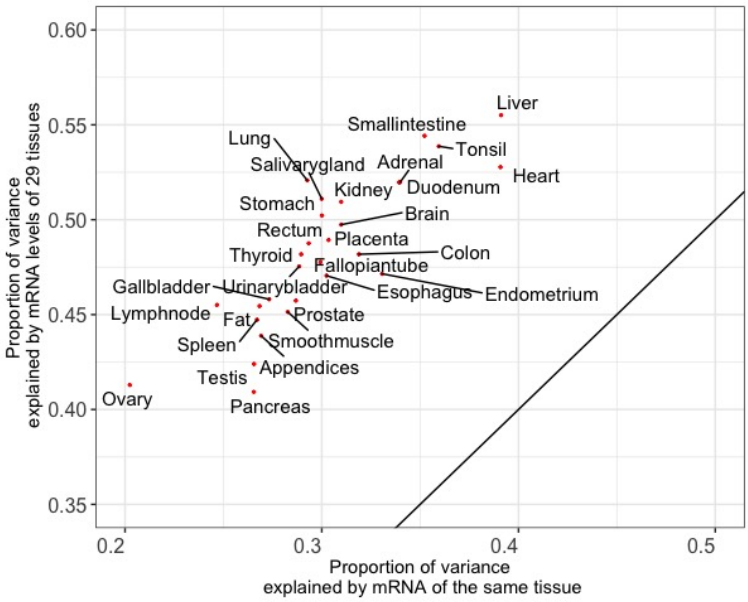


Figure 3.5: Explained variance in protein levels by own versus all tissues' mRNA levels. Proportion of the variance in protein levels explained by mRNA levels of the same tissue (x-axis) versus the explained variance by all 29 tissues' mRNA level measurements (y-axis).

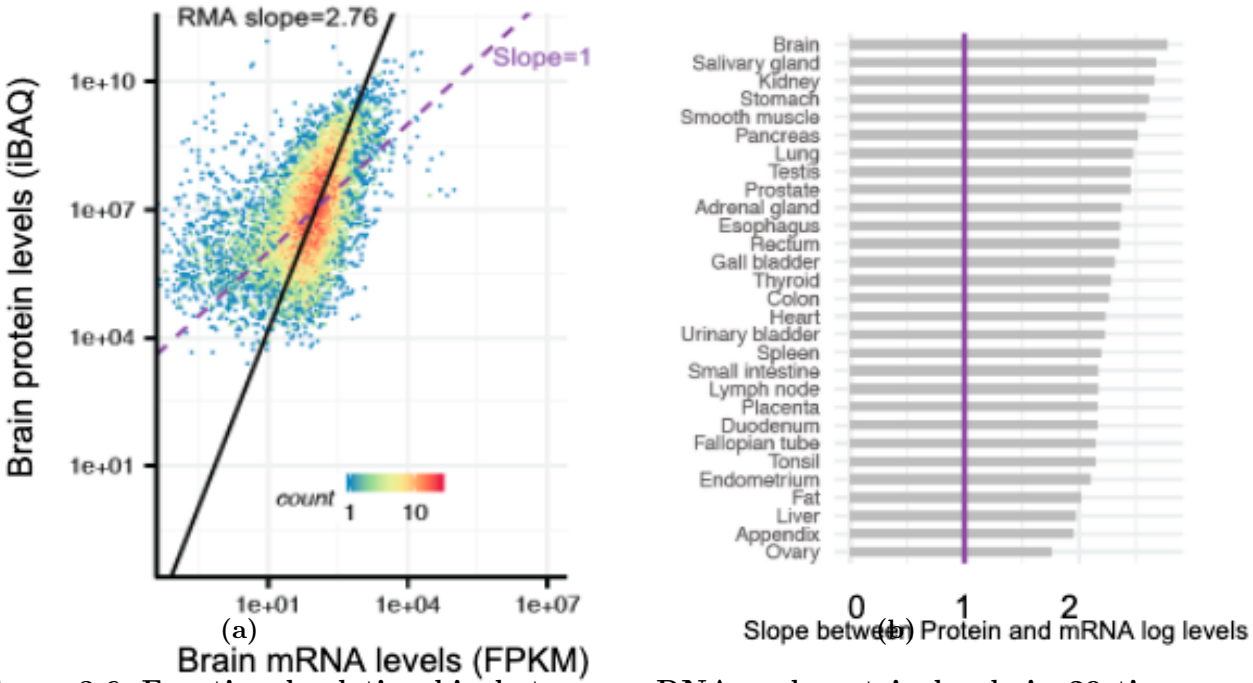


Figure 3.6: Functional relationship between mRNA and protein levels in 29 tissues. a) Scatterplot displaying the mRNA (x-axis) and protein (y-axis) gene measurement values in Brain tissue. Black line displays the ranged major-axis (RMA) regression fit with a slope of 2.76. **b)**The slopes of the tissue specific RMA regression fits between protein and mRNA levels for 29 tissues.

In addition to the explained variance in protein levels by mRNA levels, one can also question the functional relationship between these two phenotypes. According to the model shown in Equations 3.1 and 3.2, the ratio of the protein level to the mRNA level of gene i should be proportional to $\frac{k_{ri}}{k_{pi}}$ at the steady-state. If this fraction were equal for all genes, we would observe a linear relationship between mRNA and protein levels. However, we observed that there is a superlinear relationship between them in all 29 tissues (Figure 3.6) we analyzed. The approximately quadratic relationship between protein and mRNA levels across genes [2] results in a larger dynamic range of expression among proteins than mRNAs: dynamic range of mRNA levels is about 125 folds while the dynamic range of protein levels is about 794 folds (Figure 3.7).

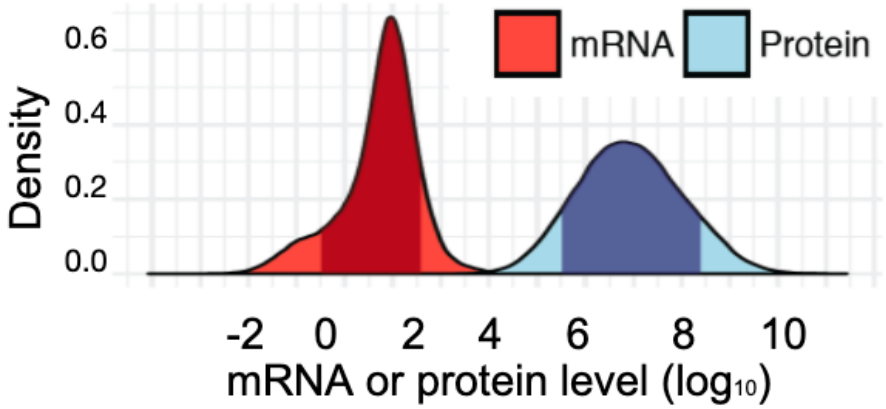
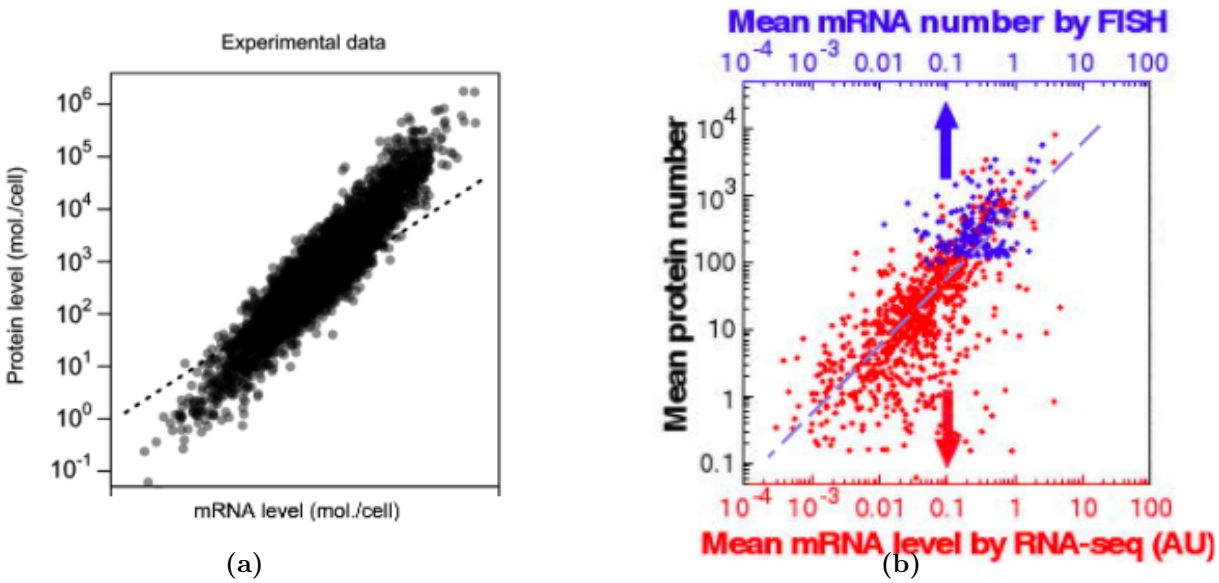


Figure 3.7: Density histograms of mRNA and protein levels of 29 tissues. Almost quadratic functional relationship between the mRNA and the protein levels results in a dynamic range difference between them. Darker red and blue regions display the 25%-75% quantile intervals.

This superlinear relationship holds true for other eukaryotes, such as yeast Saccharomyces cerevisiae (Figure 3.8-a) [39], while it is absent in prokaryotes (Figure 3.8-b) [84] for which the transcription and translation are coupled [85, 86]. In contrast to eukaryotes, there is no nucleus in prokaryotes that separates the transcription and translation process. Therefore, prokaryotic transcription and translation occur simultaneously in the cytoplasm. This reduces the effect of the translational regulation in bacteria, explaining the linear relationship between mRNA and protein levels.



(a) (b)

Figure 3.8: Difference in functional relationship between mRNA and protein levels in eukaryotes and prokaryotes. a) Experimental values of mRNA and protein levels in Saccharomyces cerevisiae. Dashed line displays the line with slope equal to 1. Figure is taken from [39]. **b)** Experimental values of mRNA and protein levels in Escherichia coli. Dashed line displays the line with slope equal to 1. Figure is taken from [84].

The superlinear functional relationship in eukaryotes is attributed to high-expression genes showing signs of more efficient translation [87, 11, 82], validated by measurements of translational activity by the ribosome profiling experiments. In these studies increased density of ribosomes on highly expressed mRNAs suggests increased rates of translation initiation as the major contributor [41, 88]. Current views attribute the positive correlation between translation rates and mRNA levels to the natural selection; that is, highly expressed genes have evolved to have sequences favoring higher translation rates [39, 82]. For example the comparisons of mRNA secondary structures of genes with different expression levels have shown that, the stability of mRNA structures in the 5' region weakens as mRNA expression level increases, favoring more efficient translation initiation [88]. Nevertheless, further studies should be designed to investigate whether there are also mechanistic factors in the cell that enable exponential amplification of protein levels when mRNA levels of the genes increase. For example, one can hypothesize that mRNA concentration and localization of the mRNAs of the highly-expressed genes would be effecting the ribosome recycling rates of these transcripts, thereby increasing the translation initiation rates. Targeted expresiments should be designed to question and refute such possible scenarios.

3.2 mRNA - protein level variations across tissues

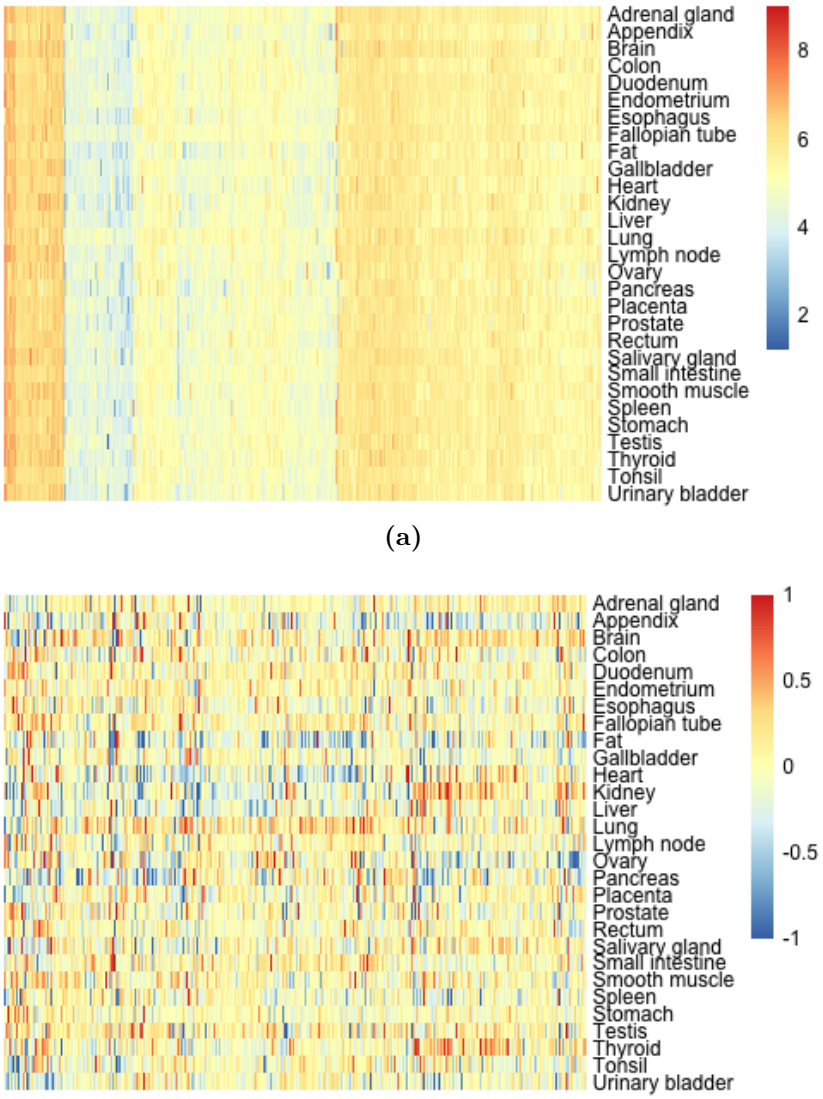


Figure 3.9: Difference between the dynamic range of PTR ratios across genes and across tissues. a) PTR ratios ($\log_{10}$) of the 4,506 genes that are measured at the transcriptome level and at the proteome level in 29 tissues. **b)** PTR ratio fold-change ($\log_{10}$) per gene across tissues. Values shown in (A) are centered per gene based on their median values across tissues.

Variation of the PTR ratio per gene across different tissues is more relevant for understanding the tissue-specific post-transcriptional regulation of protein expression than the variation between different genes of a single tissue. Our analysis shows that the variation of the PTR ratio of single genes across tissues was small in comparison with the variation of PTR ratios across different genes (Figure 3.9).

In order to understand how much of the across tissue protein variation of single genes could be predicted by the across tissue variation of their corresponding mRNA levels, we utilized latent space arithmetics between the latent space of the mRNA levels and protein levels of 29 tissues as follows:

For a $t \times g$ matrix $\mathbf{M}$ of mRNA level measurements of g genes (11,575) in t tissues (29), the SVD decomposition of the $\mathbf{M}^\intercal\mathbf{M}$ covariance matrix provides us $min(t, p)$ number of eigenvectors and eigenvalues which aligns with the $min(t, p)$ major directions of variation of g genes across t tissues (Eqn 3.3). Let $\mathbf{L}$ be the eigenvector matrix of $\mathbf{M}^\intercal\mathbf{M}$ and $\mathbf{P}$ be the $t \times g$ matrix of protein levels. When $g > t$, as it is in our case, $\mathbf{M}$ and $\mathbf{P}$ are high dimensional, low rank matrices and $\mathbf{L}$ is a $g \times t$ matrix of t dimensions. Accordingly, $\mathbf{P} \cdot \mathbf{L}$ is the projected protein levels in the t dimensional latent space of the mRNA levels. Therefore, by reconstructing back the protein matrix by mutiplying the projected data points $\mathbf{P} \cdot \mathbf{L}$ with $\mathbf{L}^\intercal$, we get the protein levels predicted by the corresponding mRNA levels (Eqn 3.4).

$$\mathbf{M}^\intercal\mathbf{M} \cdot \mathbf{L} = \mathbf{M}^\intercal\mathbf{M} \cdot \Lambda \tag{3.3}$$

$$\mathbf{P}_{\mathbf{pred}} = \mathbf{P} \cdot \mathbf{L} \cdot \mathbf{L}^\intercal \tag{3.4}$$

We observe that R-squared (R^2) value based on this prediction is equal to 0.24 (Figure 3.10) meaning that about 24% of the protein level variance across tissues is explained by the variance of the corresponding mRNA levels. Alternatively, our application of Multi-Omics Factor Analysis [89] showed that the latent factors explaining 60% of the across-tissue variance of mRNA levels were only able to explain 35% of the variance in PTR ratios (Figure 3.11 - a). Moreover, most of these latent factors were specific to either mRNA or PTR ratio level indicating that joint likelihood optimization failed to find significant factors that capture the shared covariation between mRNA and PTR ratio across tissues (Figure 3.11 - b). Together, these observations suggest that a substantial amount of the regulation of PTR ratios is independent of the mRNA level regulation, as it was also reported in the across tissue analysis done by Franks et al in 2017. ([32]).

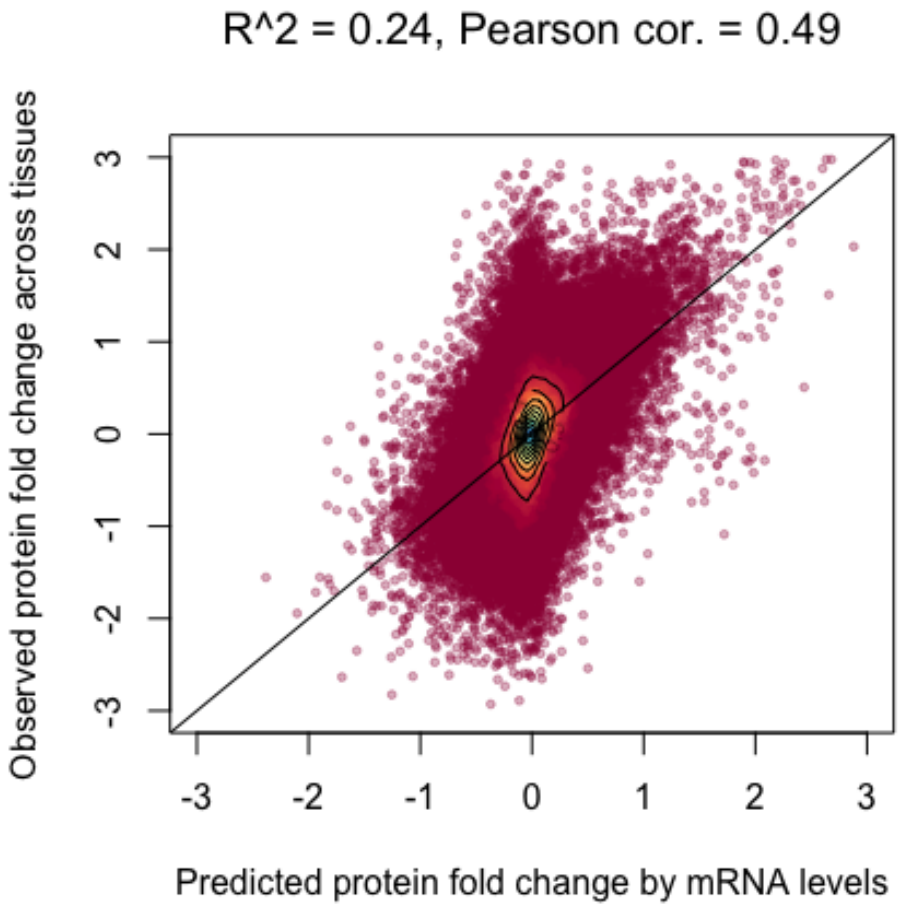
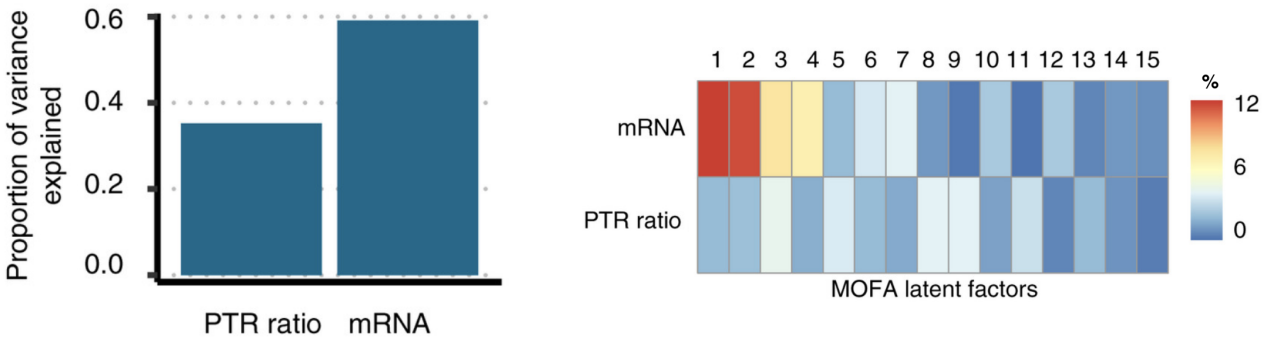


Figure 3.10: Predicted (x-axis) versus observed (y-axis) protein fold changes around the mean values across tissues.

Among the considered genes, housekeeping genes defined by the Human Protein Atlas, which are abundantly expressed in general, had fairly similar PTR ratios across tissues. Gene set enrichment analysis ($FDR < 0.1$) performed with DAVID [90, 91] revealed that cellular protein complex assembly, negative regulation of protein metabolic process, and regulation of cytoplasmic transport were some of the biological processes enriched for genes with low PTR ratio standard deviation. Also, proteins localized in certain cellular components such as chaperonin-containing T-complex, whole membrane, and cytoskeleton had significantly low PTR ratio standard deviation across tissues. In contrast, genes with strongly varying PTR ratios across tissues were enriched in biological processes that point toward tissue-specific and cell-specific biology and include cilium organization, glycolipid biosynthetic process, single–multicellular organism process, and inflammatory response and in cellular localizations that include extracellular space, intrinsic component of membrane, and secretory vesicles and granules.



(a) (b)

Figure 3.11: Explained variance in mRNA levels and PTR ratios by the common latent factors obtained by Multi-Omics Factor Analysis. a) Proportion of variance across tissues of PTR ratio (left) and mRNA (right) explained by the 15 latent factors fitted by joint optimization of the likelihood of both data modalities [89]. **b)** Explained variance by each latent factor. Factors that are active in both mRNA and PTR ratio capture shared covariation across tissues, and factors that are active in only one capture the signal specific to that modality.

4 Sequence determinants of protein-per-mRNA amount in 29 human tissues

The methodology, results and figures presented in this section are part of the manuscript "Quantification and discovery of sequence determinants of protein per mRNA amount in 29 human tissues" from Eraslan and Wang et al. 2019 [1] and "A deep proteome and transcriptome abundance atlas of 29 healthy human tissues" from Wang and Eraslan et al. 2019 [2]

4.1 Integrative analyses of multi-omics data to identify the sequence determinants of protein-to-mRNA ratio

Multiple post-transcriptional stages, including mRNA processing, nuclear export and localization, mRNA stability, and translation of mature mRNA molecules, protein satibility and secretion regulate the steady state mRNA and protein levels. A diverse set of mechanisms operating at the translation initiation, as well as during elongation and termination and even after termination, regulate the translation itself.

To identify and quantify sequence determinants of protein-to-mRNA ratio, we derived a model predicting tissue-specific PTR ratios from mRNA and protein sequence alone. The model is a multivariate linear model that includes a comprehensive set of mRNA-encoded and protein-encoded sequence features known to modulate translation initiation, elongation, and termination, as well as protein stability. In our model, we considered known post-transcriptional regulatory elements and identified novel candidates in the 5' UTR, coding sequence, and 3' UTR, by means of systematic association testing.

To interpret our findings related to different layers of gene expression regulation; that is, mRNA degradation [92], translation, and protein degradation, we included mRNA half-life measurements [93, 94, 95], in addition to human ribosome profiling of 17 independent studies [96, 23] as well as protein half-life measurements from immortal and primary cell lines [97, 98] (Figure 4.1).

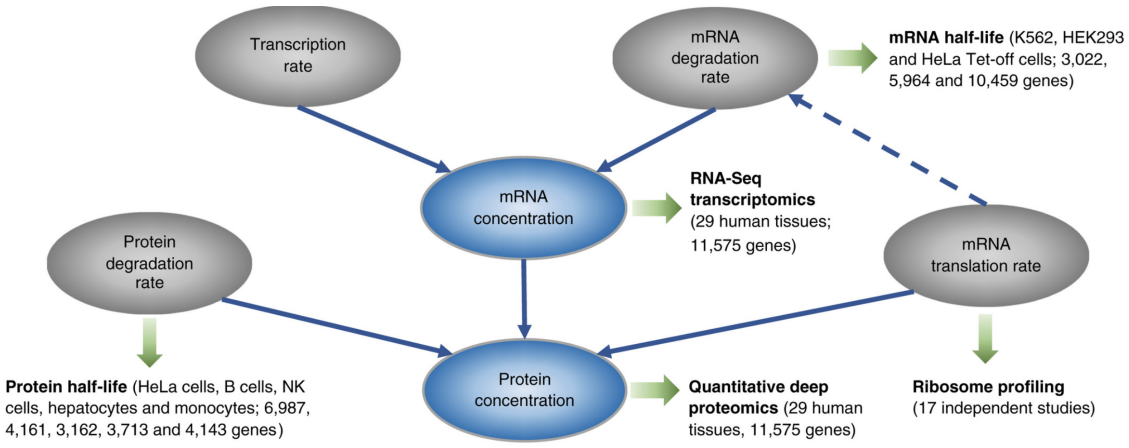


Figure 4.1: Integrated datasets to interpret our findings related to different layers of gene expression regulation Overview of the datasets analyzed in this study. We analyzed the protein-to-mRNA ratios by considering a dataset of matched proteome and transcriptome of 29 human tissues [2]. We further interpreted our findings with respect to ribosome occupancy datasets, reflecting translation elongation, protein half-life datasets, and mRNA half-life datasets. Solid lines represent the dependencies in the basic gene expression kinetic model. Dashed line represents the coupling between mRNA elongation and degradation rates [92].

In the below sections the results of the statistical analyses performed to decipher the associations between the regulatory sequence elements present in the 5' UTR region, coding region, 3' UTR region and the PTR ratio are presented. Details of the feature extraction methods and the applied statistical and bioinformatics analyses are explained in detail in Chapter 6.

4.1.1 Sequence features in the 5' Untranslated Region

Translation initiation regulation is mainly mediated via different sequence elements present in the 5' UTR and upstream coding regions of the transcripts; secondary structures can impede the detection of AUG initiation codons due to a blockage of the scanning ribosome, internal ribosome entry sites (IRESs) can stimulate cap-independent translation, binding sites of RNA binding proteins that either repress or facilitate translation, non-AUG initiation codons, the start codon sequence context that affects efficiency of AUG recognition, and upstream AUG codons which are sometimes followed by an in frame termination codon located upstream or downstream of the canonical start codon, thereby forming upstream open reading frames.

4.1.1.1 mRNA secondary structures

mRNA secondary structures affect the translation initiation, elongation and ribosome recycling rates in various ways [99, 100, 101]. Those structures occuring in the 5' UTR and

upstream of coding region are especially effective in regulating the translation initiation rates [101].

We tested the assocciation of secondary structures around canonical start codon region and upstream 5' UTR regions with PTR ratio by computing the negative minimum RNA folding energy, a computational proxy for RNA secondary structure, in 51-nt sliding windows in the [-100, +100] bp region around the start codon. We observed that these negative folding energy values associated with a lower PTR ratio around the start codon (Figure 4.2, up to 9% decrease, $FDR < 0.1$). This observation was in line with the mechanistic studies in E.Coli which had shown that secondary structures around the start codon impair translation by sterically interfering with the recruitment of the large ribosome subunit [41].

In contrast, negative minimum folding energy in 51-nt windows associated positively with the PTR ratio about 48 nt downstream of the start codon (Figure 4.2, up to 7% increase, $FDR < 0.1$) This positive association is consistent with experiments showing that hairpins located downstream of the start codon facilitate start codon recognition of eukaryotic ribosomes in vitro [102], presumably by providing more time for the large ribosome subunit to be assembled.

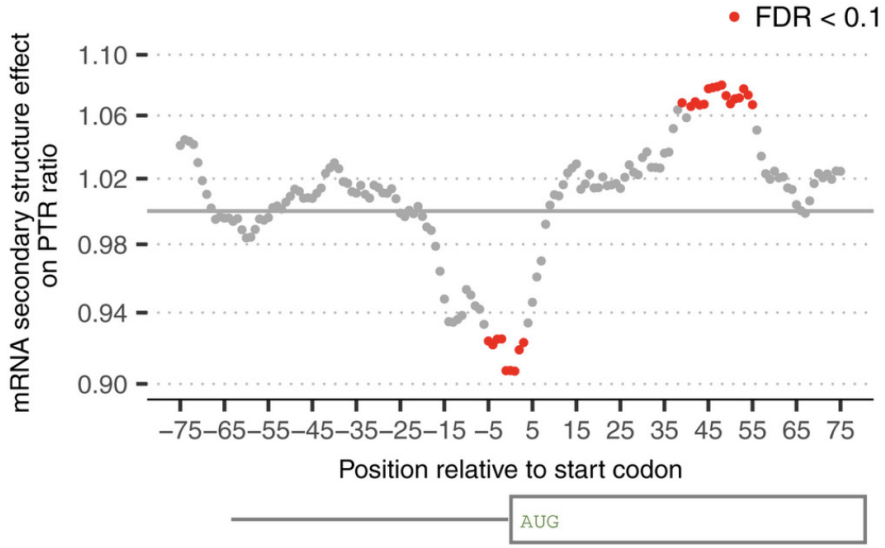


Figure 4.2: Effects of mRNA secondary structures around the canonical start codon on the PTR ratio. Effect of log2 negative minimum folding energy of 51-nt window on median log10 PTR ratio across tissues corrected for all other sequence features that are included in the model (y-axis) versus position of the window center relative to the first nucleotide of the canonical start codon (x-axis) for genes with a 5' UTR and a coding sequence longer than 100 nt. Statistically significant effects at $P < 0.05$ according to Student's t-test and corrected by the Benjamini–Hochberg methods are marked in red.

4.1.1.2 Upstream AUG codons and open reading frames

In eukaryotic mRNAs, one or more start codons may precede the start codon of the main coding region. When these upstream start codons (uAUGs) are preceded with an in-frame stop codon at the upstream or downstream of the canonical start codon, upstream open reading frames (uORFs) are formed. uAUGs and uORFs are observed to regulate the translation initiation and thus the protein expression levels through different mechanisms (reviewed in [103]).

Upon our search of the k-mers that were significantly associated with median PTR ratio across tissues 4.2, we de-novo identified AUG, the upsream start codon, for which at least one occurrence out-of-frame relative to the main ORF associated with about 18–33% lower median PTR ratios across tissues (Figure 4.3). This observation is consistent with previous reports that out-of-frame AUGs in the 5' UTR ([40]) and upstream ORFs ([104, 105, 103]) associate with lower protein-per-mRNA amounts.

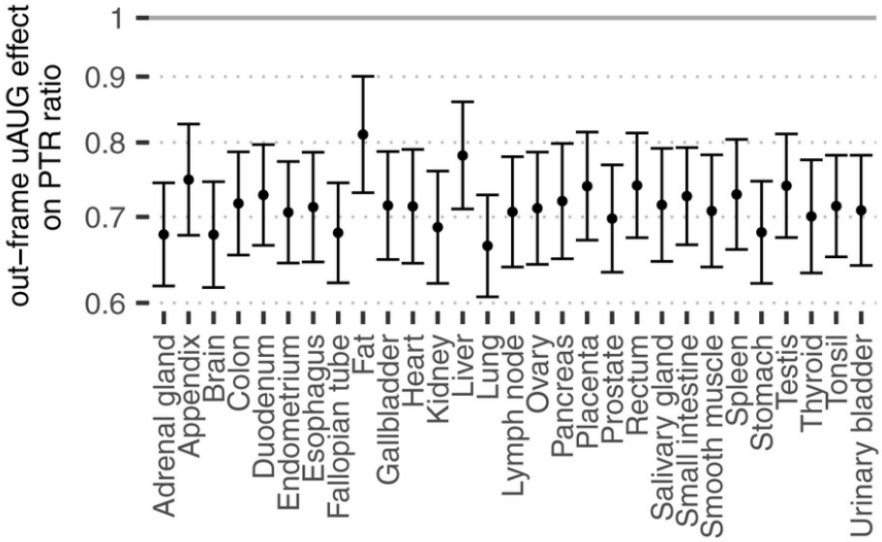
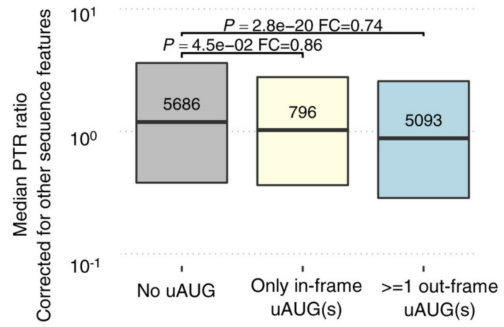
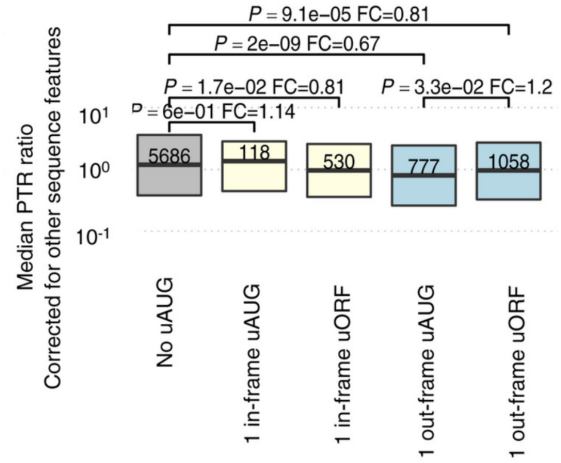


Figure 4.3: Effects of uAUGs. Effect estimate (dot) and 95% confidence interval (bar) of the presence of at least one out-of-frame AUG in 5' UTR on log10 PTR ratio corrected for all other sequence features listed in (A) (y-axis) per tissue (x-axis).

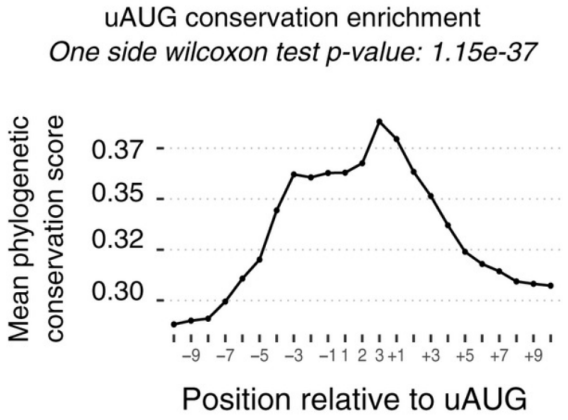
No significant associations could be found for the 796 transcripts with only in-frame uAUGs (Figure 4.4-a). Among 2,483 transcripts with a single uAUG or uORF, a single out-of-frame uAUG is associated with a 20% reduced PTR ratio compared to a single out-of-frame uORF (Figure 4.4-b), possibly because ribosomes can re-initiate translation downstream with high efficiency after translating a uORF ([104]). These uAUGs are significantly conserved (one-sided Wilcoxon test, $P = 1 \times 10^{-37}$) compared to background flanking regions according to the PhastCons score [106] computed across 100 vertebrates (Figure 4.4-c), which is consistent with earlier conservation analyses of AUG triplets in mammalian and yeast 5' UTRs [107].



(a)



(b)



(c)

Figure 4.4: Upstream AUGs and ORFs. a) Transcripts having at least one out-of-frame uAUG have significantly smaller PTR ratios across 29 tissues (corrected for other sequence elements) compared to transcripts with either no uAUG or only with in-frame uAUG(s) (Wilcoxon test, $fold - change = 0.75$, $P = 6.8 \times 10^{-17}$). Shown are the quartiles (boxes and horizontal lines). **b)**Transcripts with only one out-of-frame uAUG which is not followed by an in-frame stop codon are associated with 22% smaller PTR ratios (corrected for other sequence elements) compared to transcripts with only one out-of-frame uORF (Wilcoxon test, $fold - change = 0.78$, $P = 3.9 \times 10^{-2}$). Shown are the quartiles (boxes and horizontal lines). **c)** Average 100-vertebrate PhastCons score (y-axis) per position relative to the uAUG instances in 5' UTR (x-axis). P-values assess significance of the average 100-vertebrate PhastCons scores at the motif sites compared to the two 10-nucleotide flanking regions.

4.1.1.3 Canonical start codon context

Significant associations of individual nucleotides with the PTR ratio were detected in the 12 nt interval around the canonocal start codon ($FDR < 0.1$). At nearly every position of the start codon context, the nucleotide of the consensus sequence gccRccAUGG [42] showed the strongest association, indicating selection for efficient start codon recognition. The strongest effects were found at the third position upstream of the start codon (27% lower PTR ratio for C than for the consensus A), recapitulating mutagenesis data [42], and at the second nucleotide downstream of the start codon (23% lower PTR ratio for A than for the consensus C). Moreover, effects of the start codon context on the PTR ratio were largely independent of the tissue (Figure 4.5) consistent with a ubiquitous role of the start codon context likely due to structural interaction with the ribosome [108].

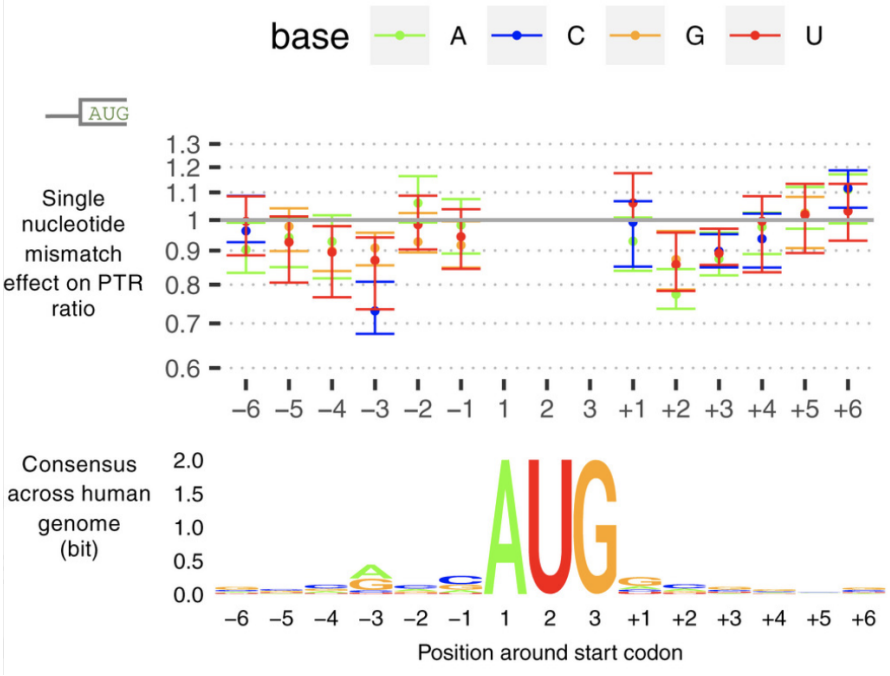


Figure 4.5: Effects of canonical start codon context on the PTR ratio. Median effect (dot) and range across 29 tissues (bar) of a single nucleotide mismatch relative to consensus sequence in a 12 nt window centered at first nucleotide of the canonical start codon (top). Position weight matrix logo showing information in bits (y-axis) computed across all 11,575 transcripts (bottom).

4.1.2 mRNA coding region sequence features

4.1.2.1 Codon usage

Codon usage frequency regulates protein function and PTR ratios in various ways. First, synonymous codon usage modulates translation efficiency [44, 109, 110, 111, 45], where preferred codons increase the rate of elongation, while non-optimal codons decrease the elongation rate. This translation rate in-turn affects co-translational protein folding,

and thus the protein function [109]. On the other hand, amino acid identity affects translation speed [112, 45] and protein half-life [113, 97].

Among all investigated sequence features, amino acid frequency had the largest predictive power for PTR ratio in every tissue (explained variance between 12 and 17%, median 15%) (Figure 4.6).

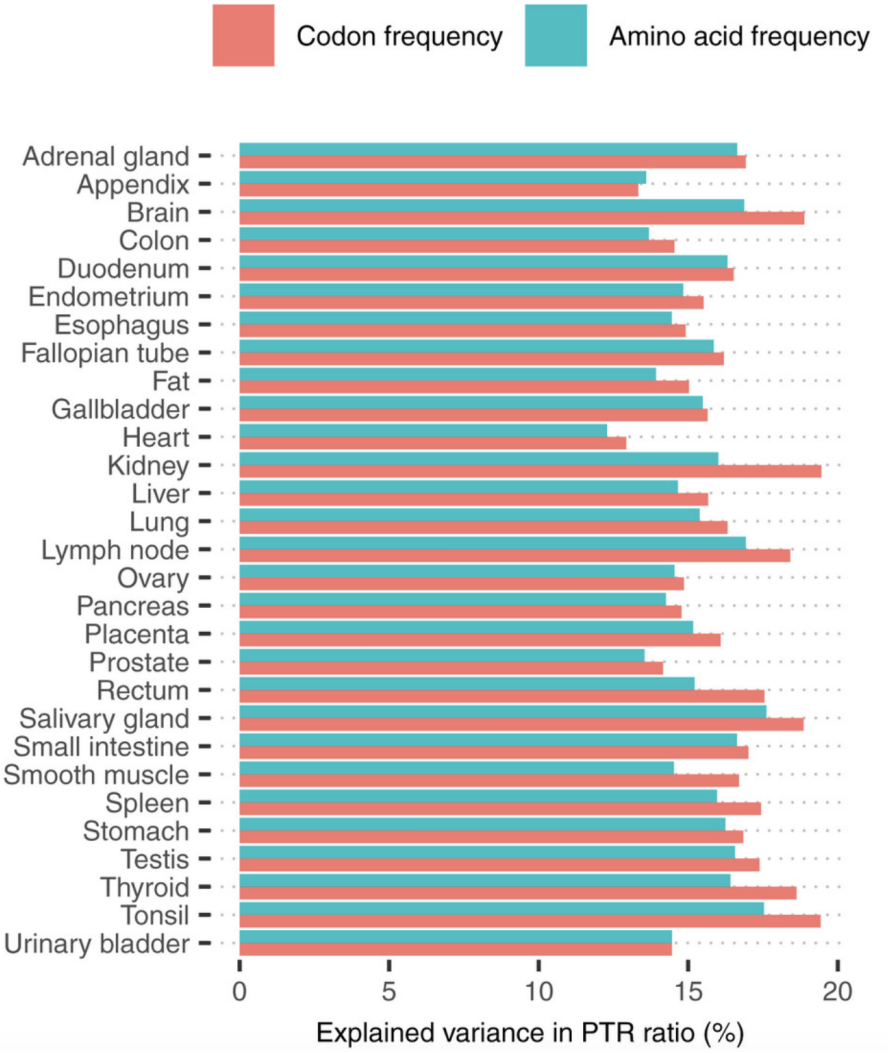


Figure 4.6: Comparison of the explained variance in PTR ratio by amino acid and codon usage. Log2-transformed frequencies of 20 amino acids in the coding region explain on average 15% of the variance in tissue-specific PTR ratios (min 12%, max 17%). In comparison, log2-transformed frequencies of 61 codons, which inherently encode for amino acid frequency and synonymous codon usage, explain on average 16% of the variance in PTR ratios (min 13%, max 20%).

In our linear model, we defined the amino acid effect on PTR ratio as the PTR ratio fold-change associated with doubling the frequency of an amino acid in a gene. The amino acid effects were large with a twofold increase in amino acid frequencies

associating with 40% lower PTR ratio for serine (S) and 50% higher PTR ratio for aspartic acid (D) (Figure 4.7 - a).

Codon frequency, which inherently encodes amino acid frequency and synonymous codon usage, increased that explained variance on average by only 1% (explained variance between 13 and 20%, median 16% , Figure 4.6). We defined the protein-to-mRNA ratio adaptation index (PTR-AI) as the PTR ratio fold-change associated with doubling the frequency of a codon in a gene. Synonymous codons coding for the same amino acids displayed different PTR-AIs (Figure 4.7 - b). Moreover, the PTR-AI of individual codons showed consistent amplitudes and directions across tissues (Figure 4.7 - b), which contests the hypothesis of widespread tissue-specific post-transcriptional regulation due to a varying tRNA pool among different tissues [114, 115].

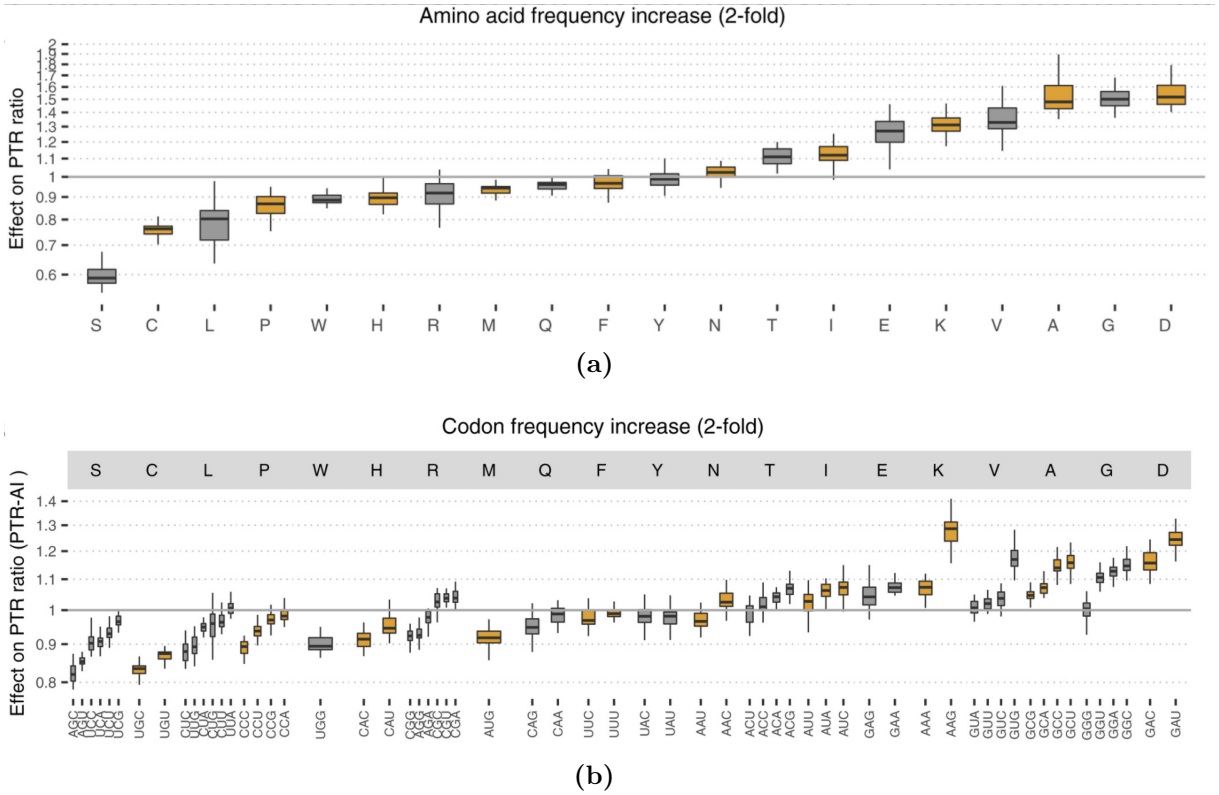


Figure 4.7: Distribution of the amino acid and codon usage effects on PTR ratio. a) Distribution of the amino effects on PTR ratio per tissue, which is the PTR ratio fold-change associated with doubling the frequency of the amino acid. Shown are the quartiles (boxes and horizontal lines) and furthest data points still within 1.5 times the interquartile range of the lower and upper quartiles (whiskers). **b)**Same as (A) for codons (PTR-AI). The codons are grouped by the amino acid they encode and are sorted first by increasing amino acid effect, then by increasing synonymous codon effect.

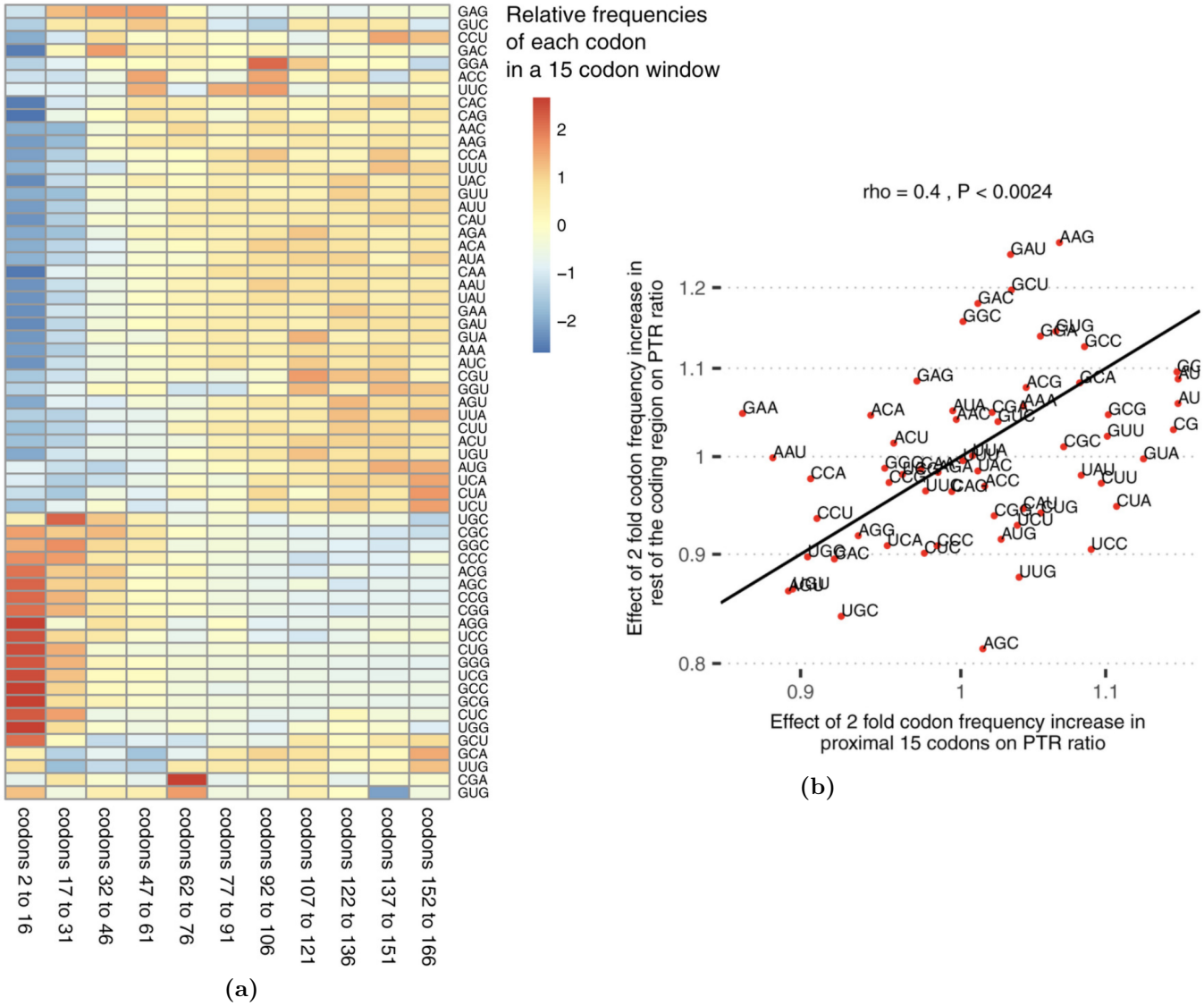
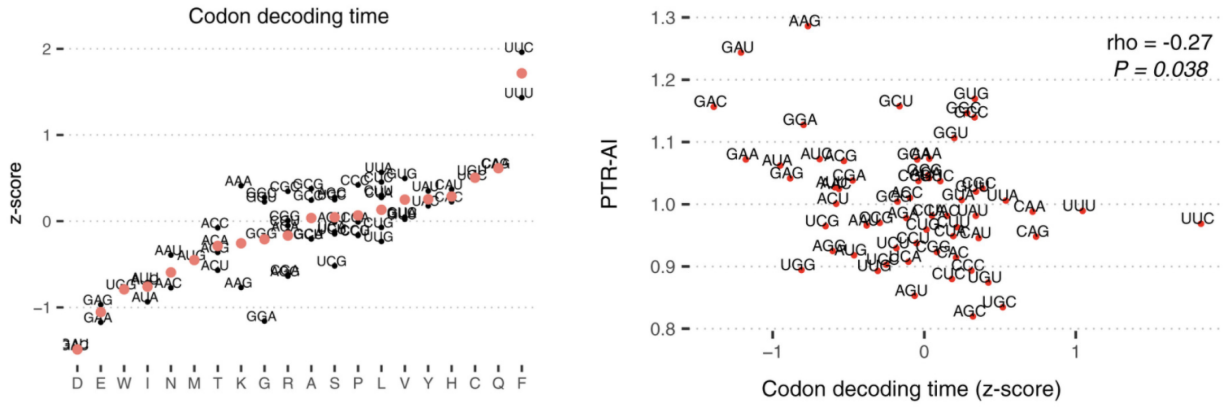


Figure 4.8: Codon frequency distribution at the upstream coding region. a) Relative codon frequencies per bin of 15 codons (columns). Codon frequency is differentially distributed in the 5' end of the CDS (from codon 2 to codon 31) compared to the rest of the coding region. **b)**In spite of the distinctive codon frequency composition of the coding region 5' end as displayed in (C), twofold codon frequency increase effect in the proximal 15 codons significantly correlates with the twofold codon frequency increase effect in the rest of the coding region ($Spearman's correlation = 0.4, P < 0.0024$).

We observed differences of codon frequency in the 5' end of the coding sequence compared to the rest of the coding region (Figure 4.8-a). However, when we modelled the codon usage of the upstream and downstream coding regions idependently, we observed that the effects of the codon usage on the PTR ratio correlated significantly between

these two regions (Figure 4.8-b). We therefore did not distinguish the 5' end region of the coding sequence from the rest of the coding sequence when considering codon frequencies in our model.

To relate the amino acid and synonymous codon effects to translation and protein degradation, both of which contribute to PTR ratios, we first investigated codon decoding times, whereby long decoding times would lead to lower translation output [116, 44, 22, 109, 110, 111, 45]. We considered codon decoding time as the typical time ribosome takes to decode a codon [117], also sometimes referred to as ribosome dwell time [23]. We computed median codon decoding times across 17 ribosome profiling datasets [96, 23]. Notably, amino acid identity explained 70% median codon decoding time variance (Figure 4.9-a, Appendix Figure A.1), consistent with the dominant role of amino acids on PTR ratio. The strong association between amino acid identity and codon decoding time may be in part reflecting that the amino acid content of the nascent polypeptide chain influences translation elongation [118]. PTR-AIs correlated significantly negatively with median codon decoding times ((Figure 4.9-b), Spearman's correlation $= -0.27, P = 0.03$).



(a) **(b)**

Figure 4.9: Codon usage effects on PTR ratio based on its effect on translation efficiency. a) Median codon decoding time (transformed to z-scores) across 17 independent ribosome profiling datasets (y-axis), grouped per amino acid (x-axis). Red dots display the average amino acid decoding time. **b)** Median codon decoding time estimates (z-scores) across 17 independent human Ribo-Seq datasets (x-axis) significantly negatively correlate with average PTR-AI across tissues (y-axis).

We also found that median PTR-AIs correlated significantly positively with predicted effects of codons on mRNA stability in K562 (Spearman's correlation $= 0.47, P = 4.7 \times 10^5$, [95]), in HEK293 (Spearman's correlation $= 0.48, P = 9 \times 10^5$, [94]), and in HeLa Tet-off cells (Spearman's correlation $= 0.52, P = 3 \times 10^5$, [93]) (Figure 4.10). This agreement of PTR-AIs and predicted effects of codons on mRNA stability is consistent with the fact that codon composition is causally affecting mRNA degradation [119, 120, 121, 122] in a way that is mediated by translation [92]. Together, these results indicate that PTR-AIs capture the effect of codons on translation.

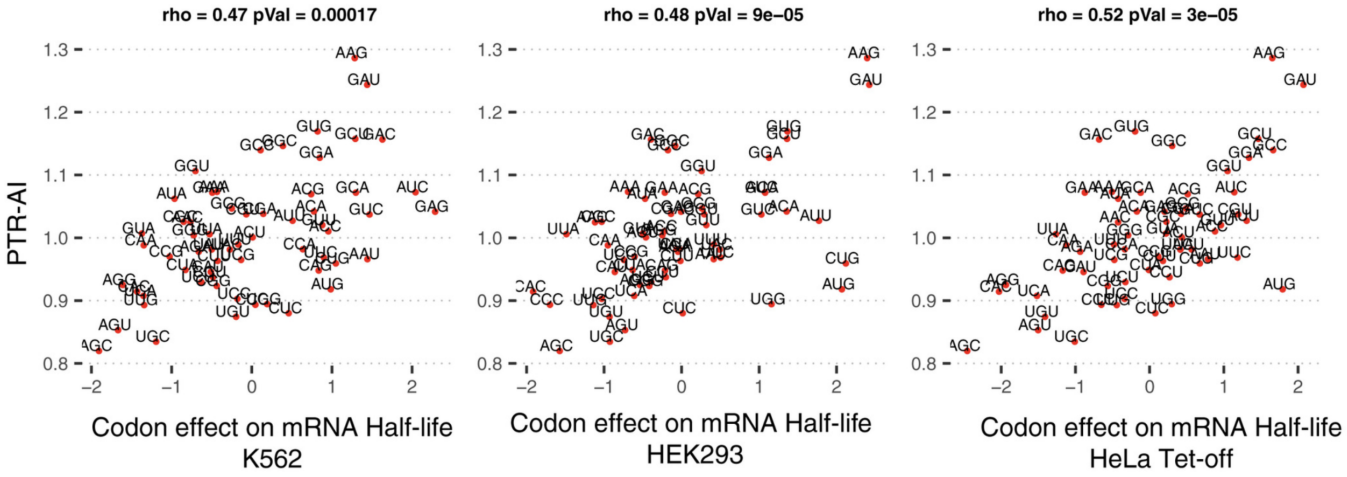


Figure 4.10: PTR-AI correlation with mRNA half-lives Median PTR-AI across tissues highly correlates with mRNA half-life fold-changes associated with twofold frequency increase of codons in K562 [95], HEK293 [94], and HeLa Tet-off cells [93].

We then asked whether our amino acid effects on PTR ratios captured the effects of amino acids on protein degradation. To this end, we first performed a linear regression of protein half-lives measured in HeLa cells [97], B cells, NK cells, hepatocytes, and monocytes [98] on amino acid frequency. We defined the amino acid effect on protein half-life as the protein half-life fold-change associated with doubling the frequency of an amino acid in a gene. The amino acid effects on protein half-life agreed well among these datasets (Figure 4.11-a)) with proportions of explained variance varying from 9% for monocytes to 19% for NK cells. Moreover, the amino acid effects on protein half-life significantly correlated with the effects of single amino acid substitutions on protein thermodynamic stability ([123], Figure 4.11-b); Spearman's Correlation $= 0.18, P = 0.002$) and with amino acid hydrophobicity values (Figure 4.11-c; Spearman's Correlation $= 0.42, P = 0.04$), a major force stabilizing the folding of proteins [124]. This suggests that the associations of amino acids with protein half-lives are in part functional and due to the role of amino acids on protein thermodynamic stability, a strong determinant of protein cytoplasmic degradation [125].

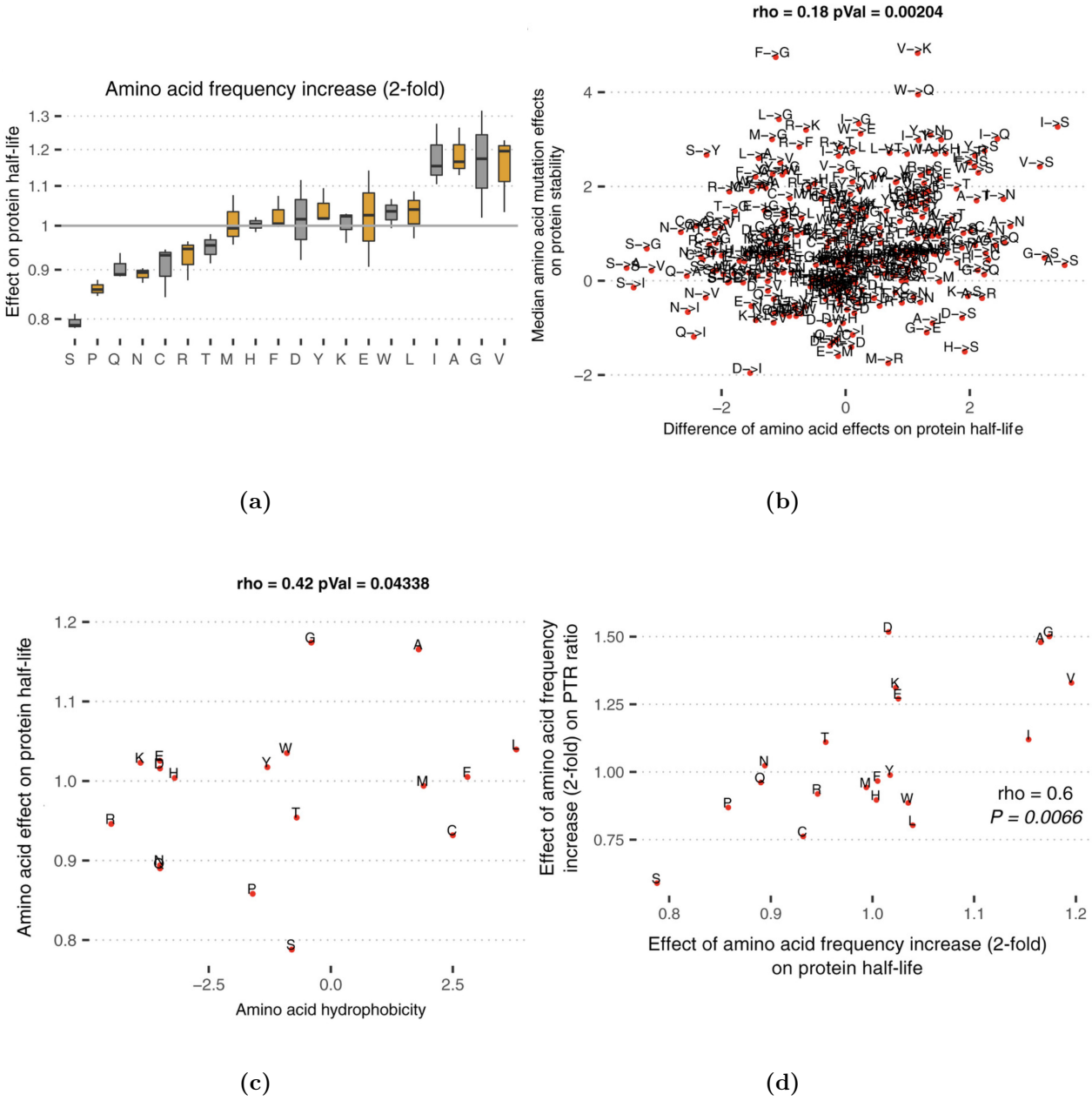


Figure 4.11: Codon usage effects on PTR ratio based on its effect on protein half-lives. **a)** Distribution of the amino acid effect on protein half-lives, which is the protein half-life fold-change associated with doubling the frequency of the amino acid, for five different cell types: HeLa cells [97], B cells, NK cells, hepatocytes, and monocytes [98]. Shown are the quartiles (boxes and horizontal lines) and furthest data points still within 1.5 times the interquartile range of the lower and upper quartiles (whiskers). **b)** Differences of amino acid effects on protein half-life (x-axis) significantly correlate with effects of single amino acid substitutions on protein thermodynamic stability (y-axis; citeDehouck2009). **c)** Amino acid effect on protein half-life significantly correlates with the amino acid hydrophobicity value. **d)** Amino acid effect on protein half-lives (x-axis) significantly positively correlates with amino acid effect on PTR ratio (y-axis).

Overall, the amino acid effects on PTR ratio correlated significantly with both the amino acid effects on protein half-life (Figure 4.11-d; Spearman's correlation $= 0.6, P = 0.006$) and the average amino acid decoding time (Figure 4.12 ; Spearman's correlation $= -0.41, P = 0.03$). However, average amino acid decoding times did not correlate significantly with the amino acid effects on protein half-life (Figure 4.12; Spearman's correlation $= -0.22, P = 0.35$). Analogous results were obtained by taking a codon-centric rather than an amino acid-centric point of view. Specifically, PTR-AI correlated significantly with codon effects on protein half-life (Spearman's correlation $= 0.56, P = 4.7e-06$) on the one hand, and with codon decoding time (Figure 4.12 ; Spearman's correlation $= -0.27, P = 0.04$) on the other hand. However, codon decoding time did not correlate significantly with codon effects on protein half-life (Spearman's correlation $= -0.09, P = 0.45$). Hence, PTR-AI appears to capture a combination of apparently independent effects of codon frequency on translation elongation and amino acid frequency on protein stability.

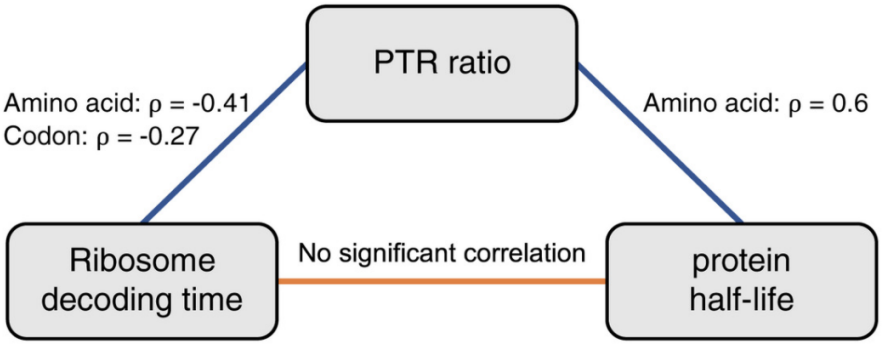
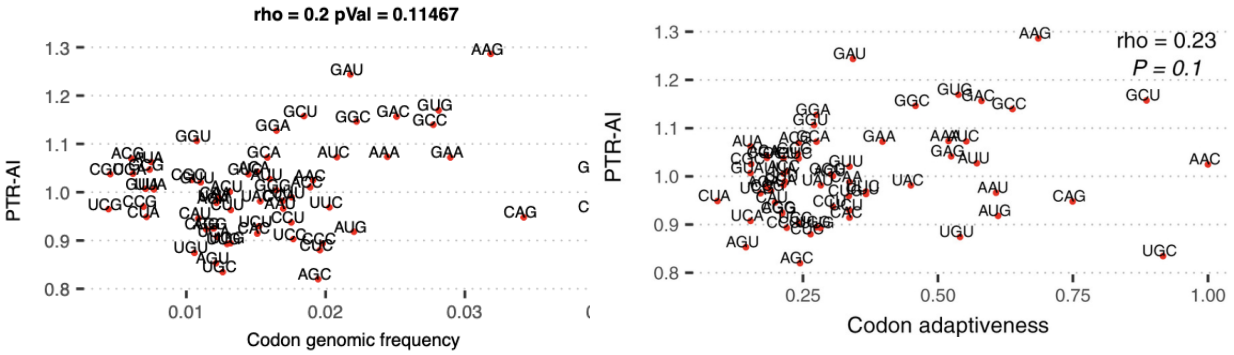


Figure 4.12: Codon adaptiveness based on its effects on translation efficiency and protein stability. Correlation network of the amino acid or codon frequency when applicable on PTR ratio, codon decoding time, and protein half-life. Significant Spearman correlations ($P < 0.05$) are found between the effects on PTR ratio and codon decoding time, and between the effects on PTR ratio and protein half-life but not between codon decoding time and protein half-life.

Notably, PTR-AI did not correlate well with previous codon optimality measures, including the frequency of codons in human coding sequences (Figure 4.13-a, Spearman's correlation $= 0.2, P = 0.11$) and species-specific codon absolute adaptiveness ([51]; 4.13-b, Spearman's correlation $= 0.23, P = 0.1$), which are based on genomic or transcriptomic data and strong modeling assumptions. Altogether, these results indicate that a PTR ratio-based measure of codon optimality, which captures the combined effects of protein production and degradation, is an attractive alternative to existing codon optimality measures and could help resolving some of the debates about the role of codon optimality in human cells.



(a) **(b)**

Figure 4.13: PTR-AI did not correlate well with previous codon optimality measures. a) PTR-AI (2 fold codon frequency increase effect on PTR ratio) does not significantly correlate with human genomic codon frequencies. **b)** Codon tRNA adaptiveness (x-axis), a widely used codon optimality metric, does not significantly correlate with PTR-AI (y-axis), which may be a new optimality metric reflecting the combined effect of amino acid and synonymous codon usage on protein synthesis and degradation.

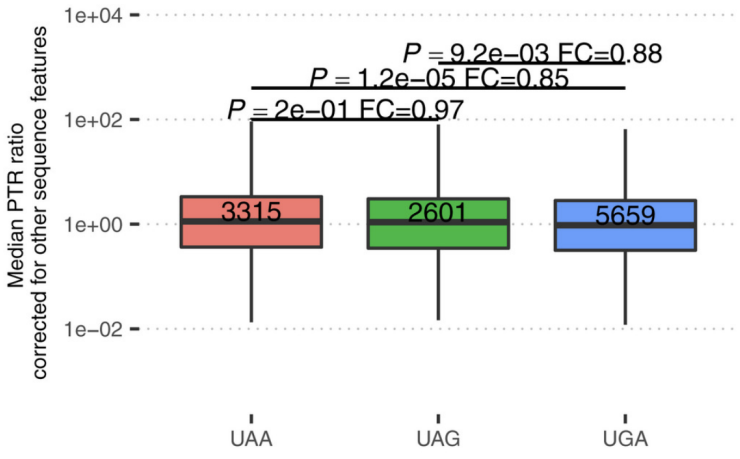
4.1.2.2 Stop codon context

Stop codon itself is only part of the translation termination signal and the context in which it resides has significant effects on the efficiency of the translation termination and the ribosome recycling rate [56].

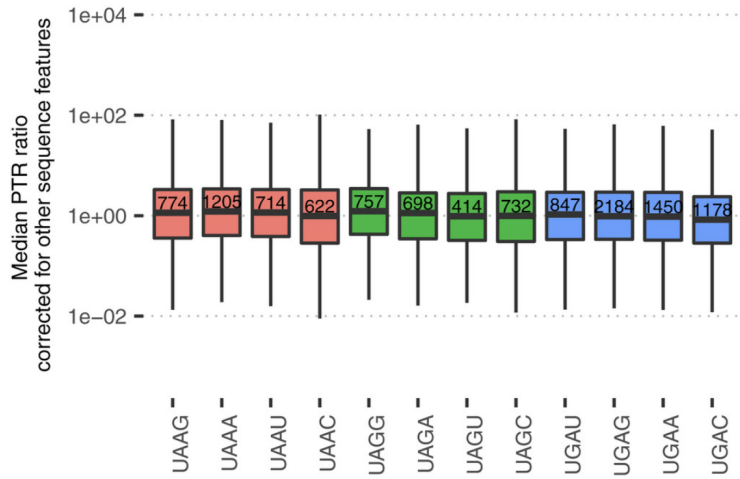
The opal stop codon UGA was significantly associated with the lowest median PTR ratio having in median 15% lower PTR ratios than the ocher stop codon UAA (Figure 4.14-a; P = 1.210^5).

Around the stop codon, the two most influential positions were the +1 nucleotide at which a C associated with 15% lower PTR ratios than the consensus G, and the −2 nucleotide, at which a G associated with 19% lower PTR ratios than the consensus A in median across tissues (Figure 4.15). The inhibitory effect of a C at the +1 nucleotide, which was observed for all three stop codons (Figure 4.14-b), is in line with previous studies in prokaryotes and eukaryotes [53, 54, 55, 56]. Also, structural data show that a C following the stop codon interferes with stop codon recognition [126], thereby leading to stop codon read-through. Moreover, our data indicate that the nucleotide at the −2 position, which is also reported to be highly biased in E. coli [127], is significantly associated with PTR ratio and deviation from the consensus nucleotide A is associated

with a reduced PTR ratio. Altogether, the start and stop codon contexts demonstrate the sensitivity of the PTR ratio analysis in detecting contributions to translation down to single-nucleotide resolution.



(a)



(b)

Figure 4.14: Comparison of stop codons. a) Transcripts with an opal (UGA) canonical stop codon have significantly smaller median PTR ratios across 29 tissues (corrected for other sequence elements) compared to transcripts with an ochre (UAA) or an amber (UAG) stop codon. Shown are the quartiles (boxes and horizontal lines), the furthest data points still within 1.5 times the interquartile range of the lower and upper quartiles (whiskers), P-values for two-sided Wilcoxon test (P), and fold-change (FC). **b)** Transcripts with a cytosine at the +1 position relative to the stop codon have significantly smaller median PTR ratios across 29 tissues (corrected for other sequence elements) independently of the stop codon type. Shown are the quartiles (boxes and horizontal lines) and furthest data points still within 1.5 times the interquartile range of the lower and upper quartiles (whiskers).

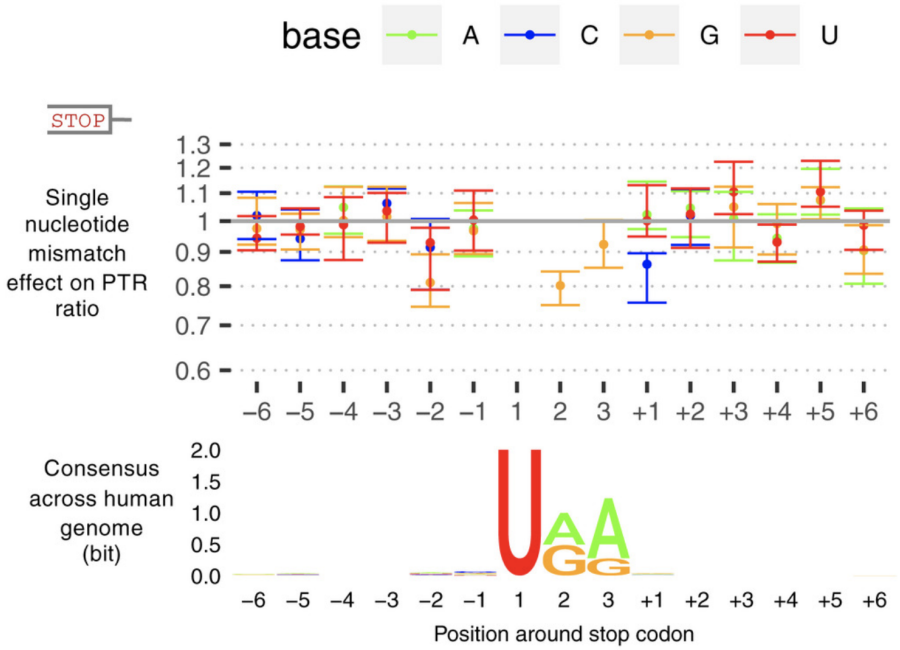


Figure 4.15: Effects of stop codon context on the PTR ratio. Median effect (dot) and range across 29 tissues (bar) of a single nucleotide mismatch relative to consensus sequence in a 12 nt window centered at first nucleotide of the canonical stop codon (top). Position weight matrix logo showing information in bits (y-axis) computed across all 11,575 transcripts (bottom).

4.1.3 mRNA 3' UTR sequence features

4.1.3.1 RNA binding proteins

RNA binding proteins (RBPs) are among the major factors controlling protein translation. They regulate post-transcriptional processes such as splicing, cleavage and polyadenylation, and the editing, localization, stability and translation of mRNAs [128]. Even though RBPs bind to not only to 3' UTR region but also 5' UTR and coding regions, the ones that bind to the 3' UTR region are mostly involved in controlling translation efficiency and mRNA half-life [12, 128]. This is why we include our results about the RBP analysis in this subsection.

Exploiting our comprehensive protein expression measurements across 29 human tissues, we investigated tissue-specific expression of RNA binding proteins. Overall, 1,233 out of 11,575 inspected genes were among the 1,542 RNA binding proteins manually curated by Gerstberger et al [60]. Of these, 825 RBPs were measured in all 29 tissues(Figure 4.16) . According to tissue specificity scores defined by Gerstberger et al, 135 out of 1,233 RBPs were defined as being tissue-specific based on our RNA-Seq dataset, which was consistent with the general observation that the majority of the RBPs are ubiquitously expressed and typically at higher levels than average cellular proteins [129, 60]. The 135 tissue-specific RBPs were significantly enriched in spermatogenesis, the multi-organism

reproductive process, DNA modification, and meiotic nuclear division and localized in germ plasm, pole plasm, and P granule ($FDR < 0.1$).

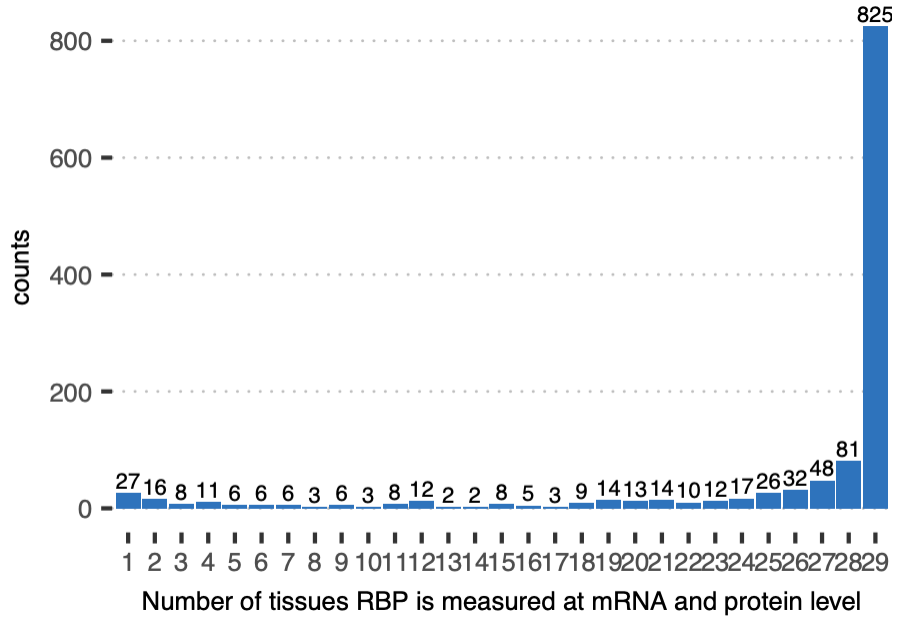


Figure 4.16: Number of tissues the 1,233 RNA binding proteins are measured Distribution of the number of tissues 1,233 RNA binding proteins are measured both at transcriptome and proteome level.

Disentangling the effects of each of the 1,542 RBPs on PTR ratios is challenging not only because the binding targets of these RBPs are poorly charted, but also because binding sites of RBPs and miRNAs often co-occur due to cooperative and competitive binding [130, 131, 132, 133]. Nevertheless, we included the binding evidence of our 11,575 genes to 112 RNA binding proteins (RBPs) by exploiting the enhanced CLIP (eCLIP) data set published by Von Nostrand et al [12] to observe the explained variance of the PTR ratios by these 112 RBPs. The proportion of variance in PTR ratio explained by the binding evidence to 112 RNA binding proteins [12] varied from 3 to 6% across tissues (median 5%) (chapter 6). Overall, these RBPs appeared to be ubiquitously expressed since 81 out of the 112 RBPs (77%) were detected expressed at the proteome level and at the mRNA level in all tissues. Ubiquitous expression of RBPs and the frequent co-binding of RBPs and miRNAs may be two reasons why tissue-specific effects of RBP binding on PTR ratio did not show significant correlations with the corresponding tissue-specific RBP expression levels (Appendix Figure A.2). Nevertheless, in 16 of these RBPs, there was a significant difference between their across-tissue covariation with their target and non-target genes (Figure 4.17).

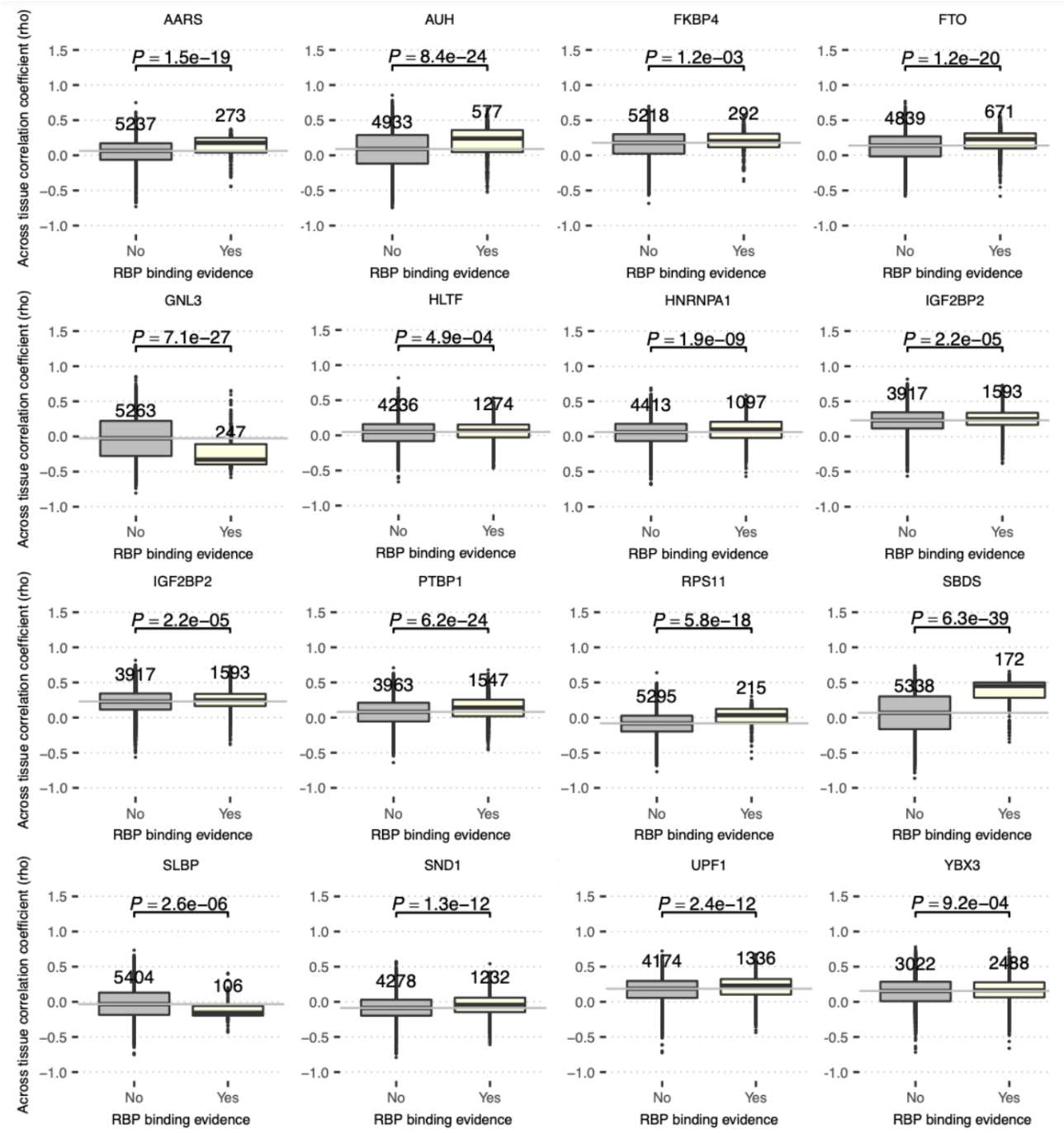


Figure 4.17: Expression levels of RBPs correlate with the expression levels of their target genes. Distribution of Spearman's rho between RBP protein level expressions and target and non-target genes' PTR ratios across tissues. Only RBPs which are expressed in at least 15 tissues and with protein level standard deviation greater than 0.1 are taken into account. Similarly for the calculation of correlation coefficients, only genes with which are expressed in at least 15 tissues and with PTR ratio standard deviation greater than 0.1 are considered.

4.1.3.2 microRNAs

MicroRNAs (miRNAs) are 23-nucleotide RNAs that favourably bind to 3' UTR sites in the mRNAs of protein-coding genes to downregulate the mRNA staibility and/or the protein translation efficiency [57]. Previous studies has shown that miRNAs reduce the protein output by mainly destabilizating the target mRNAs, and their effect on translational efficiency is moderate [59].

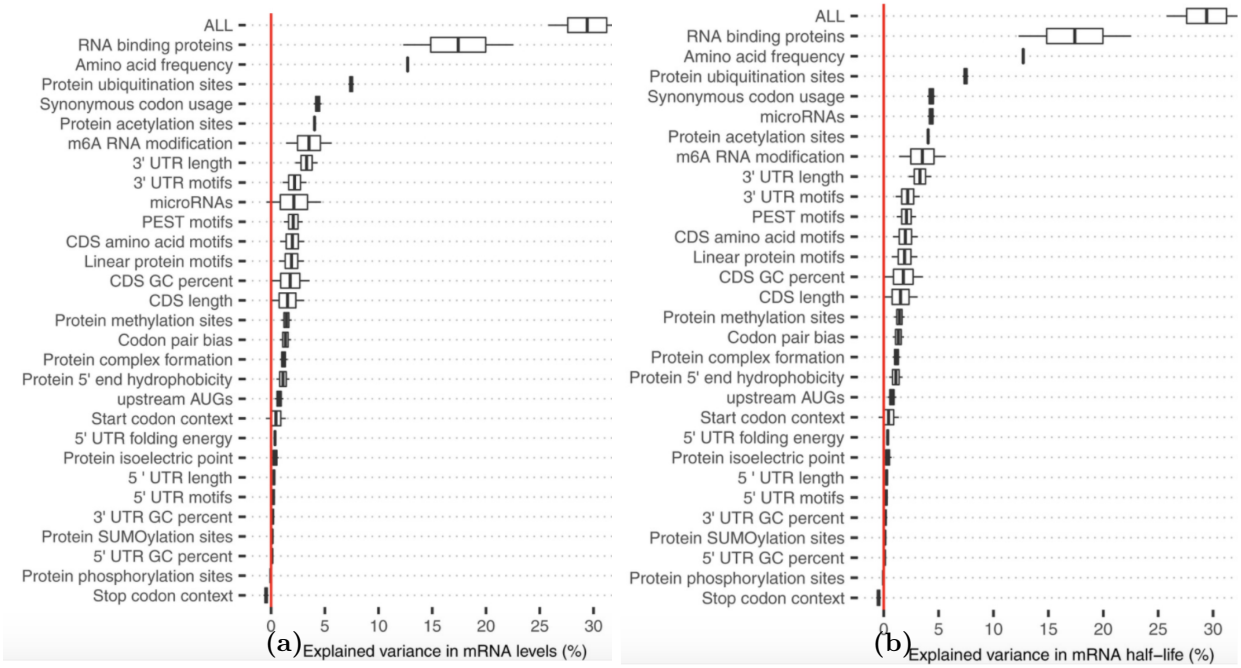


Figure 4.18: Explained variances in mRNA levels and mRNA half-lives. a) Distributions of explained variances by linear models of individual sequence feature groups in mRNA levels of 29 tissues. The distribution of the total explained variance by the linear model that combines all of the listed sequence features is displayed in the first line. Shown are the quartiles (boxes and vertical lines) and furthest data points still within 1.5 times the interquartile range of the lower and upper quartiles (whiskers). **b)** Same as (a), with the difference that the response variable is set to be Ensembl Transcript ID matched mRNA half-lives in K562 [95], HEK293 [94], and HeLa Tet-off cells [93].

Currently there are 2,599 catalogued human miRNAs [134], most of which display cooperative/competitive binding profiles with the RNA binding proteins. Therefore, estimating each of their individual effects on PTR ratios is a challenging goal. Nevertheless, in order to explore the degree to which the prediction of the PTR ratio from sequence could be improved in principle, we included the binding evidence for 296 miRNAs from the miRTarBase database with more than 200 targets in our dataset as features in our predictive model (chapter 6). 150 latent variables of these 296 miRNAs' binding evidence (chapter 6) explained on average only 1% of the PTR ratio variance across genes, while these 150 variables were able to explain on average 5% (min 4%, max 7%) of the variance in tissue-specific mRNA levels.

The binding evidence of RBPs and miRNAs were among the top mRNA features explaining the tissue-specific mRNA levels and mRNA half-lives of three different cell types (Figure 4.17-a,b). The binding of the considered 112 RBPs explained on average 18% (min 13%, max 21%) and features representing miRNA binding explained on average 5% (min 4%, max 7%) of the variance in tissue-specific mRNA levels (Figure 4.17-a). Consistent with that, RBP binding explained on average 17% of the variance (min 12%, max 22%) in mRNA half-lives of K562, HEK293, and HeLa Tet-off cells (Figure 4.17-b). Likewise, features representing miRNA binding explained 5% (median, min 4%, max 5%) of the variance in mRNA half-lives of these three cell lines. Altogether, the differences and similarities in the explained variances of mRNA levels, mRNA half-life, PTR ratio that the RBPs and miRNAs considered in our model may be more effective in regulating mRNA stability rather than PTR ratios.

4.1.4 Protein sequence features

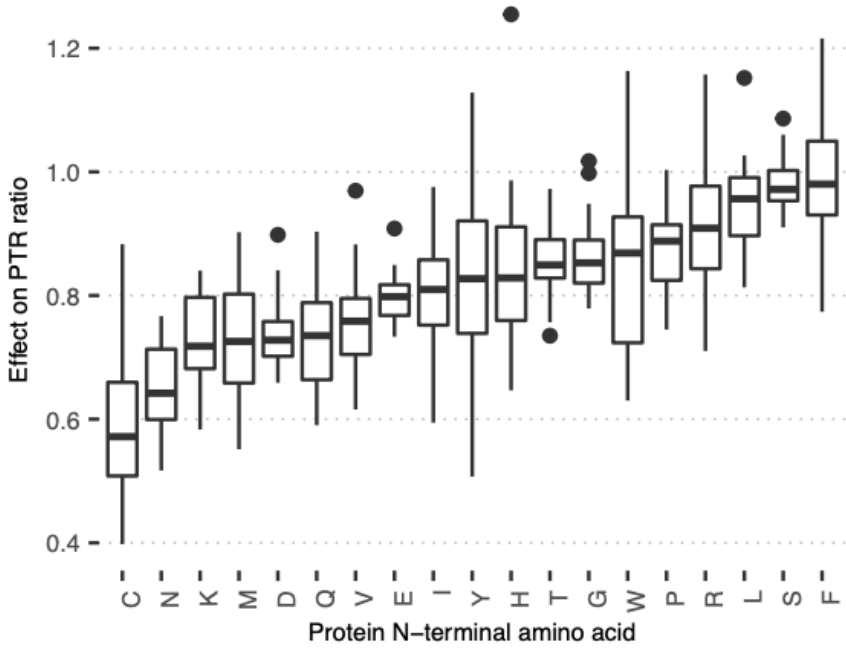


Figure 4.19: N-terminal residue effect on PTR ratio. Distribution of the effects of protein N-terminal residues (with respect to Alanine) on PTR ratio across tissues.

Protein sequence and structural features, as well as its post-translational modifications, not only determine its cellular function, but also its cellular localization and degradation rate. Therefore, these features also play significant roles in regulating the PTR ratios. We analyzed the effect of well known protein degradation signals on the PTR ratios and also searched for amino acid k-mers that significantly associated with the PTR ratios and the protein half-lives.

Although the N-terminal amino acid, which is known to affect protein stability via the N-end rule pathway, significantly associated with the PTR ratio (Figure 4.19), the N-terminal amino acid was not significant in the joint model, possibly because the effect was confounded with the start codon context.

A recent study by Kats and colleagues [65] in yeast indicated that the mean hydrophobicity of the first 15 amino acids plays a more important role in protein stability than the N-end rule pathway. We observed that mean hydrophobicity of the first 15 amino acids significantly associated with the PTR ratios of 8 tissues (3% higher PTR ratio on average, $FDR < 0.1$; Figure 4.20), however positively, in apparent contradiction with its negative effect on protein stability in yeast [65]. This may be due to the multiple roles of the 5' end of the coding region in gene expression regulation, which also includes a role in translation [135].

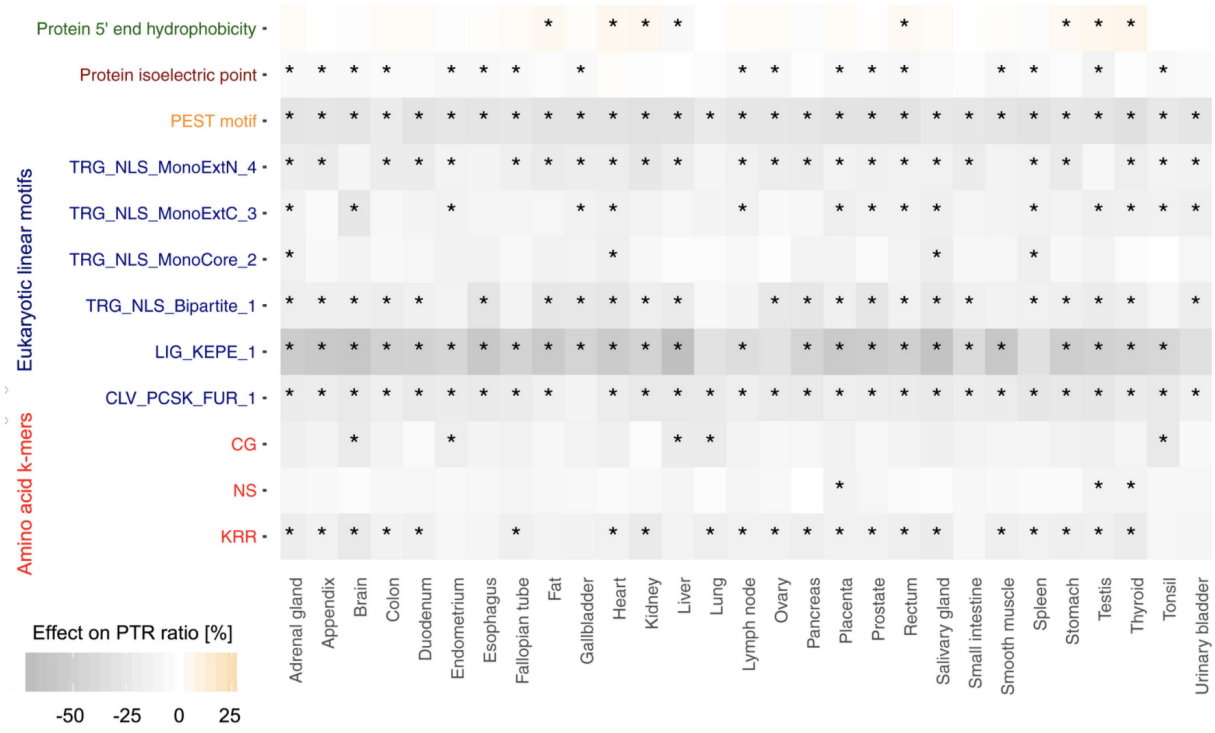


Figure 4.20: Protein sequence features that significantly associate with PTR ratios. Heatmap showing tissue-specific associations of protein sequence features with higher (red gradient) or lower (grey gradient) PTR ratios. Stars represent tissue-specific significance of the sequence feature with $FDR < 0.1$. Eukaryotic protein motif acronyms are *CLV_PKCS_FUR_1* (Furin (PACE) cleavage site), *LIG_KEPE_1* (Sumoylation site), *TRG_NLS_BIPARTITE_1* (classical bipartite nuclear localization signal), and three classical monopartite nuclear localization signals: *TRG_NLS_MonoCore_2*, *TRG_NLS_MonoExtC_3*, and *TRG_NLS_MonoExtN_4*.

We also considered protein surface charge–charge interactions because they can affect protein stability [136, 137], and because the charged polypeptides in the ribosome exit

tunnel can influence ribosome elongation speed [138]. Consistently, we observed that a one unit increase in the protein isoelectric point had a significant negative association with the PTR ratio (median 5%) in several tissues (Figure 4.20; $FDR < 0.1$). Our analysis also confirmed, genome-wide, the negative effect on PTR ratios of PEST regions, which are degrons that are rich in proline (P), glutamic acid (E), serine (S), and threonine (T) [139] that were present in 4,592 proteins , and estimated its median effect across tissues to a 26% lower PTR ratio (Figure 4.20; $FDR < 0.1$).

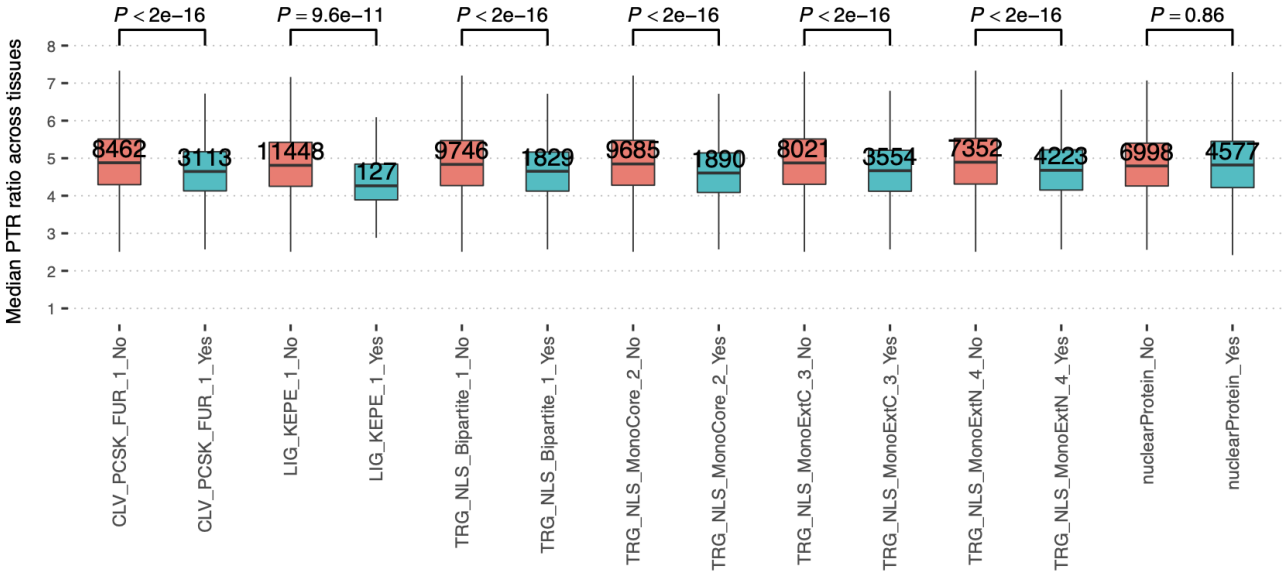


Figure 4.21: Identified eukaryotic linear protein motifs that associate with PTR ratio variation. Distribution of median PTR ratio across 29 tissues for genes with and without the eukaryotic linear protein motifs. *CLV_PKCS_FUR_1* (Furin (PACE) cleavage site), *LIG_KEPE_1* (Sumoylation site), *TRG_NLS_BIPARTITE_1* (classical bipartite nuclear localization signal), three classical monopartite nuclear localisation signals: *TRG_NLS_MonoCore_2*, *TRG_NLS_MonoExtC_3*, *TRG_NLS_MonoExtN_4*, and nuclear proteins (GO:0005634) in general. Four nuclear localization signals were associated with less median PTR ratio even though there is no significant PTR ratio difference between nuclear and non-nuclear proteins.

When we searched for other known protein motifs that associated with PTR ratios, we identified 6 linear protein motifs out of the 267 motifs from the ELM database [140] using a feature selection method (chapter 6). These 6 linear protein motifs contained 4 nuclear localization signals of the ELM database which associated negatively with PTR ratios. It is unclear why these four nuclear localization signals were associated negatively with PTR ratio even though there is no significant PTR ratio difference between nuclear (GO:0005634) and non-nuclear proteins (Figure 4.21). One possibility is that these linear motifs are destabilizing elements. Indeed, these 6 linear protein motifs were significantly associated with shorter protein half-lives (Figure 4.22). Also,

nuclear proteins with the four nuclear localization signals were associated with shorter half-lives compared to nuclear proteins without these signals (Figure 4.22). We also note that these linear motifs are KR-rich and this could reflect either that stretches of positively charged amino acid slow down translation or a technical bias due to the usage of trypsin as the protein digestion enzyme.

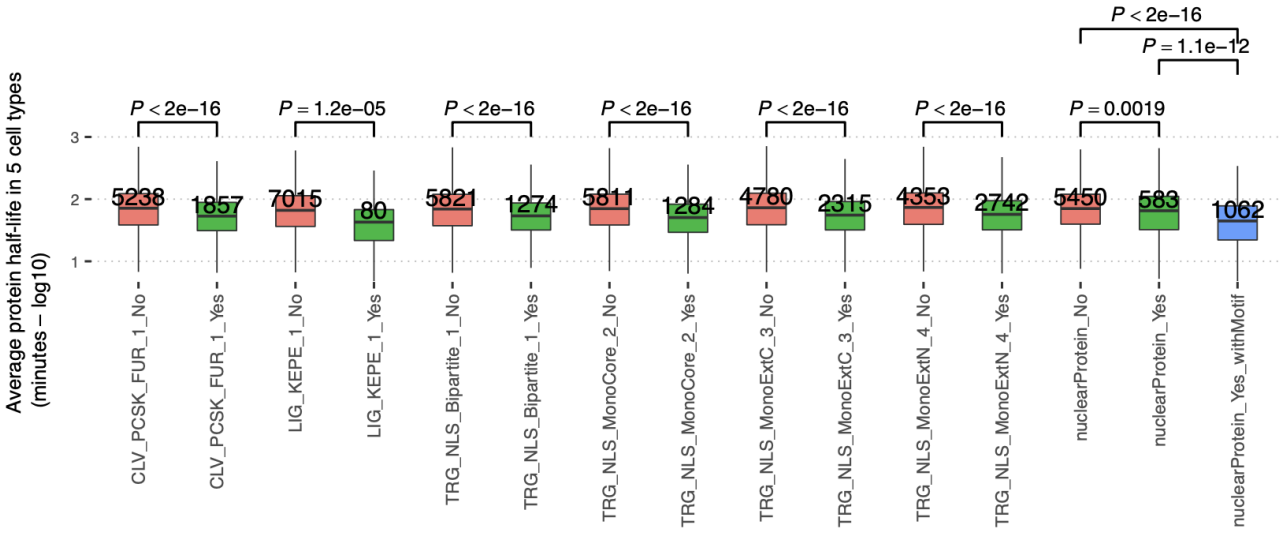


Figure 4.22: Identified eukaryotic linear protein motifs also display similar associations with protein half-life measurements. Same as Figure 4.20, for average protein half-lives across HeLa cells, B cells, NK Cells, Hepatocytes and Monocytes [97, 98]. Nuclear proteins with the four nuclear localization signals we have identified have even shorter protein half-lives compared to other nuclear proteins.

4.2 De-novo discovery of sequence motifs that are predictive of tissue-specific protein-to-mRNA ratios

In addition to analyzing the associations of the known mRNA and protein sequence elements with the PTR ratios, we also did de-novo motif searches to find the k-mers that are associated with the PTR ratio variation across genes. We executed these searches for 5' UTR, coding region and 3' UTR regions independently where the response variables were the PTR ratios of the 29 tissues we considered (chapter 6). Our search strategy produced fruitful results, de-novo identfying many known important post-transcriptional motif sequences as well as suggesting new ones. In the following sections we describe the found motifs along with the additional analyses we have done to characterize the motif site phylogenetic conservation scores and the gene set enrichment analyses of the genes having the consensus motifs.

4.2.0.1 5' UTR Motifs

Investigating every 3- to 8-mer in the 5' UTR, while controlling for occurrence of other k-mers, revealed 6 k-mers significantly associated with median PTR ratio across tissues, as well as 19 further k-mers associated with tissue-specific PTR ratio at a false discovery rate $FDR < 0.1$. The 6 k-mers that were significantly associated with median PTR ratio across tissues include AUG, the canonical start codon, for which at least one occurrence out-of-frame relative to the main ORF associated with about 18–33% lower median PTR ratios across tissues (Figure 4.23).

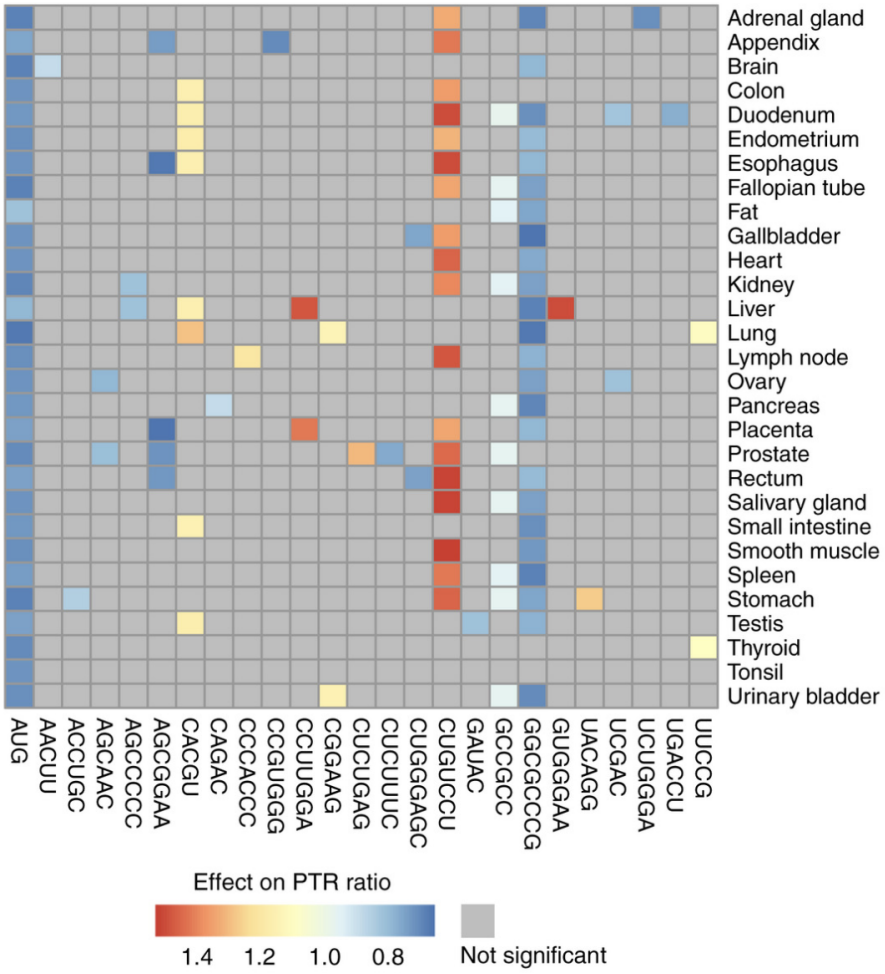


Figure 4.23: Identified motifs in 5' untranslated region. Estimated effect of PTR ratio in each tissue (row) of the 25 5' UTR k-mers (column) associating with either median PTR ratio across tissues or tissue-specific gene-centered PTR ratios. Color scale ranges from blue (negative effect) to red (positive effect). Gray marks non-significant ($FDR \geq 0.1$) associations.

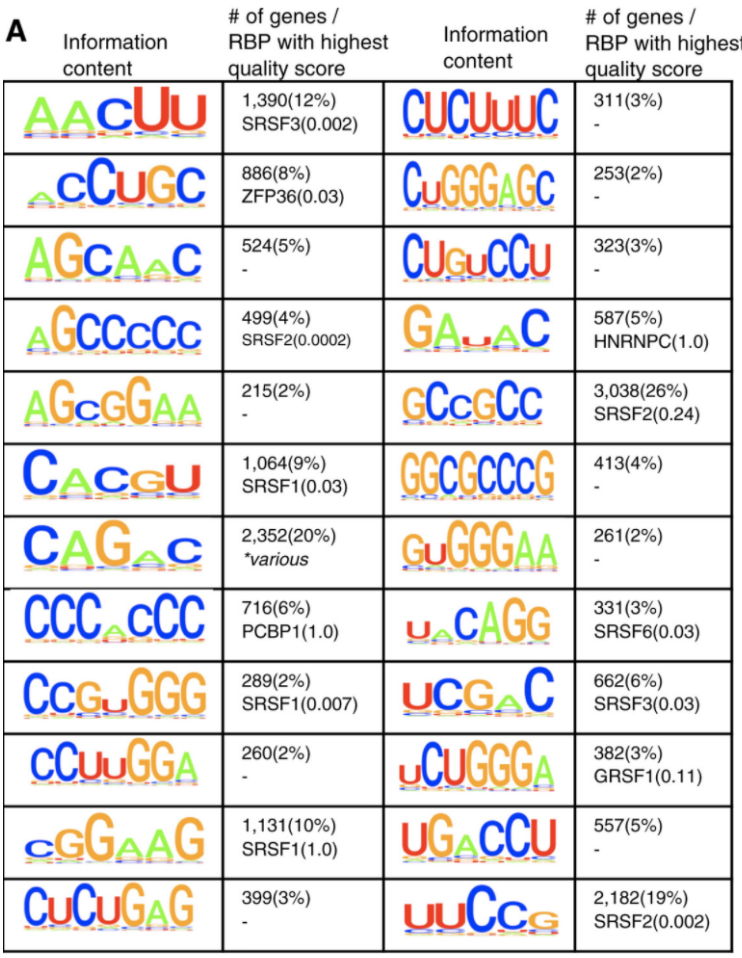


Figure 4.24: Identified eukaryotic linear protein motifs also display similar associations with protein half-life measurements. First and third columns: Motif information content logos for the 25 k-mers, obtained by motif consensus sequence search in 11,575 5' UTR sequences allowing for one mismatch. Second and fourth columns: number and percentage of transcripts consensus motif sequence among the 11,575 transcripts (first line) and best significantly matching RNA binding protein motif of the database ATtRACT [141] together with the ATtRACT motif quality score Q (value between 0 and 1, the higher the better).

AUG 3-mer, a strong positive control, was de-novo identified by our search approach. While the out-of-frame uAUG associated significantly with decreased PTR ratio in all 29 tissues, the other 24 5' UTR k-mers showed significant effects on PTR ratio ($FDR < 0.1$) only in certain tissues (Figure 4.23). These 24 k-mers were found in between 215 transcripts (2%) for AGCGGAA and 3,038 transcripts (26%) for GCCGCC (Figure 4.24). To search for possible proteins binding these k-mers, we queried the ATtRACT database ([141], which is, to our knowledge, the most extensive database of RNA binding motifs and contains 3,256 position weight matrices collected for 160 human RNA binding proteins. However, no obvious association between these k-mers and RNA binding motifs

could be drawn as most of the matches remain very distant ($ATtRACTqualityscore < 0.1$). A potential reason is that the ATtRACT database covers a small fraction of all human RBPs, which could consist of more than 1,500 proteins [60]. Nonetheless, 11 out of 24 of our k-mers were significantly more conserved than their flanking regions ($FDR < 0.1$, (Figure 4.25), Appendix Figure A.0)) and 10 showed significant enrichment for Gene Ontology (GO) terms, supportive for a potential regulatory role (Appendix Figure A.0). The appendix provides a comprehensive description of these results.

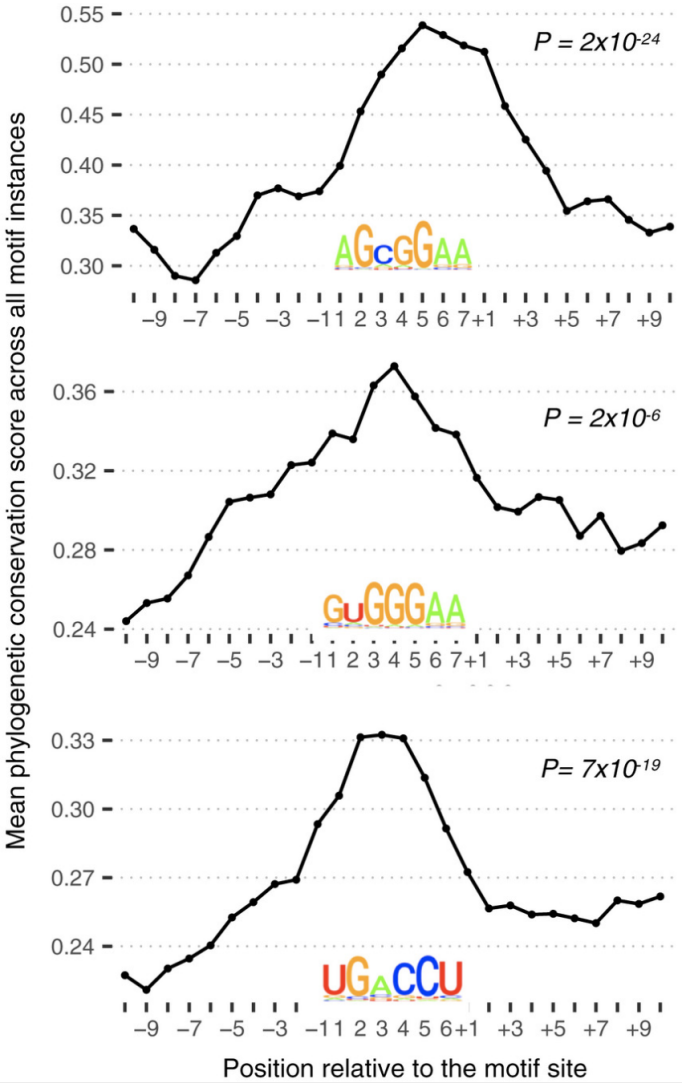


Figure 4.25: Identified motifs in 5' untranslated region. Average 100-vertebrate PhastCons score (y-axis) per position relative to the exact motif match instances in 5' UTR (x-axis) for three example k-mers that are significantly predictive of PTR ratios in specific tissues. P-values assess significance of the average 100-vertebrate PhastCons scores at the motif sites compared to the two 10-nucleotide flanking regions. The motif logos are constructed using all matches of the considered k-mer up to one mismatch in the 5' UTR sequences.

4.2.0.2 3' UTR Motifs

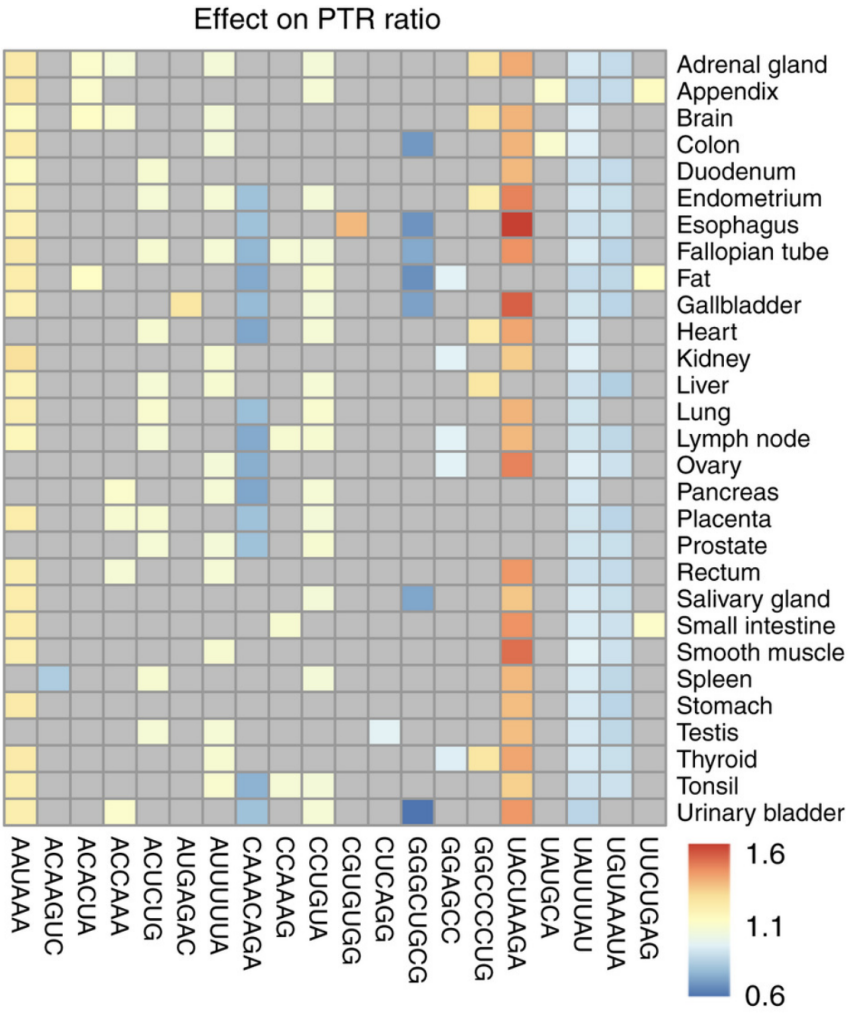


Figure 4.26: Identified motifs in 3' untranslated region. Estimated effect of PTR ratio in each tissue (row) of the 20 3' UTR k-mers (column) associating with either median PTR ratio across tissues or tissue-specific gene-centered PTR ratios. Color scale ranges from blue (negative effect) to red (positive effect). Gray marks non-significant ($FDR \geq 0.1$) associations.

De novo motif searching in the 3' UTR revealed 20 k-mers significantly associated with median PTR ratios across tissues or with tissue-specific PTR ratio ($FDR < 0.1$; Figure 4.26, 4.27). This recovered 4 well-known mRNA motifs: the polyadenylation signal AAUAAA [142], the AU-rich elements UAUUUAU [143, 144] and AUUUUUA [145], and the binding site of the Pumilio family of proteins UGUAAAUA [146]. The polyadenylation signal AAUAAA associated with between 13 and 28% increased PTR ratio across tissues (median 21%, $FDR < 0.1$; Figure 4.26), consistent with one role of polyadenylation signals in translation [147]. The AU-rich element UAUUUAU was found in 3,158 genes (27%) and associated with lower PTR ratios by about 9% consistently across tis-

sues, in agreement with its function in mRNA destabilization and translational silencing [143, 144]. The Pumilio motif UGUAAAUA was found in 1,320 genes (11%) and is the binding target of members of the Pumilio family of proteins which regulate translation and mRNA stability in a wide variety of eukaryotic organisms [146].

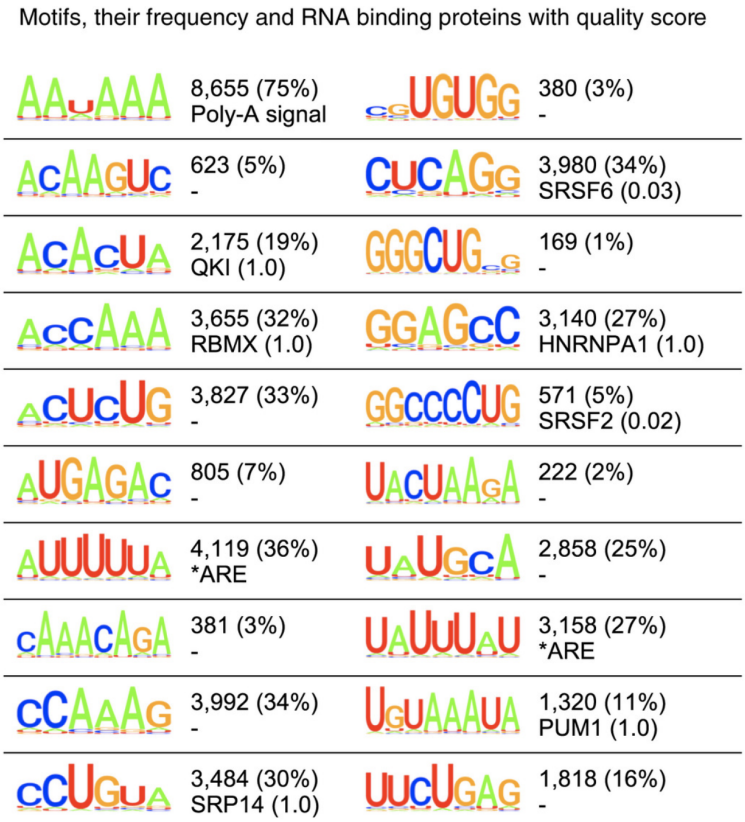


Figure 4.27: Identified eukaryotic linear protein motifs also display similar associations with protein half-life measurements. First and third columns: motif information content logos for the 20 k-mers of panel A, obtained by motif consensus sequence search in 11,575 3' UTR sequences allowing for one mismatch. Second and fourth columns: number and percentage of transcripts consensus motif sequence among the 11,575 transcripts (first line) and best significantly matching RNA binding protein motif of the database ATtRACT [141] together with the ATtRACT motif quality score Q (value between 0 and 1, the higher the better.

In addition to these four evolutionarily conserved motifs (Appendix Figure A.-2), we identified 7 motifs, namely ACCAAA, CCAAAG, CUCAGG, GGGCUGCG, GGAGCC, GGCCCUG, and UUCUGAG; these are also significantly conserved with respect to the background flanking regions (Appendix Figure A.-2). While some of these conserved motifs were not previously reported in the literature, some of the obtained k-mers may possibly be the binding motifs of RBPs with a post-transcriptional role. One notable example is the k-mer ACACUA, which matches a recognition site of the QKI protein according to the ATtRACT database (quality score = 1.0), which is highly enriched in the

brain (Human Protein Atlas; [148]) and important for myelinization [149], mRNA stability, and protein translation [150]. Another example is the well-conserved ACCAAA, present in 3,655 genes (32%), possibly being the target motif of RBMX (ATtRACT quality score = 1.0) which plays several roles in the regulation of post-transcriptional processes [151]. The appendix provides a full analysis per motif based on their number of occurrences, phylogenetic conservation scores (Appendix Figure A.-2), and a gene set enrichment analysis for the genes having the consensus motif (Appendix Figure A.-3).

4.2.1 Validation of the identified motifs

We assessed the effects of motifs in a dual reporter assay in which the nine tested motifs (GGCCCCUG, UACUAAGA, UAUUUAU, UGUAAAUA, CACGU, CCCACCC, CUGUCCU, GGCGCCCG, UUCCG) were inserted in the 5' UTRs or 3' UTRs of Gaussia luciferase constructs. The same plasmid also expressed a secreted alkaline phosphatase as control. This assay showed significant effects for two positive controls: the out-of-frame upstream AUG and the out-of-frame upstream ORF, i.e., an upstream AUG with an in-frame stop codon within the 5' UTR (Figure 4.28; $P < 0.0001$).

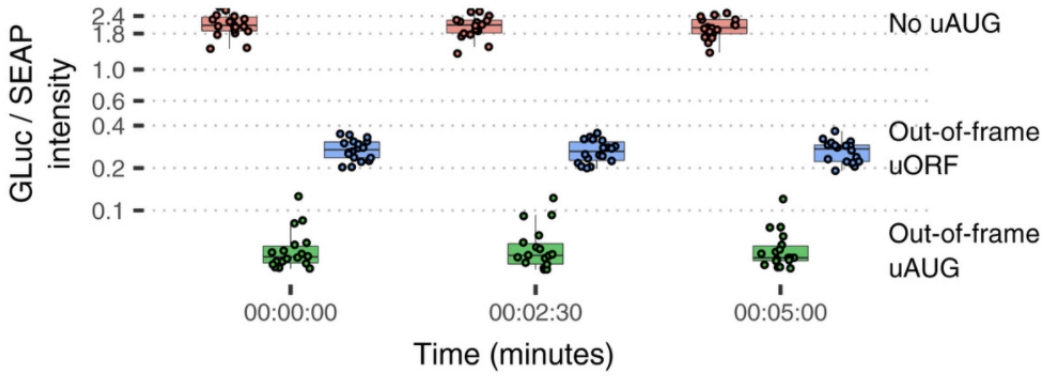


Figure 4.28: Reporter assay of the AUG in 5' UTR. Ratio of GLuc over SEAP intensities normalized per experiment (y-axis, n = 18) per time point (x-axis) and construct: no insertion (pink), inserted out-of-frame AUG (green), and inserted uORF, i.e., inserted AUG with an inserted stop codon in-frame in the 5' UTR (blue). Shown are the quartiles (boxes and horizontal lines) and furthest data points still within 1.5 times the interquartile range of the lower and upper quartiles (whiskers).

For the remaining tested motifs, control constructs containing scrambled versions of the tested motif were also assayed (Appendix Figs S15 and S16, Table EV10). Two tested motifs (UUCCG and CUGUCCU) showed significant effects in the direction predicted by the model (Figure 4.29, $FDR < 0.1$). Most motifs had small predicted effects, so that significance was difficult to attain in such assays. Taking this into account, four further motifs, including two positive controls, the AU-rich 3' UTR motif UAUUUAU and the

Pumilio response elements, as well as the new motifs CCCACCC and GGCCCCUG, showed effects consistent with the model prediction (Appendix Figures A.-2 and A.-1) both in direction and in amplitude.

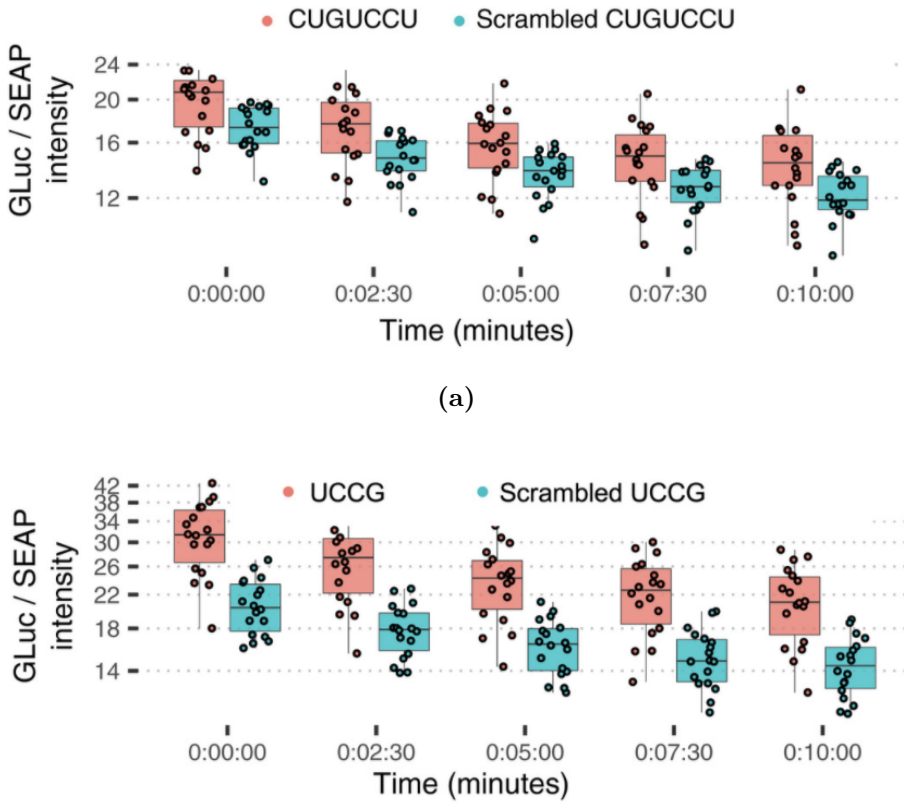


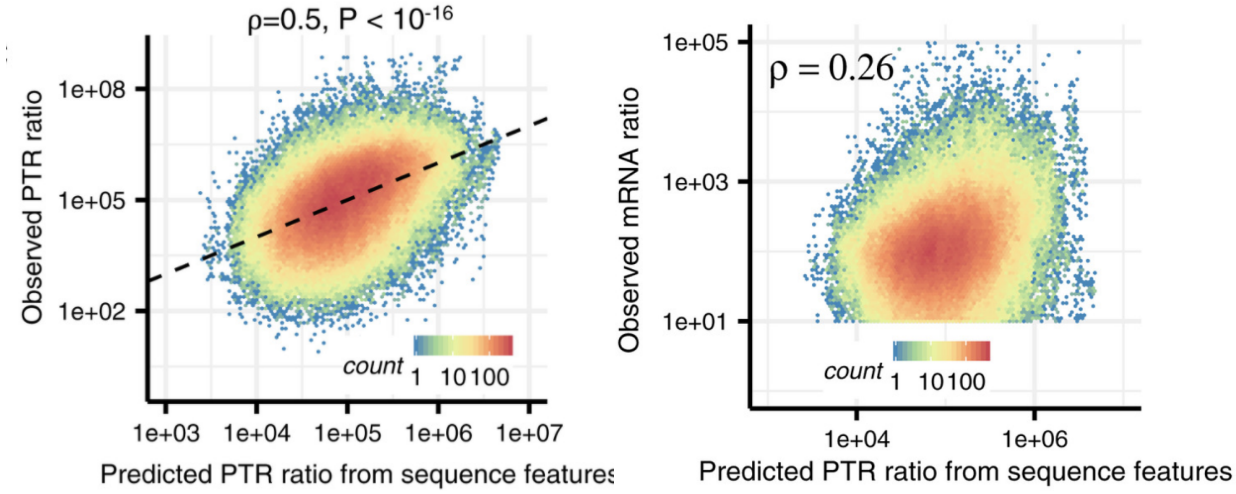
Figure 4.29: Reporter assay results for tested motifs. a) Reporter assay of the 5' UTR motif CUGUCCU. Ratio of GLuc over SEAP intensities normalized per experiment (y-axis, n = 18) per time point (x-axis) and construct: inserted motif (pink), inserted scrambled motif UUUGCCC (blue). Shown are the quartiles (boxes and horizontal lines) and furthest data points still within 1.5 times the interquartile range of the lower and upper quartiles (whiskers). **b)** As in (b) for inserted 5' UTR motif UUCCG (pink) or a scrambled version of it (blue, CUUCG).

5 Prediction of protein-to-mRNA ratios from sequence features

The methodology, results and figures presented in this section are part of the manuscript "Quantification and discovery of sequence determinants of protein per mRNA amount in 29 human tissues" from Eraslan and Wang et al. 2019 [1] and "A deep proteome and transcriptome abundance atlas of 29 healthy human tissues" from Wang and Eraslan et al. 2019 [2]

5.1 An interpretable model explaining PTR ratios from sequence

The multivariate linear model combining all these sequence features predicted PTR ratios at a median relative error of 3.2-fold on held-out data (10-fold cross-validation), which is small compared to the overall variation of PTR ratios (200-fold for the 80% equi-tailed interval). This model explained 22% (median across tissues) of the variance (Figure 5.1 -a). Moreover, we observed that the predicted PTR ratios moderately positively correlated with the mRNA levels (Figure 5.1 - b; $Spearman'srho = 0.26$). Hence, our model supports the hypothesis that highly transcribed genes also have optimized sequences for post-transcriptional up-regulation, hence yielding higher amounts of proteins, which is consistent with earlier work by Vogel and colleagues [152]. Combining these sequence features together with the mRNA profiles in a single linear model explained 58% of the variance of tissue-specific protein levels in average (minimum 49% in pancreas, maximum 63% in liver), increasing the proportions of variance of tissue-specific protein levels explained with mRNA profiles alone by 10% in average ($P = 3 \times 10^{-10}$, Wilcoxon test).



(a) **(b)**

Figure 5.1: Explained variances in mRNA levels and mRNA half-lives. a) Observed PTR ratios of all tissues (y-axis) versus predicted PTR ratios by the interpretable sequence model (x-axis) which includes 18 sequence feature groups representing 204 post-transcriptional regulatory elements. **b)** Observed mRNA levels of all tissues (y-axis) correlates with predicted PTR ratios by the interpretable sequence model (x-axis) which includes 18 sequence feature groups representing 204 post-transcriptional regulatory elements. This observation supports the hypothesis that genes that are highly transcribed are also optimized for post-transcriptional regulation leading to higher protein levels.

5.2 Extended model with experimentally characterized elements

There are thousands of further sequence elements that could play a role in controlling the PTR ratios, including the binding sites of any of the 2,599 catalogued human miRNAs [134], the binding sites of the estimated 1,542 RNA binding proteins [60], and elements subject to mRNA modifications and post-translational modifications of certain amino acids. In this context, derivation of a more comprehensive yet interpretable model of PTR ratio from sequence is difficult. One reason is that the sequence determinants driving the binding of these factors and these modifications are poorly charted. Another reason is that binding sites of RBPs and miRNAs often co-occur due to cooperative and competitive binding [130, 131, 132, 133], which makes untangling the effects of individual sequence elements difficult. Nevertheless, in order to explore the degree to which the prediction of the PTR ratio from sequence could be improved in principle, we considered a model that was not based on sequence alone, rather also including experimental characterization of such interactions and modifications of mRNA and proteins.

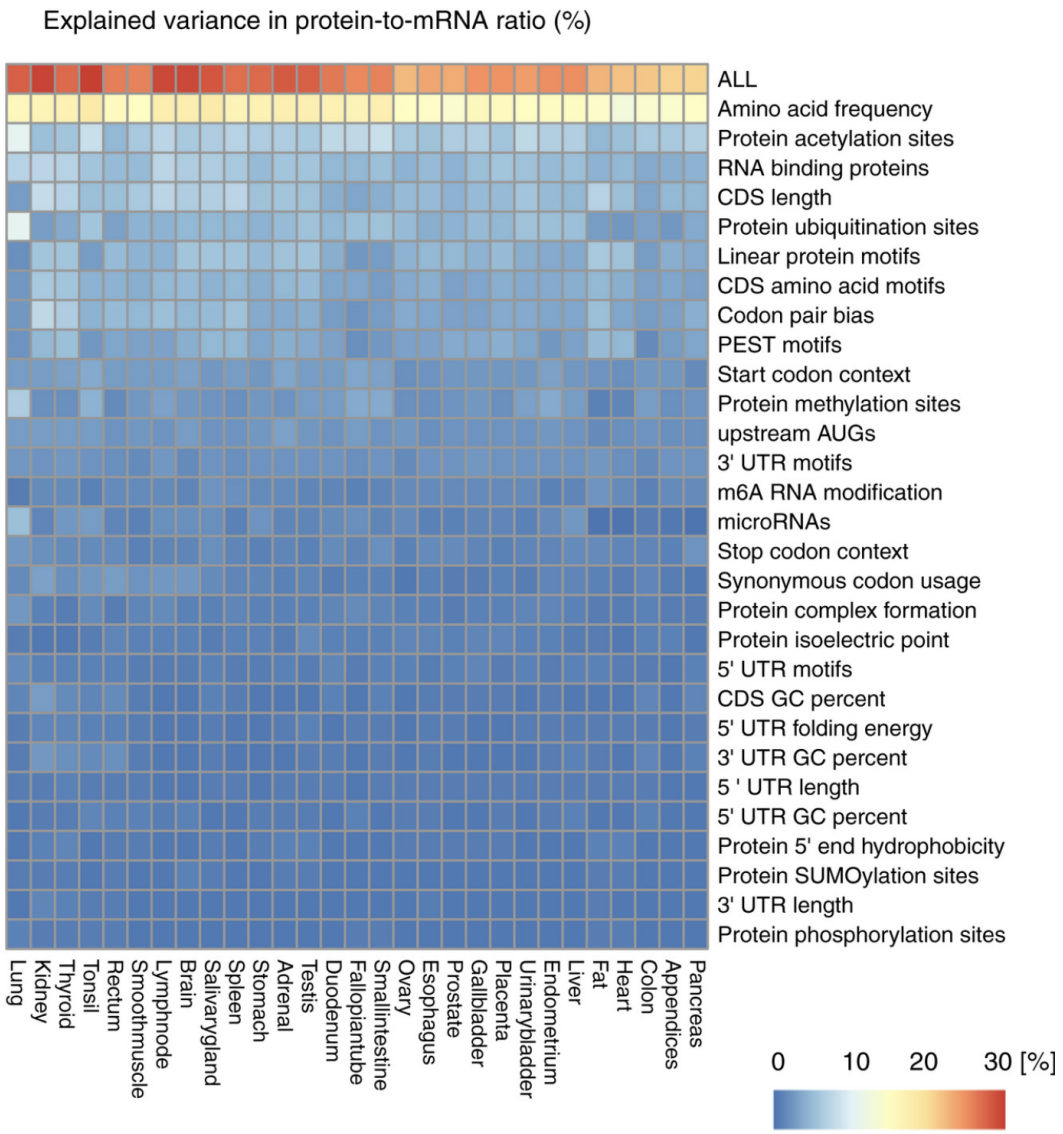


Figure 5.2: Explained variances in tissue-specific PTR ratios by the inspected sequence elements. Proportion of variance in tissue-specific PTR ratios explained (R2) by separate linear models representing one sequence feature group in each tissue. The first row (labeled "ALL") corresponds to the linear model combining all of the features displayed in the consecutive rows.

This extended model included (i) N6-methyladenosine (m6A) mRNA modification, an abundant modification enhancing translation [153]; (ii) binding evidence for 296 miRNAs from the miRTarBase database [134] with more than 200 targets in our dataset; (iii) whether proteins are part of protein complexes, which is known to stabilize proteins [154, 155]; (iv) binding evidence to 112 RNA binding proteins (RBPs) [12]; and (v) phosphorylation, methylation, acetylation, SUMOylation, and ubiquitination of certain amino acids [156]. This analysis showed that with the inclusion of these experimentally characterized features, the proportion of variance of PTR ratio increased to a

median across tissues of 27% (Figure 5.2; min 24%, max 31%). Moreover, combining the extended set of features together with the mRNA profiles in a single linear model explained 62% of the variance of tissue-specific protein levels in average (minimum 53% in pancreas, maximum 68% in tonsil). However, these increased proportions of variance explained do not imply that these experimentally characterized features are not driven by regulatory elements encoded in sequence. Rather, they may reflect that our primary regression of PTR ratio on sequence features was not powerful enough to capture those underlying, potentially complex, regulatory sequence elements.

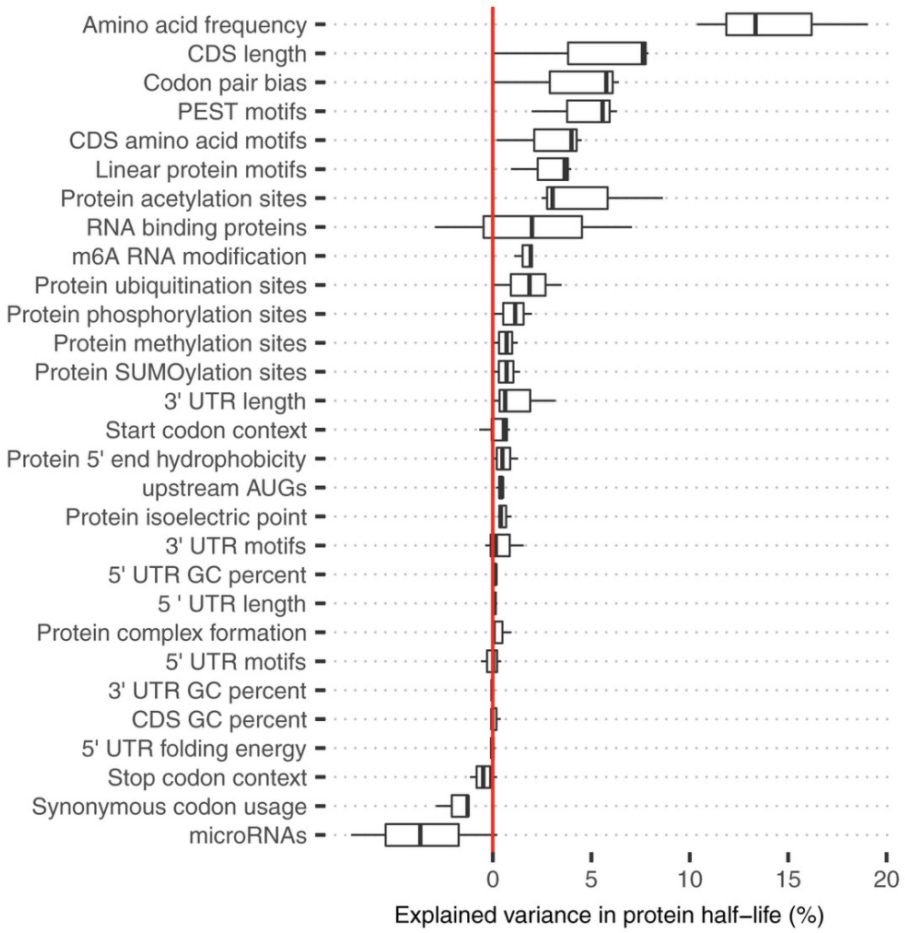


Figure 5.3: Distributions of explained variances by linear models of individual sequence feature groups in protein half-lives of HeLa cells [97], B cells, NK cells, hepatocytes, and monocytes [98]. The distribution of the total explained variance by the linear model that combines all of the listed sequence features is displayed in the first line. Shown are the quartiles (boxes and vertical lines) and furthest data points still within 1.5 times the interquartile range of the lower and upper quartiles (whiskers).

Analysis of explained variance of individual feature groups indicated that amino acid frequency alone explained on average 15% of the variance in PTR ratios (min 12%, max 15%; Figure 5.2 and Appendix Figure A.0). This is followed by protein acetylation sites, binding sites of 112 RBPs [12], CDS length, protein ubiquitination sites, and linear protein motifs. These results suggest that sequence elements affecting protein stability may be the dominant features predictive of PTR ratios. In line with this possibility, we observed that the explained variance in PTR ratio by these sequence features highly correlated ($Spearman's \rho = 0.59, P = 0.001$) with their explained variances in protein half-lives (Figure 5.4) in five cell types [98, 97].

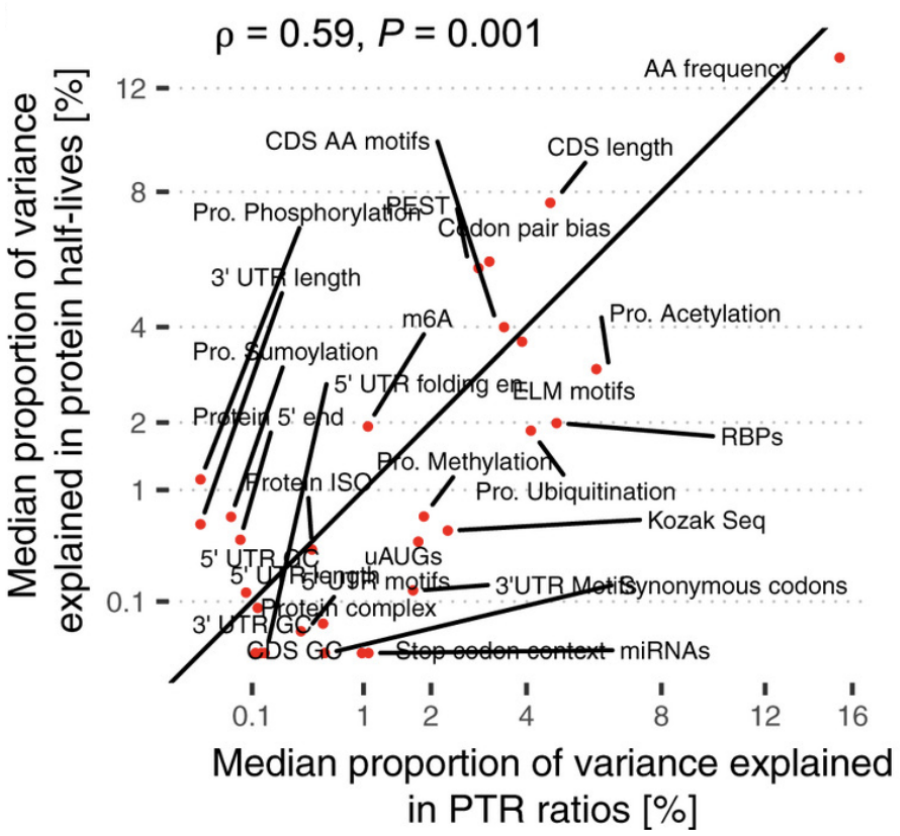


Figure 5.4: Sequence elements affecting protein stability may be the dominant features predictive of PTR ratios. Median proportion of variance in tissue-specific PTR ratios explained (x-axis, R2) by each sequence feature group shown in (D) highly correlates with median proportion of variance explained in protein half-lives of five different cell types (y-axis). Most of the explained variance in PTR ratios is dominated by sequence elements that are highly predictive of protein half-lives.

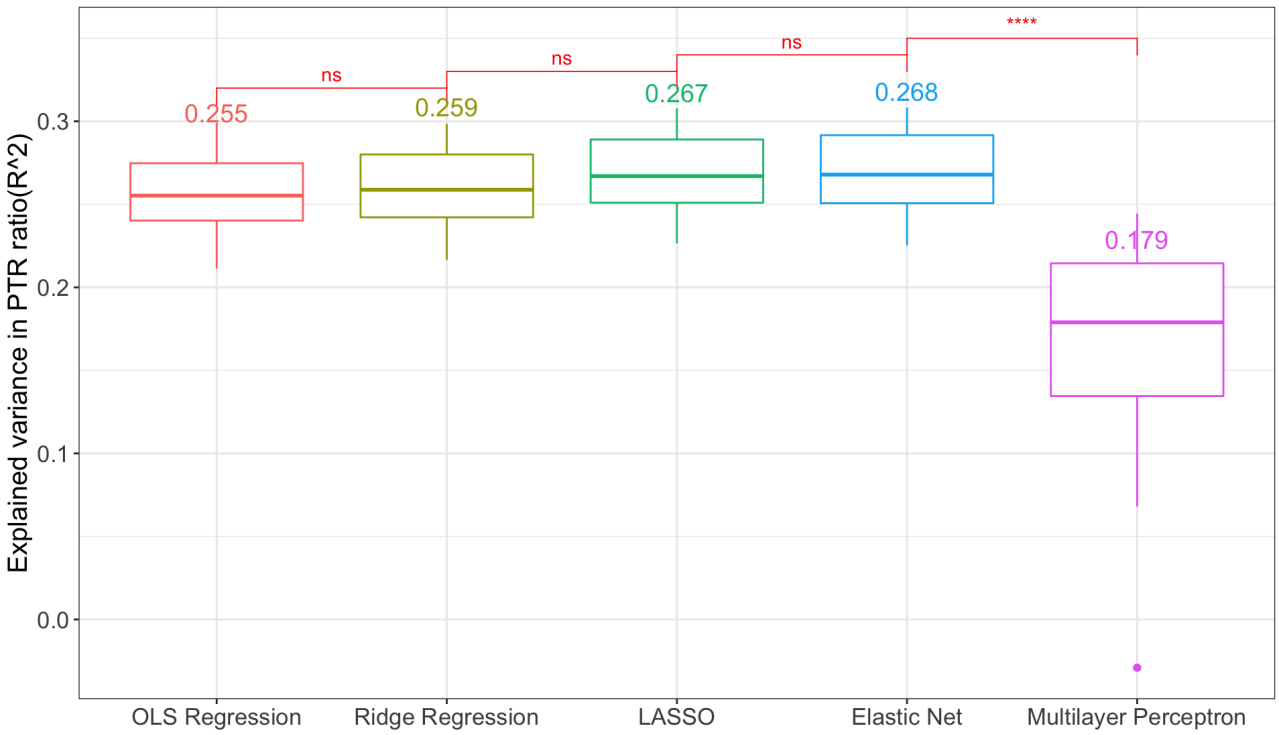
The proportion of variance in PTR ratio explained by the binding evidence to 112 RNA binding proteins [12] varied from 3 to 6% across tissues (median 5%), while 150 latent variables of 296 miRNAs' binding evidence explained on average only 1%. Overall,

these RBPs appeared to be ubiquitously expressed since 81 out of the 112 RBPs (77%) were detected expressed at the proteome level and at the mRNA level in all tissues. Ubiquitous expression of RBPs and the frequent co-binding of RBPs and miRNAs may be two reasons why tissue-specific effects of RBP binding on PTR ratio did not show significant correlations with the corresponding tissue-specific RBP expression levels (Appendix Figures A.2, A.0). Nevertheless, in 16 of these RBPs, there was a significant difference between their across-tissue covariation with their target and non-target genes (Figure 4.17). The binding of these regulatory elements was among the top mRNA features explaining the tissue-specific mRNA levels and mRNA half-lives of three different cell types (Figure 4.18). The binding of the considered 112 RBPs explained on average 18% (min 13%, max 21%) and features representing miRNA binding explained on average 5% (min 4%, max 7%) of the variance in tissue-specific mRNA levels (Figure 4.18-a). Consistent with that, RBP binding explained on average 17% of the variance (min 12%, max 22%) in mRNA half-lives of K562, HEK293, and HeLa Tet-off cells (Figure 4.18-b). Likewise, features representing miRNA binding explained 5% (median, min 4%, max 5%) of the variance in mRNA half-lives of these three cell lines. Altogether, the differences and similarities in the explained variances of mRNA levels, mRNA half-life, PTR ratio, and protein half-life suggest that the RBPs and miRNAs considered in our model may be more effective in regulating mRNA stability rather than PTR ratios.

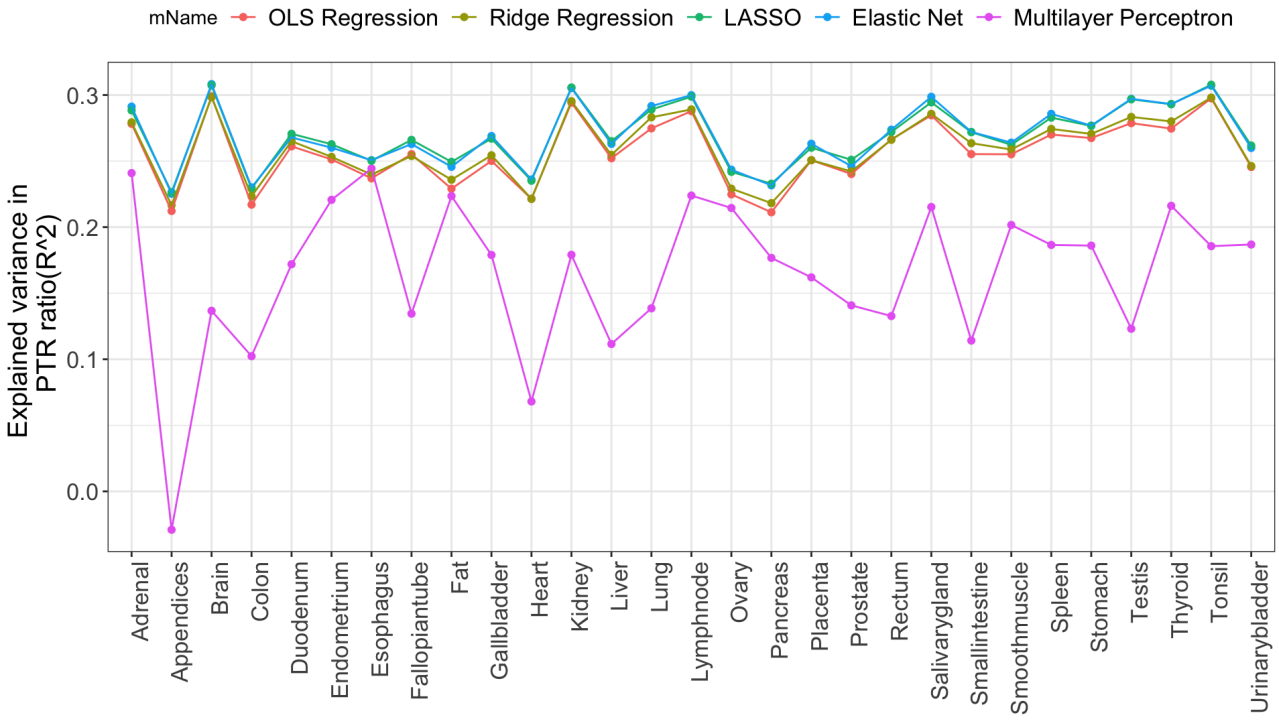
Of note, the proportion of variance explained is driven by the combination of effect size, frequency, and variability of the features across genes. Hence, sequence features which play a crucial role for translation, like the Kozak sequence, can only explain 3% of the genome-wide PTR ratio variation by itself because it is already optimized for most of the genes in the genome. Also, the 5' and 3' UTR motifs explain a small fraction of the variance between genes although their effect size can be large because they typically occur in a small number of genes.

5.2.1 Model comparison with the full set of features

In order to determine the best model with the whole set of sequence and experimentally characterized features, we compared the prediction performance of various regression methods (Figure 5.5) with their optimized hyperparameter values. Here we observed that for the full set of 472 features that are characterized either based on the sequence or the experimental measurements, the models with regularization effects, such as ridge regression, lasso and elastic net, scored higher explained variance (R^2). On the other hand models like multilayer perceptrons which are able to capture nonlinear relationships between the covariates, but at the same time have higher number of parameters, overfitted and performed poorly in generalization in the 10 fold cross-validation setting. Given these results, we see that modelling more complex relationships between 472 features was not feasible with the available PTR values of 9,000 genes for each tissue. Furthermore, we even observed that applying some regularization on the parameters of the simple linear regression slightly increased the explained variance by preventing overfitting, which means even modelling the additive effects of 472 features was not feasible with this amount of data.



(a)



(b)

Figure 5.5: Comparison of the different regression model performances. a) Boxplots displaying the distribution of the explained variances in tissue specific PTR ratios by five different regression models with optimized hyperparameter values. **b)** Comparison of explained variances in PTR ratio of each tissue by the considered regression models with optimized hyperparameter values.

6 Methods

6.1 Methods used for the analysis of the translational gene expression regulation in maturating human dendritic cells

6.1.1 Segmenting the genome based on the transcriptome data

Most human genes have multiple transcription start and polyadenylation sites, which are fine tuned in a tissue-specific manner [157, 158, 159]. In order to determine the boundaries of the transcriptional units in human dendritic cells, we segmented the genome with a two state hidden markov model (HMM) (Figure 6.1) applied on the coverage of combined iDC, 4h, 24h transcriptome data and adjusted the resulting segments with min-length max-gap algorithm as previously described in Zacher et al. [160]. Here two of the hidden states correspond to the transcribed and untranscribed regions, while the observed variables are the RNA read occupancy profiles in 1 base-pair resolution.

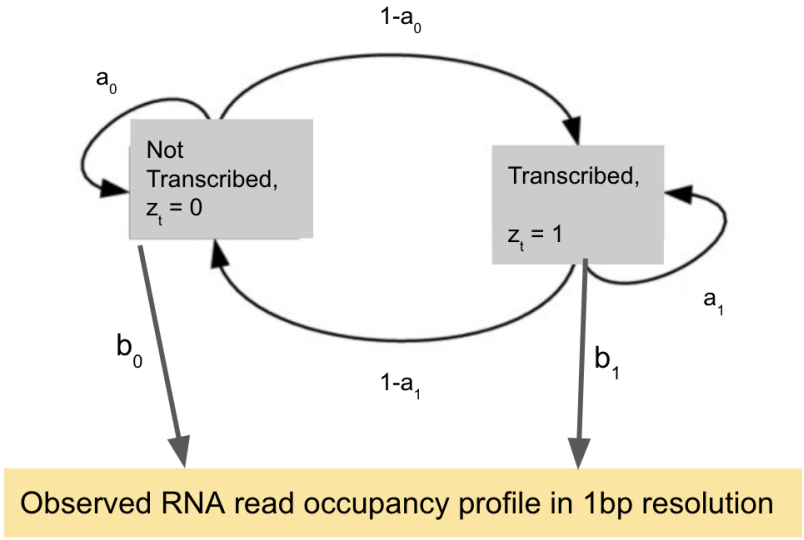


Figure 6.1: Two state hidden markov model for segmenting the transcribed regions. In the used HMM, two hidden states correspond to the transcribed and untranscribed regions of the genome. a_0, $1-a_0$, a_1, $1-a_1$ are the state transition probabilities and b_0, b_1 are the probabilities of the observed 1bp RNA occupancy profile given the "not transcribed" and "transcribed" states respectively.

After segmenting the genome based on the total transcriptional profiles at iDC, 4h and 24h, the segments standing for the transcribed regions of the genome were mapped to the human RefSeq annotation hg38 to find the 5' and 3' untranslated regions (UTRs) of the genes. We defined the gene 5' UTR to be the region between the start of the mapped segment and the first annotated start codon, while the region between the last annotated stop codon and the end of the mapped segment is accepted to be the 3' UTR region of the gene. Defining the gene untranslated regions as such enabled us to make sure that the ribosome profiling reads mapped to these locations do not belong to the coding region of any of the alternative isoforms of the gene.

6.1.2 Computation of ribosome density values in 5' UTR, 3' UTR and coding regions

Ribosome protected fragments signal the total number of ribosomes that are employed in protein synthesis, which is dependent on both the number of mRNAs in the cell and the number of ribosomes that translate each transcript [22]. Therefore, ribosome density of a specific region of the genome can be described as the ratio of the number of ribosome profiling reads to the number of transcriptome reads mapped to this region. We computed the log2 ribosome density at 3' UTR, 5' UTR and coding regions by taking the log2 ratio of the library size normalized counts of ribosome profiling reads over transcriptome reads mapped to the exonic intervals of these regions. While counting the number of ribosome profiling reads mapped to a region, we considered the P-site of the ribosomes which is at 13 nucletides downstream of 5' end of each ribosome protected fragment [161].

6.2 Methods used for the analysis and the prediction of the sequence determinants of protein per mRNA amounts in 29 human tissues

The methodology, results and figures presented in this section are part of the manuscript "Quantification and discovery of sequence determinants of protein per mRNA amount in 29 human tissues" from Eraslan and Wang et al. 2019 [1] and "A deep proteome and transcriptome abundance atlas of 29 healthy human tissues" from Wang and Eraslan et al. 2019 [2]

6.2.1 Preprocessing of the protein levels, mRNA levels, and PTR ratios

The protein data in MaxQuant [162] output file 'proteinGroups.txt' were filtered such that the Reverse, Only.identified.by.site, and Potential.contaminant columns were not equal to "+". Moreover, we restricted to unambiguously identified gene loci by requiring the number of Ensembl Gene IDs in the Fasta.headers column to equal 1. To calculate

protein expression levels, IBAQ values equal to zero were set as missing values (NA). Next, IBAQ values were adjusted to have in each tissue the same median than the overall median by adding in the logarithmic scale a tissue-specific constant.

About 10% of the genes were reported to have 2 or more transcript isoforms in the MaxQuant file 'proteinGroups.txt'. We defined as major transcript isoform per gene the transcript isoform reported in the MaxQuant file 'proteinGroups.txt' that had the largest sum of IBAQ values across all tissues. We used these major transcript isoforms for all tissues, to compute all sequence features and to compute mRNA levels.

For each tissue, only the mRNA replicates which had a matching protein sample were used throughout the analysis. Paired-end raw read files were quality-checked with FastQC software (Babraham Bioinformatics – FastQC A Quality Control tool for High Throughput Sequence Data), and the overrepresented adapter sequences were trimmed using the Trim Galore software (Babraham Bioinformatics – Trim Galore!). After that, resulting read files were checked again with FastQC and the reads were mapped with STAR alignment software [163] to human genome annotation Hg38.83, with the parameter of maximum number of multiple alignments allowed for a read to be equal to 1 (–outFilterMultimapNmax).

To estimate the mature mRNA levels, for each sample (each replicate in each tissue) the number of reads that map to exonic and intronic regions of the transcript (which was decided to be used based on the major protein isoform) was counted separately and then normalized by the total exonic and intronic region lengths, respectively. Next, the intronic counts normalized by the intronic region length were subtracted from exonic counts normalized by the exonic region length. The resulting normalized exonic counts per sample (i.e., each replicate of each tissue) were corrected by the library size factor obtained with the Bioconductor package DESeq2 and further log-transformed (log10). Finally, technical replicates were summarized by taking the median value. We set a cutoff of 10 reads per kilobase pair for a transcript to be treated as transcribed, which further improved the correlation between mRNA and proteins, possibly because of the poorer sensitivity of proteomics for lowly expressed genes or because of higher technical noise in low ranges of expression for RNA-Seq and for proteomics. Tissue-specific PTR ratios were computed as the logarithm in base 10 of the ratio of the normalized protein levels over the normalized mRNA levels.

6.2.2 mRNA isoform level quantification

In order to obtain tissue-specific mRNA transcript isoform FPKM levels, we used Kallisto [164] with Gencode annotation Hg38.83 using default parameters. For each gene, the major isoform in a specific tissue was defined to be the isoform with the largest FPKM value among all isoforms with FPKM ¿ 1. Thereon, the number of major isoforms across the 29 tissues with unique Ensembl Transcript IDs was counted per gene.

6.2.3 Explained variance of protein levels and relative protein levels by mRNA levels

For protein levels, we performed a linear regression of log-transformed protein levels against either log-transformed mRNA levels of the matching tissue or the complete log-transformed mRNA levels across all tissues. Explained variance was reported as adjusted R2, and statistical significance was assessed using the chi-square test for nested linear models. For relative protein levels, the same was done for log-transformed and median-centered protein levels against log-transformed and median-centered mRNA levels.

6.2.4 Multi-omics factor analysis on mRNA levels and PTR ratios

Multi-omics factor analysis [89] was applied to mRNA levels and PTR ratio matrices (7,822 by 29) of the 7,822 genes detected expressed at the mRNA level and at the protein level in at least 15 tissues. The mRNA levels and PTR ratios were mean-centered per gene across tissues before the fitting was performed.

6.2.5 Feature engineering for the multivariable predictive model

In order to develop an interpretable predictive model of PTR ratios from sequence, in addition to the de-novo search for novel mRNA and protein motifs which could be the binding sites of several factors such as RNA binding proteins, miRNAs, protein localization and degradation factors, I did a deep literature search of all the sequence elements that are known to be important for post-transcriptional gene expression regulation and included them in my model. However, including each of these features to the covariate matrix needs meticulous preprocessing and feauture engineering steps in order to achive a model with the best performance. This feature engineering step was also important for keeping the model as interpretable as possible because one of our main objectives was to gain more biological insights in post-transcriptional regulation through our integrated model. Therefore, intead of applying standard feature extraction techniques used in black-box machine learning techniques, we carefully crafted each of the sequence features. In the sections below, these feature engineering techniques are presented for each of the sequence feature we have included.

6.2.5.1 5' UTR folding energy (secondary structure proxy)

The sequence spanning 100 nt 5' and 100 nt 3' of the first nucleotide of the canonical start codon was extracted for all transcripts with a valid PTR value in at least one tissue. The folding energies were computed via Vienna-RNAfold package [165] with 51-nt-wide sliding window for each center position in $[-75, +75]$ nt relative to the first nucleotide of the canonical start codon. The effect and P-values of the log2-transformed negative minimum folding energy values at each position on median PTR across tissues were assessed individually with a linear regression model, in which all the analyzed sequence

features were included as covariates. P-values were corrected for multiple testing using Benjamini–Hochberg correction.

6.2.5.2 Kozak sequence and stop codon context

In order to find the effect size of 1bp nucleotide change at Kozak sequence or stop codon context on PTR ratios, a simple linear regression model was used where the feature matrix X is contains the categorial nucleotide information in a $[-6, +6]$ nt window around the canonical start and stop codons. Base reference level for each position is taken to be the most frequent nucleotide at this position among the inspected 11,575 genes.

6.2.5.3 Codon frequency

Codon frequency was encoded as the log2 of the frequency of each of the 61 coding codons (number of codons divided by coding sequence length). Using the frequency in natural scale led to a decreased explained variance by 1%. In addition, codon pair frequencies were modeled in the design matrix as the first 2 principal components of the codon pair frequency matrix consisting of 3,721 features.

6.2.5.4 Linear protein motifs

Linear protein motifs were downloaded from the ELM database [140] as regular expressions. We classified proteins as containing an ELM motif if the regular expression matched at least once in the protein sequence. Thereafter, we selected the ELM motifs significantly associating with PTR ratios in at least one tissue by utilizing LASSO feature selection where the PTR ratios were corrected for the core sequence features, which we defined as the motifs identified de novo, the 5' UTR folding energies at positions 0 and +48, start codon context, codon frequencies, codon pair bias indicators, stop codon context, UTR and CDS region lengths, PEST motifs, protein isoelectric point, and protein N-end hydrophobicity.

6.2.5.5 N-terminal residue

The second residue of the protein sequence was extracted.

6.2.5.6 Protein 5' end hydrophobicity

The mean hydrophobicity value of the amino acids 2–16 at the 5' end of the protein was calculated by the hydropathy index per amino acid values reported in Kyte and Doolittle [166].

6.2.5.7 Protein isoelectric point

Protein isoelectric points for 11,575 protein considered in our model were computed with the IPC-Isoelectric Point Calculator software (Kozlowski, 2016).

6.2.5.8 PEST-region

We classified protein sequences as 'PEST-region containing' if the EMBOSS program epestfind [139, 167] identified at least one 'PEST-no-potential' hit.

6.2.5.9 m6A mRNA modification

We classified mRNAs as m6A-modified if at least one m6A peak for the same gene locus in untreated HepG2 cell line was reported in Supplementary Table 6 of Dominissini et al [168].

6.2.5.10 Protein complex membership

We classified each protein as a protein complex member if it was a subunit of at least one annotated protein complex in the CORUM [169] mammalian protein complex database (release version 02.07.2017).

6.2.5.11 Protein post-translational modification

We downloaded protein acetylation, methylation, phosphorylation, SUMOylation, and ubiquitination data from the Phosphosite database (release version 02.05.2018) [156] and calculated the number of modification sites per modification type for each protein. For proteins whose modification information was not available in the downloaded dataset, we assigned 0 instead. The covariate for each of these features was defined as the log2 of the number of modifications plus 1 (pseudocount).

6.2.5.12 RNA binding protein targets

We classified transcripts as targets of 112 RBPs if they contained at least one peak in the eCLIP dataset of Van Nostrand et al [12] as processed earlier [170].

6.2.5.13 miRNA targets

Many miRNAs in the miRTarBase database (Chou et al, 2018) have very few reported targets, leading to no improvement explained variance. Therefore, we filtered for the miRNAs which have at least 200 experimentally validated target genes in our dataset and classified the genes accordingly as targets for these miRNAs. Due to high collinearity between binding evidences of different miRNAs, we applied PCA to the 11,575 $\times$ 296 binding evidence matrix and selected as features the 150 first principal components that explained 95

6.2.5.14 De novo motif Identification

Similar to Eser et al [171], de novo motif identification was performed separately for 5' UTR, CDS, and 3' UTR regions by using a linear mixed model in which the effect of each individual k-mer on the median PTR ratios across tissues was assessed while controlling

for the effect of the other k-mers (random effects) and region length and region GC percent (fixed effects). In order to identify k-mers which display more tissue-specific effects, the same approach was applied to tissue-specific median-centered (median being taken per gene across tissues) log-transformed PTR ratios. The model was fitted with the GEMMA software [172]. Motif search was executed for k-mers ranging from 3 to 8, and the P-values were adjusted for multiple testing with Benjamini–Hochberg's false discovery rate computed across the P-values of all tissues jointly. Significant motifs at $FDR < 0.1$ were subsequently manually assembled based on partial overlap.

6.2.6 Developed models

6.2.6.1 Interpretable multivariate linear model

The multivariate linear model we used for quantifying the tissue-specific effects of the considered sequence elements is:

$$y_{ij} = {\beta_j}^0 + {\boldsymbol{x}_i}^T \beta_j + \epsilon_{ij} \tag{6.1}$$

$$\hat{\beta}_j = \arg\min_{\beta_j}({\|X_j\beta_j - Y_j\|_2}^2) \tag{6.2}$$

where y_{ij} is the tissue-specific PTR ratio (log10) of gene i and tissue j, and ${x_i}^T$ is the i th row of the matrix X of sequence feature predictors which contains:

- 61 features for individual codon frequencies (in log2 scale)
- 36 features for Kozak sequence position–nucleotide pairs
- 39 features for stop-codon-context position–nucleotide pairs
- three features for CDS, 5' UTR and 3' UTR lengths (in log2 scale)
- three features for CDS, 5' UTR and 3' UTR GC percentages
- 20 features for 3' UTR motifs
- 25 features for 5' UTR motifs (including upstream AUG)
- three features for CDS amino acid motifs
- six features for linear protein motifs
- three features for 5' UTR folding energy
- two features for codon pair bias
- one feature for PEST motifs
- one feature for protein isoelectric point

- one feature for protein N-terminal hydrophobicity

The intercept ${\beta_j}^0$ and the vector β_j of the model coefficients for the j th tissue were estimated by ordinary least squares, that is, minimizing the squared of the errors ϵ_{ij}.

To predict the tissue-independent effects of the sequence features, we considered the simple linear regression model:

$$y_{ij} = {\beta_j}^0 + {x_i}^T \beta + \epsilon_{ij} \tag{6.3}$$

where the intercepts ${\beta_j}^0$ varied by tissue while the coefficients of the sequence features (the vector β) were kept equal across tissues. The intercept ${\beta_j}^0$ and the vector β of the model coefficients were estimated by ordinary least squares, i.e., minimizing the squared of the errors ϵ_{ij}. The explained variance R^2 of the PTR ratio by the sequence features was obtained by 10-fold cross-validation where in each fold the held-out data were used to have the PTR ratio predictions based on the linear regression model fit obtained from the remaining nine partitions.

6.2.6.2 Extended regularized model

We used elastic net for the model with extended set of features where the covariate matrix X consists of both the sequence based and the experimentally characterized regulators of the post-transcriptional regulation. Elastic net is a penalized linear regression model which applies both L1 and L2 regularization on model parameters in the loss function:

$$\hat{\beta_j} = \arg\min_{\beta_j}(\|X_j\beta_j - Y_j\|_2^{\ 2} + \lambda(\frac{(1-\alpha)}{2}\|\beta_j\|_2^{\ 2} + \alpha\|\beta_j\|_1)) \tag{6.4}$$

Elastic net is the same as lasso when $\alpha = 1$ and is the same as ridge regression when $\alpha = 0$. For other values of α, the penalty on the model parameters interpolates between the L1 norm of β and the squared L2 norm of β. We used grid search with 10 fold cross-validation to find the optimal values for the λ and α hyperparameters (R caret package [173]).

6.2.7 Processing and modelling of external data sets

6.2.7.1 mRNA half-life

To estimate codon effects on mRNA half-life, for K562 cells we first called a major isoform as the highest expressed isoforms of Gencode v24 coding transcripts in the total RNA samples of Schwalb and colleagues [95] according to Kallisto [164]. The half-life was estimated as the ratio of 5 min labeled TT-seq sample over total RNA-Seq sample (two replicates) after correcting library size with spike-in. For HeLa Tet-off cells, we used the isoforms reported by the authors [93], and for HEK293 cells [94], we used the dominant major isoforms across the 29 tissues we have inspected. We then fitted a linear model with log10 mRNA half-life as response variable against log2 frequency of codons with region length and GC content of 5' UTR, CDS, and 3' UTR as further covariates.

6.2.7.2 Protein half-life

Protein half-lives for B cells, NK cells, hepatocytes, and monocytes [98] were identified only by gene name and not by isoforms, and those for HeLa cells [97] by gene names and UniProt protein identifiers. We therefore mapped our transcript isoforms to these datasets by gene identifiers. We estimated the associations of the sequence features with protein half-life by multivariate regression where the response variable was the cell-type-specific log10-transformed protein half-life.

6.2.7.3 Independently matched transcriptome–proteome dataset

We used data from Kremer et al [174]. As originally reported, these data showed strong technical effects. To be on the safe side, we restricted the analysis to six samples (sample IDs: 65126, 73804, 78661, 80248, 80254, and 81273) that belonged to the same cluster.

6.2.8 Additional analyses

6.2.8.1 Motif analysis

Tissue-specific motif effects:

In the design matrix, all of the de novo identified motifs except 'AUG' and 'AAUAAA' are encoded as the number of motif sites in the sequence of the mRNA region (i.e., 5' UTR, CDS, 3 'UTR). 'AUG' and 'AAUAAA' are encoded as binary, hence whether the motif is available in 5' UTR and 3' UTR regions, respectively. The tissue-specific effect of the motif is assessed by fitting all sequence features considered jointly in the linear model, with the tissue-specific PTR ratios being the response variables.

Gene ontology enrichment:

Enrichment for gene ontology categories [175] as of January 21, 2016, was performed using the Fisher exact test and corrected for multiple testing using the Benjamini–Hochberg correction.

Systematic motif search in RNA binding protein databases:

Motif consensus sequences are searched in the RNA binding protein database ATtRACT [141] by using the database Web interface at https://attract.cnic.es/searchmotif. The RNA binding protein with the highest quality score, if any, was reported as binding candidate of the motif.

Motif 1 nucleotide mismatch logos:

The sequences of each motif instance with at most 1 nucleotide mismatch were obtained from transcript mRNA sequences. The logos were created with R ggseqlogo package.

Motif conservation analysis:

Phylogenetic conservation scores for human annotation hg38 (phastConst100way from http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phastCons100way), which reports conservation across 99 vertebrates aligned to the human genome, were downloaded, and the conservation scores per nucleotide were extracted for each of the motif instances without any mismatch. The significance of the enrichment scores at the motif sites compared

to 10 nucleotides flanking regions was tested with a one-sided Wilcoxon test across all consensus motif occurrences in the given mRNA region (i.e., 5' UTR or 3' UTR).

6.2.8.2 RNA binding protein across-tissue covariation with target genes

Among 112 RBPs whose binding evidences were used in our integrated model, 64 of them were expressed in at least 15 tissues with an mRNA level standard deviation across tissues ¿ 0.1. In order to see across-tissue expression covariation between these RBPs and their target genes, for each RBP we calculated the Spearman's rho between its protein level expression and mRNA levels in other genes (again expressed in at least 15 tissues and with mRNA ratio standard deviation ¿ 0.1). The significance of the correlation coefficient distribution difference between target and non-target genes was assessed with two-sided Wilcoxon test.

6.2.8.3 Coding sequence 5' end codon frequency analysis

We considered the 10,778 transcripts with CDS length greater than 460 nucleotides. Starting from the second codon, the log2 frequencies of 61 coding codons are calculated in each of the 11 non-overlapping 15-codon-long windows. The frequency values are centered per codon across windows. In order to compare the effect of twofold codon frequency increase in the first window (codons from 2 to 16) versus the effect of the twofold codon frequency increase in the rest of the coding sequence, the codon frequencies of the whole coding sequence are replaced by the respective frequency values in the global interpretable model.

6.2.8.4 Explained variance of protein levels by mRNA levels and sequence features

We performed a linear regression of log-transformed protein levels against the complete log-transformed mRNA levels across all tissues and the sequence features. We also performed a linear regression of log-transformed protein levels against the complete log-transformed mRNA levels across all tissues, the sequence features, and the non-sequence features. Explained variance was reported as adjusted R^2.

6.2.8.5 Effect of amino acids on PTR ratio

To estimate the effect of doubling the frequency of an amino acid in any gene on its log10 PTR ratio, we performed a modified version of the regression defined by equation 1 (for tissue-specific effects) and a modified version of the regression defined by equation 2 (general effect), whereby the amino acid log2 frequencies were considered as features instead of the codon log2 frequencies.

6.2.8.6 PTR-AI

The tissue-specific protein-to-mRNA ratio index (tissue-specific PTR-AI) of a codon is computed as ten to the power of the estimated coefficient of the log2 frequency of this codon in the regression described by equation 1, where j is the index of the tissue of interest. It is an estimation of the fold-change on PTR ratio for a specific tissue obtained if one would double the frequency of this codon in any gene. The protein-to-mRNA ratio index (PTR-AI) of a codon is computed as ten to the power of the estimated coefficient of the codon log2 frequency in the regression described by equation 2. It is an estimation of the fold-change on PTR ratio in any tissue obtained if one would double the frequency of this codon in any gene.

6.2.8.7 Codon decoding time and average amino acid decoding time

Codon decoding times for 16 human ribosome profiling datasets were obtained from RUST values [23]. We estimated decoding time using the RUST ratio defined by the RUST A-site values over the RUST expected value (personal communication with Patrick O'Connor). We also included decoding times in the HEK293 cell line estimated by Dana and Tuller [96]. To estimate the average decoding times per codon, for each dataset i we converted the decoding times into z-scores (i.e., subtracting the mean and dividing by the standard deviation) and then used the median z-score per codon across datasets as the average normalized decoding time of the codon. The average amino acid decoding time was defined as the average codon decoding time per amino acid weighted by the codon genomic frequency.

7 Discussion

The methodology, results and figures presented in this section are part of the manuscript "Quantification and discovery of sequence determinants of protein per mRNA amount in 29 human tissues" from Eraslan and Wang et al. 2019 [1]

In this thesis I presented two studies in which I designed, implemented and executed several statistical computational analyses to gain better insight about the post-transcriptional regulation events. I displayed that through integrated analyses of high-throughput omics data sets, we are able to identify new regulatory motifs and generate novel hypotheses about the regulation mechanisms. My analyses provide a data driven system overview by connecting multiple components of the post-transcriptional regulation. Performance of our predictive models which were developed based on the known and identified regulatory elements help us to have an idea about where we stand in the way of understanding genotype-phenotype relationships.

However, I should note that even though our high-throughput data driven approaches of provide a first step in generating new hypothesis about the mRNA and protein relationship, they should be taken with a grain of salt due to the various experimental and statistical constraints. First of all, the potentially large number of transcript isoforms that can be generated from the same gene via alternative splicing presents an important complication for the comparison of protein and mRNA levels. The possibility that peptide levels may be compared to splice isoforms that do not contain the respective peptide sequence may distort the protein/mRNA correlation. The emergence of RNA-seq has greatly improved our ability to account for splicing effects compared to earlier transcript profiling methods. This is particularly relevant when evaluating protein/mRNA correlation for a single gene, as the transcript isoforms expressed may change across conditions. An important complication for measuring translation again points to alternative splicing because the efficiency of translation of different isoforms can vary greatly - e.g., by including or excluding uORFs in front of the coding sequence [176].

Our multivariate regression analysis estimated the contribution within and across tissues of 18 sequence feature groups representing 204 post-transcriptional regulatory elements. Altogether the model predicts the PTR ratio of individual genes at a median precision of 3.2-fold from sequence alone, while the PTR ratio spans about 200-fold across 80% of the genes. For most known regulatory elements, the estimated effects were consistent with the literature, such as the effects of the secondary structures in the upstream CDS, upstream AUGs, individual nucleotides in the start and stop codon context, and de novo identified 3' UTR motifs AATAAA, TATTTAT, and TGTAAATA, providing support to the functional interpretability of the model. Moreover, this analysis led to the identification of novel candidate regulatory elements in 5' UTR and 3' UTR, whose effects are estimated to be in the range of well-known canonical motifs.

Follow-up experiments provided initial functional support for these motifs. Moreover, our extended model comprising 269 additional experimentally characterized sequence features indicates that post-translational protein modifications substantially contribute to PTR ratios and would constitute an important set of features for modeling in more detail in the future.

There are limitations to this approach that should be noted. The model is additive on the logarithmic scale. However, regulatory elements likely function depending on the sequence context, the presence of other regulatory elements, and their respective distance along the transcript but also in space. Given the amount of variations across genes, such non-additive effects are very hard to be fitted. Hence, the effect of mutating a particular sequence element on a given gene may differ from the expected effect estimated by the linear model. Also, the conserved sequence elements we found associated with PTR ratio may be functional but actually play a different role because high PTR ratios correlate with other selected traits such as high mRNA levels. Experiments will help in resolving these questions. Nonetheless, our study provides interesting conserved sequence elements to follow up with mechanistic studies. We have also performed matches to the ATtRACT database, as an indication of possible RBP recognizing these motifs. ATtRACT matches, even with the highest scores, can be lenient. Also, this database suffers from the general poor charting of RBP binding sites. As a result, we shall take these matches as indicative and with caution. Another limitation is that most tissues investigated have been obtained from different donors [2]. While it is reasonable to expect that tissue-specific effects dominate the differential expression signal between these samples, one cannot exclude donor-specific effects as well.

Our regression approach led to a new codon metric, PTR-AI, for protein-to-mRNA ratio adaptation index, which estimates the effect of doubling the frequency of a codon in a gene on its protein-to-mRNA ratio. Using PTR-AI, codons, which inherently encode amino acids and synonymous codon usage, are the lead explanatory variable explaining about 16% of the PTR ratio variance across genes almost in every tissue we inspected. Amino acid frequency and synonymous codon usage affect PTR ratios via various mechanisms. Amino acid identity affects translation [112, 45] and protein half-life [113, 97] while synonymous codon usage influences translation efficiency due to variation in the translation elongation rates of different codons [116, 44, 22, 109, 110, 111, 45]. Highly expressed genes contain relatively high proportions of codons recognized by abundant tRNAs with efficient codon–anticodon base-pairing. Based on this observation, several codon optimality metrics have been suggested [177, 178, 179, 51]. However, all of these rely on some assumptions and simplifications, such as the codon adaptation index defining a set of highly expressed genes as a reference set or the tRNA adaptation index overlooking the supply and demand relationship for charged tRNAs. PTR-AI does not correlate well with codon genomic frequency or tAI adaptiveness, whereas it does correlate well with the codon decoding times estimated from several ribosome profiling datasets. Furthermore, we have shown that PTR-AI also captures the effects of amino acids on protein stability. Consequently, we suggest that PTR-AI is a more reliable codon optimality metric than previous metrics.

Our findings do not support the hypothesis of tissue-specific codon optimality. It has been suggested that there is tissue-specific codon-mediated translational control due to differential synonymous codon usage in human tissue-specific genes, which correlates with varying tRNA expression among different tissues [114, 115]. However, other studies found no evidence for optimization of translational efficiency by cell-type-specific codon usage in human tissues [180, 181]. Our tissue-specific PTR-AIs, which are estimated by fitting our model separately for each tissue, do not display high variation across tissues. This result is coherent with negligible tissue-specific enrichments of expressed codons in human transcriptomes, showing that tissue-specific expression is neither due to the transcription nor due to the translation of genes with particular codon contents. Further corroborating this finding, genes with high-effect codons tended both to have a high median level of protein expression and to be ubiquitously expressed. These genes were enriched for housekeeping functions. A possible explanation of these findings is that housekeeping genes have evolved for optimal coding sequence to reach high protein expression levels. Because of the ubiquitous role of housekeeping genes, their codon content in turn constrains the pool of tRNA to be rather constant across tissues. These explanations are consistent with the recent massive genomic editing experiment results, which show that codon bias of highly expressed genes maintains the efficiency of global protein translation in the cell [182]. The lack of tissue specificity of PTR-AIs we reported here does not contradict the differential tRNA pool regulation between proliferative and differentiating cells [183], since our tissues are essentially constituted of non-proliferative cells.

In every tissue investigated, protein-to-mRNA ratios were higher for genes with high mRNA expression levels, leading to an approximately quadratic relationship between protein and mRNA levels across genes [2] and a larger dynamic range of expression among proteins than mRNAs. Our model partially explains this apparent amplification from sequence features, thereby showing that high protein expression levels are reached because of high mRNA levels and because of genetically encoded elements favoring the synthesis and stability of proteins. Regulatory elements that affect both the mRNA levels and protein-per-mRNA copy numbers could further contribute to this apparent amplification. Codons are known to play such a dual role since they affect translation on the one hand, and mRNA stability on the other hand. The mechanistic basis for these cross-talks between translation and mRNA stability is not fully understood. It is possible that regression approaches similar to those employed by us could help in revealing further sequence elements acting on both levels. A similar super-linear relationship had been reported before for the unicellular eukaryotes in baker's yeast [184] and fission yeast [39], which appears to be absent in the prokaryote E. coli, respecting which mRNA and protein levels across genes obey a nearly linear relationship [84]. Prokaryotic transcription and translation are coupled processes, which do not allow post-transcriptional regulation to have an effective role in determining steady-state protein levels. In contrast, these two processes are highly uncoupled and have specialized mechanisms in eukaryotes, which are favored by the compartmentalization of eukaryotic cells. We suggest that the uncoupling of transcription and translation underlies a fundamental difference in the relationship between protein and mRNA levels across genes in eukaryotes compared to prokaryotes

and may allow protein copy numbers of eukaryotic cells to span a much larger dynamic range. Further matched transcriptome and proteome datasets for a larger range of prokaryotes would help to support this model.

A comprehensive post-transcriptional regulatory code is important for interpreting regulatory genetic variations in personal genomes and in genetic engineering for biotechnological or gene therapy applications. Our study provides an important contribution by modeling codon effects, identifying novel sequence elements with potential function, and giving a framework for quantifying and assessing the role of new elements on protein-per-mRNA copy number. In the future, we expect further approaches including the analysis and integration of perturbation-based data and the mapping of post-translational regulatory elements in order to complement and refine the present analysis.

A Appendix: Additional Figures

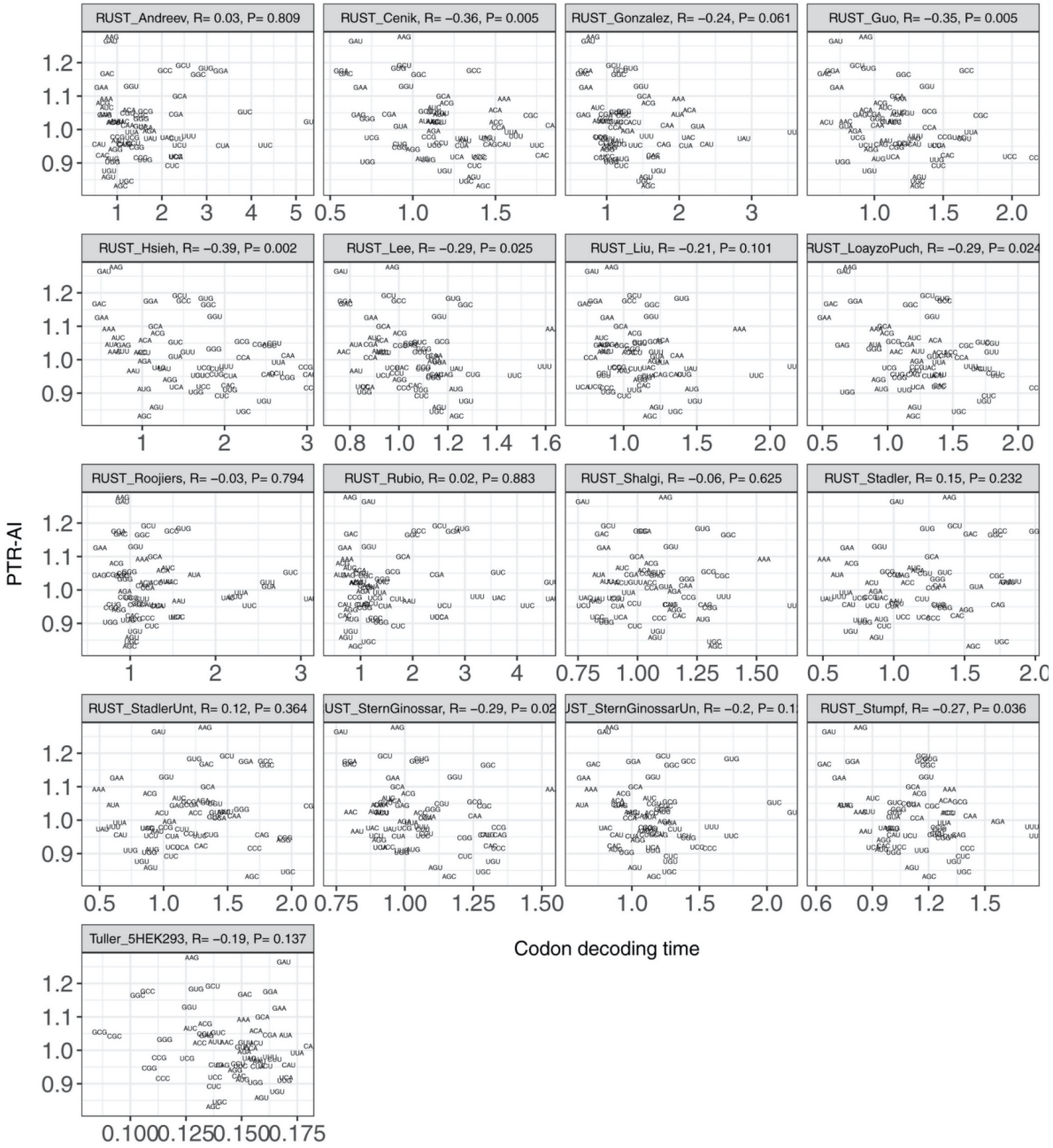


Figure A.1: Correlations between PTR-AI and the codon decoding times obtained from 17 independent ribosome profiling data sets Median PTR-AI across tissues negatively correlates with expected codon decoding times in 17 ribosome profiling datasets [23].

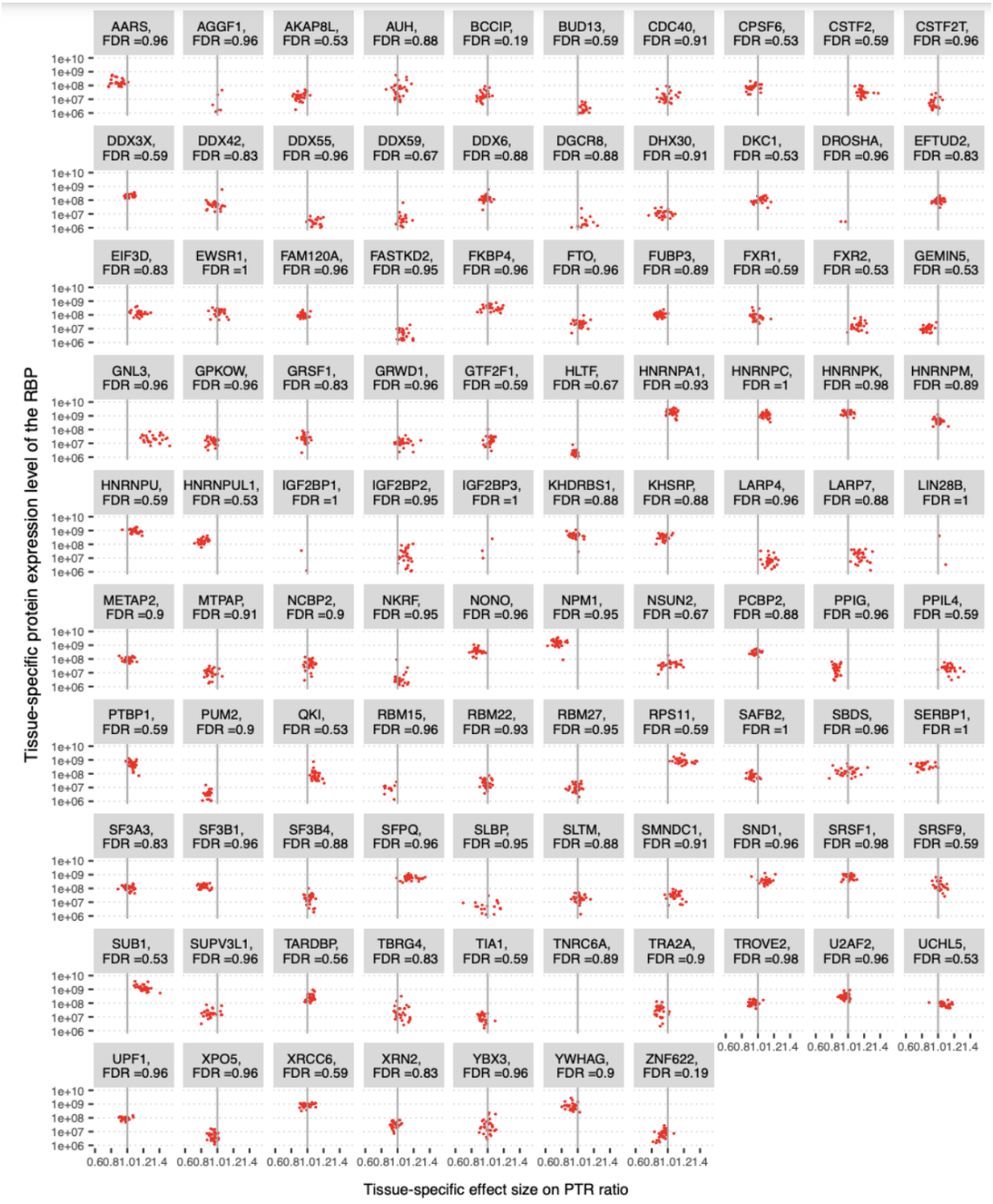
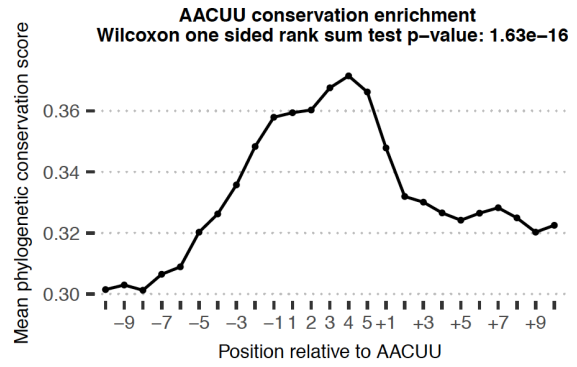
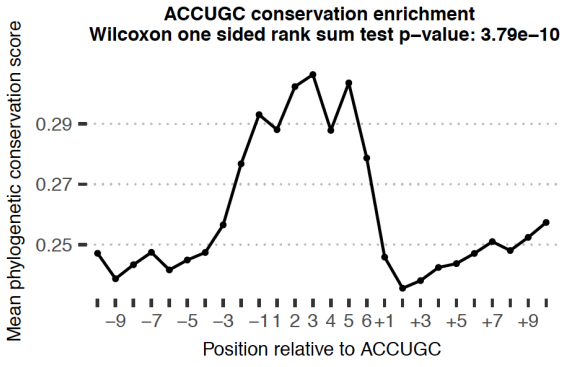


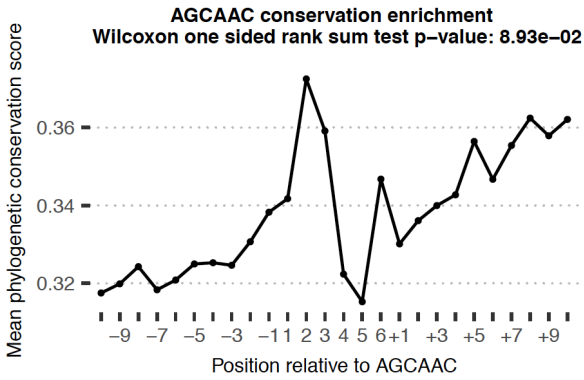
Figure A.2: Correlations between PTR-AI and the codon decoding times obtained from 17 independent ribosome profiling data sets Our data set covers matched transcriptome and proteome measurements of 97 out of 112 RBPs whose target genes were detected by Van Nostrand et al. [12] and 81 of these RBPs were measured in all 29 tissues. The tissue-specific effect sizes of the RBP binding evidences in the linear model did not significantly correlate with the tissue-specific RBP expression levels.


AACUU conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.63e−16
Mean phylogenetic conservation score
0.36
0.34
0.32
0.30
−9 −7 −5 −3 −11 2 3 4 5 +1 +3 +5 +7 +9
Position relative to AACUU

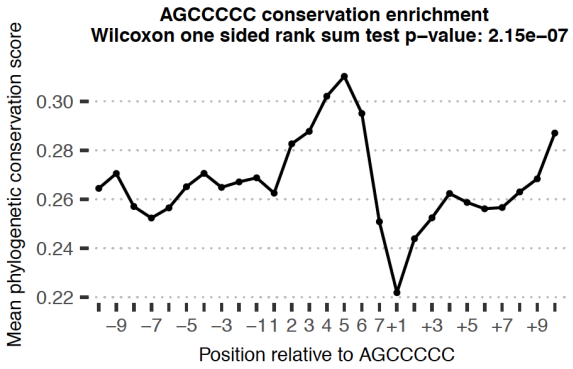
(a)


ACCUGC conservation enrichment
Wilcoxon one sided rank sum test p−value: 3.79e−10
Mean phylogenetic conservation score
0.29
0.27
0.25
−9 −7 −5 −3 −11 2 3 4 5 6 +1 +3 +5 +7 +9
Position relative to ACCUGC

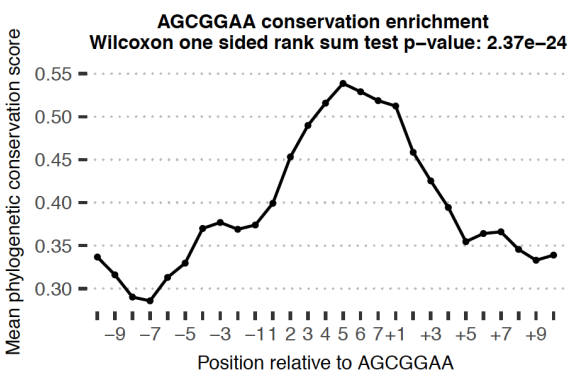
(b)


AGCAAC conservation enrichment
Wilcoxon one sided rank sum test p−value: 8.93e−02
Mean phylogenetic conservation score
0.36
0.34
0.32
−9 −7 −5 −3 −11 2 3 4 5 6 +1 +3 +5 +7 +9
Position relative to AGCAAC

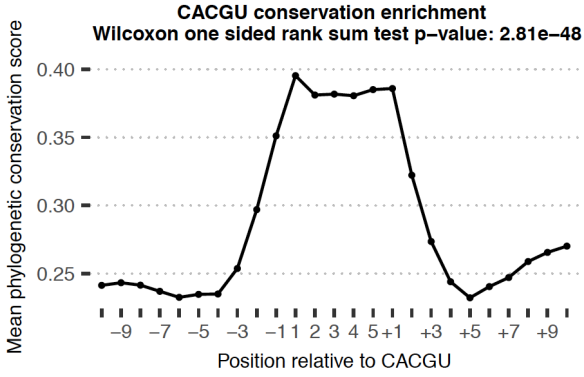
(c)


AGCCCCC conservation enrichment
Wilcoxon one sided rank sum test p−value: 2.15e−07
Mean phylogenetic conservation score
0.30
0.28
0.26
0.24
0.22
−9 −7 −5 −3 −11 2 3 4 5 6 7 +1 +3 +5 +7 +9
Position relative to AGCCCCC

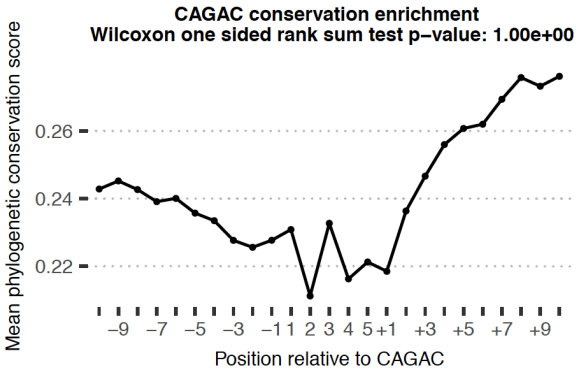
(d)


AGCGGAA conservation enrichment
Wilcoxon one sided rank sum test p−value: 2.37e−24
Mean phylogenetic conservation score
0.55
0.50
0.45
0.40
0.35
0.30
−9 −7 −5 −3 −11 2 3 4 5 6 7 +1 +3 +5 +7 +9
Position relative to AGCGGAA

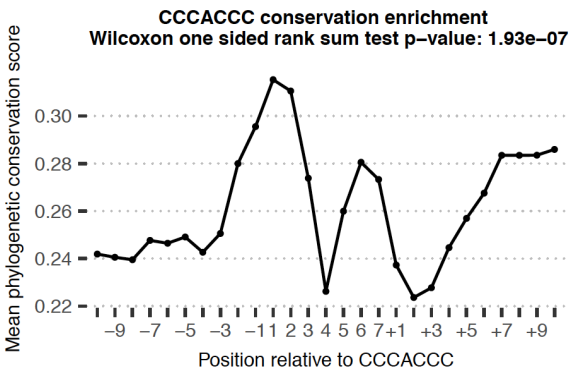
(e)


CACGU conservation enrichment
Wilcoxon one sided rank sum test p−value: 2.81e−48
Mean phylogenetic conservation score
0.40
0.35
0.30
0.25
−9 −7 −5 −3 −11 2 3 4 5 +1 +3 +5 +7 +9
Position relative to CACGU

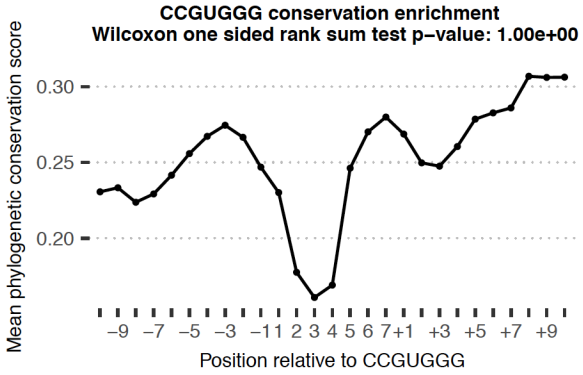
(f)


CAGAC conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.00e+00
Mean phylogenetic conservation score
0.26
0.24
0.22
−9 −7 −5 −3 −11 2 3 4 5 +1 +3 +5 +7 +9
Position relative to CAGAC

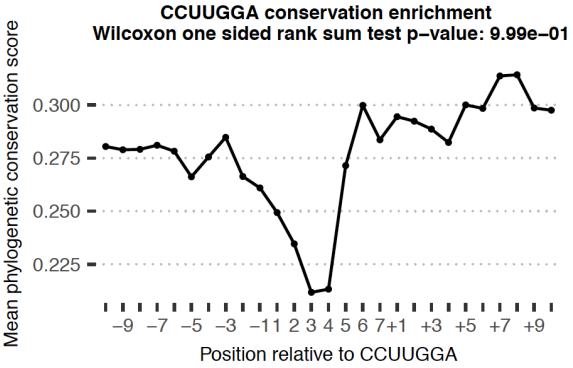
(g)


CCCACCC conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.93e−07
Mean phylogenetic conservation score
0.30
0.28
0.26
0.24
0.22
−9 −7 −5 −3 −11 2 3 4 5 6 7 +1 +3 +5 +7 +9
Position relative to CCCACCC

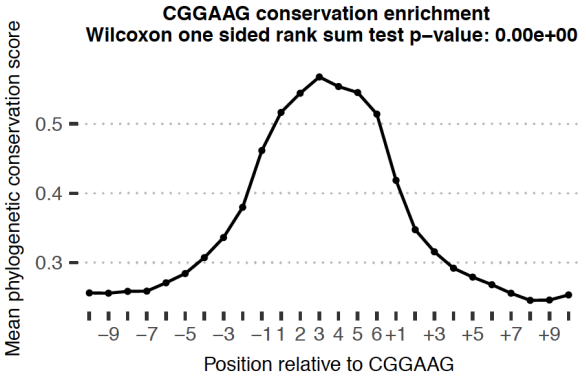
(h)


CCGUGGG conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.00e+00
Mean phylogenetic conservation score
0.30
0.25
0.20
−9 −7 −5 −3 −11 2 3 4 5 6 7 +1 +3 +5 +7 +9
Position relative to CCGUGGG

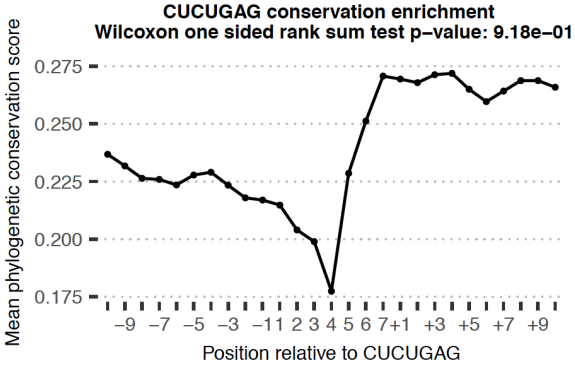
(i)


CCUUGGA conservation enrichment
Wilcoxon one sided rank sum test p−value: 9.99e−01
Mean phylogenetic conservation score
0.300
0.275
0.250
0.225
−9 −7 −5 −3 −11 2 3 4 5 6 7 +1 +3 +5 +7 +9
Position relative to CCUUGGA

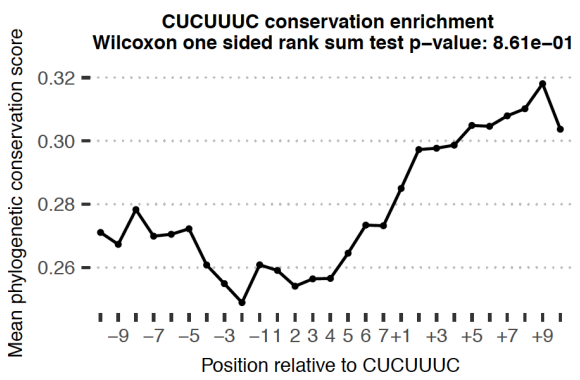
(j)


CGGAAG conservation enrichment
Wilcoxon one sided rank sum test p−value: 0.00e+00
Mean phylogenetic conservation score
0.5
0.4
0.3
−9 −7 −5 −3 −11 2 3 4 5 6 +1 +3 +5 +7 +9
Position relative to CGGAAG

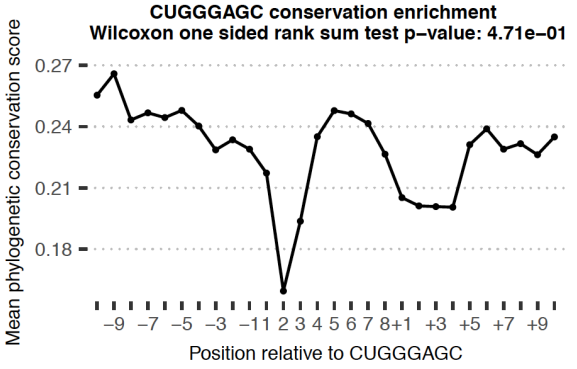
(k)


CUCUGAG conservation enrichment
Wilcoxon one sided rank sum test p−value: 9.18e−01
Mean phylogenetic conservation score
0.275
0.250
0.225
0.200
0.175
−9 −7 −5 −3 −11 2 3 4 5 6 7 +1 +3 +5 +7 +9
Position relative to CUCUGAG

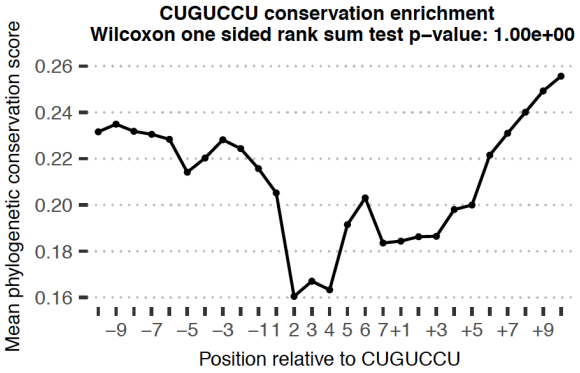
(l)


CUCUUUC conservation enrichment
Wilcoxon one sided rank sum test p−value: 8.61e−01
Mean phylogenetic conservation score
0.32
0.30
0.28
0.26
−9 −7 −5 −3 −11 2 3 4 5 6 7+1 +3 +5 +7 +9
Position relative to CUCUUUC

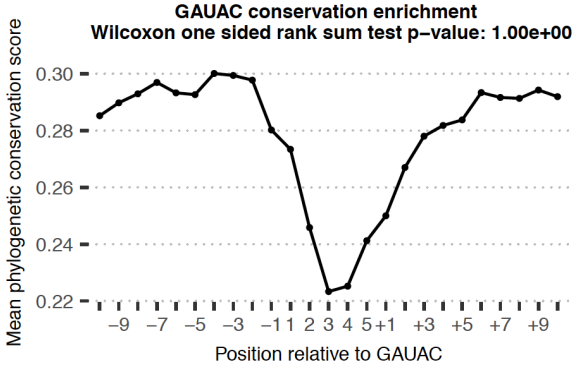
(m)


CUGGGAGC conservation enrichment
Wilcoxon one sided rank sum test p−value: 4.71e−01
Mean phylogenetic conservation score
0.27
0.24
0.21
0.18
−9 −7 −5 −3 −11 2 3 4 5 6 7 8+1 +3 +5 +7 +9
Position relative to CUGGGAGC

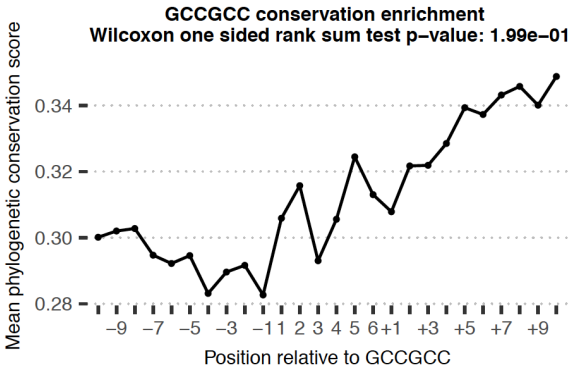
(n)


CUGUCCU conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.00e+00
Mean phylogenetic conservation score
0.26
0.24
0.22
0.20
0.18
0.16
−9 −7 −5 −3 −11 2 3 4 5 6 7+1 +3 +5 +7 +9
Position relative to CUGUCCU

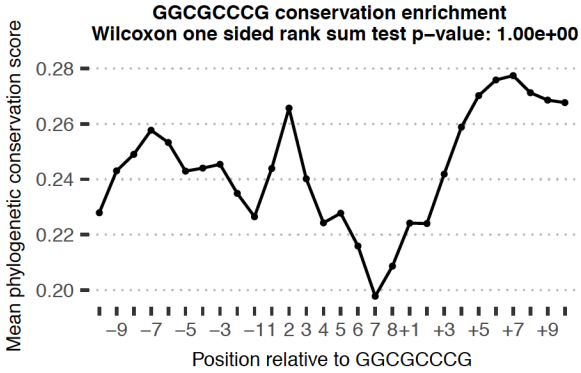
(o)


GAUAC conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.00e+00
Mean phylogenetic conservation score
0.30
0.28
0.26
0.24
0.22
−9 −7 −5 −3 −11 2 3 4 5+1 +3 +5 +7 +9
Position relative to GAUAC

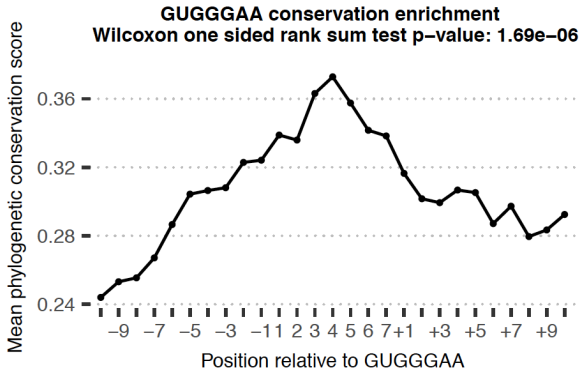
(p)


GCCGCC conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.99e−01
Mean phylogenetic conservation score
0.34
0.32
0.30
0.28
−9 −7 −5 −3 −11 2 3 4 5 6+1 +3 +5 +7 +9
Position relative to GCCGCC

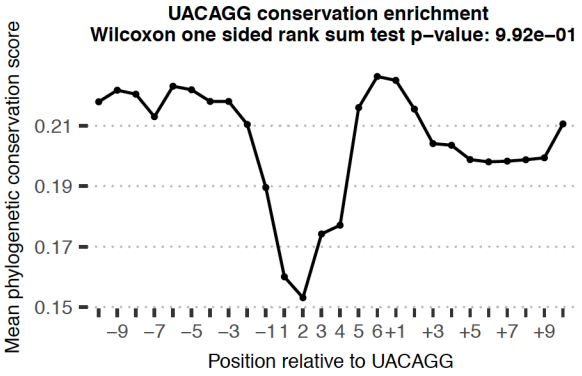
(q)


GGCGCCCG conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.00e+00
Mean phylogenetic conservation score
0.28
0.26
0.24
0.22
0.20
−9 −7 −5 −3 −11 2 3 4 5 6 7 8+1 +3 +5 +7 +9
Position relative to GGCGCCCG

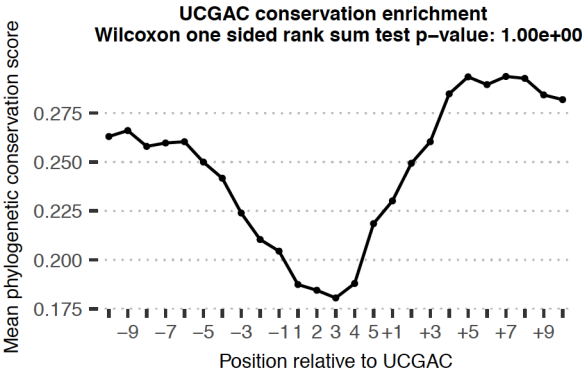
(r)



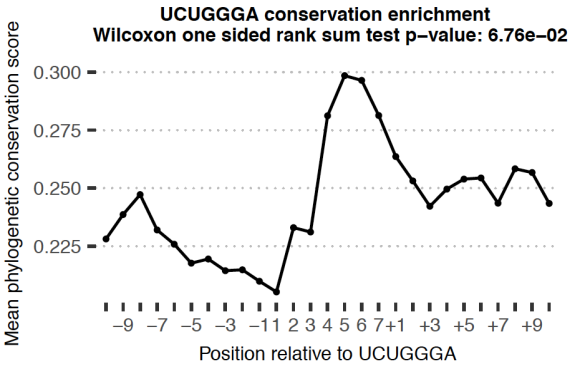
(s)



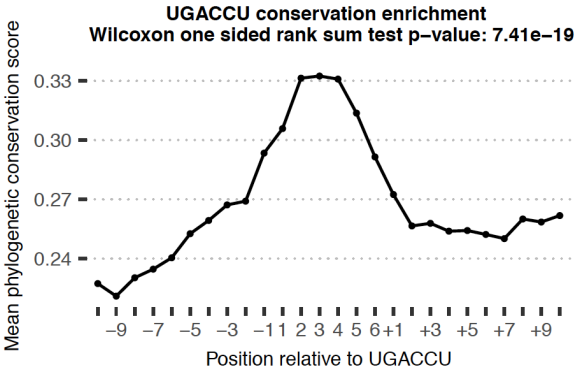
(t)



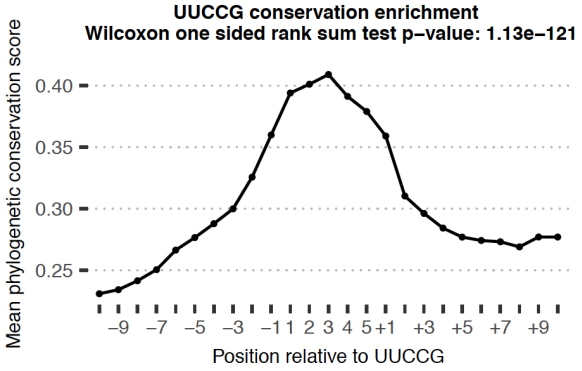
(u)



(v)

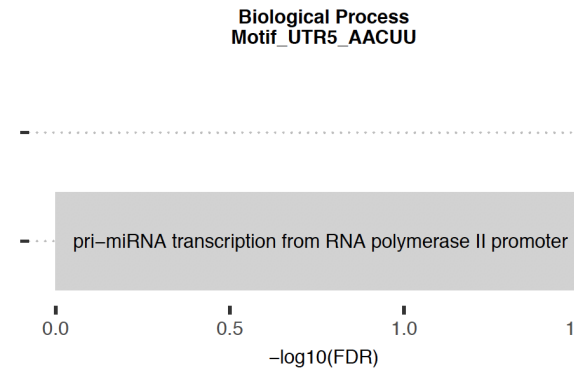


(w)

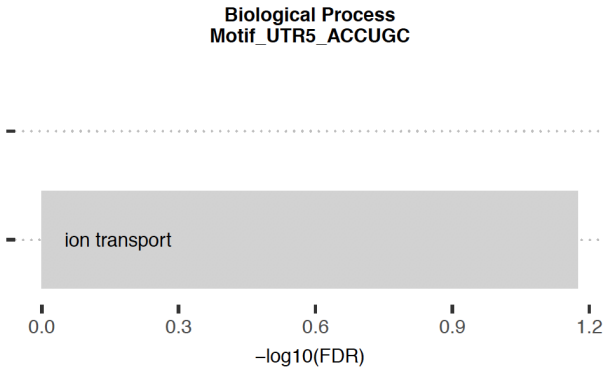


(x)

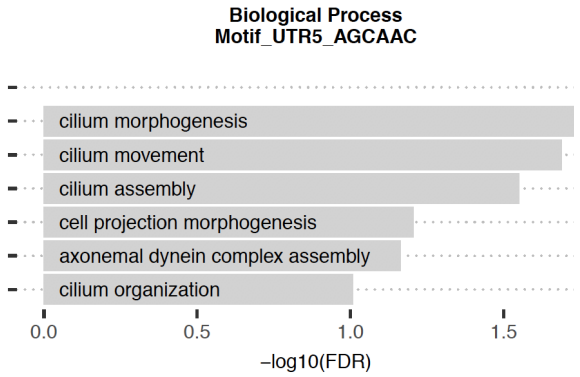
Figure A.0: 5' UTR motifs' conservation scores. Average 100-vertebrate PhastCons score (y-axis) per position relative to the exact motif match instances in 5' UTR (x-axis). P-values assess significance of the average 100-vertebrate PhastCons scores at the motif sites compared to the two 10-nucleotide flanking regions


Biological Process
Motif_UTR5_AACUU
pri−miRNA transcription from RNA polymerase II promoter
0.0
0.5
1.0
1
−log10(FDR)

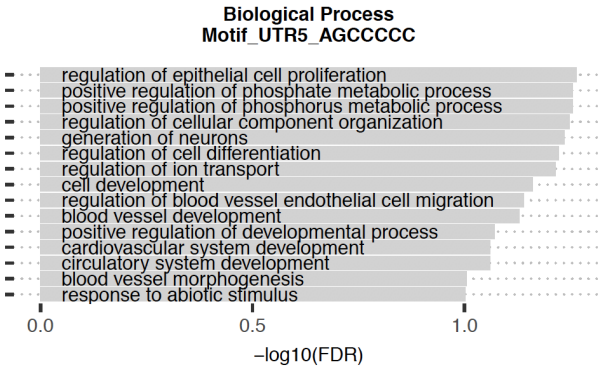
(a)


Biological Process
Motif_UTR5_ACCUGC
ion transport
0.0
0.3
0.6
0.9
1.2
−log10(FDR)

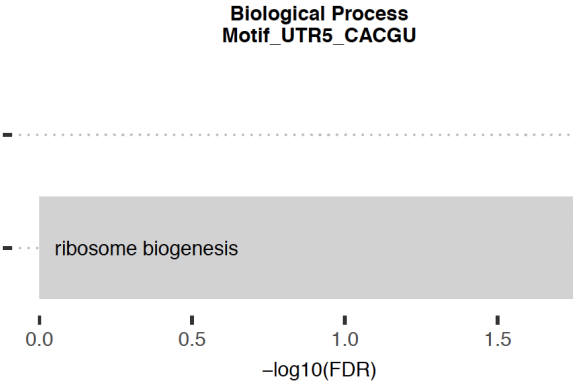
(b)


Biological Process
Motif_UTR5_AGCAAC
cilium morphogenesis
cilium movement
cilium assembly
cell projection morphogenesis
axonemal dynein complex assembly
cilium organization
0.0
0.5
1.0
1.5
−log10(FDR)

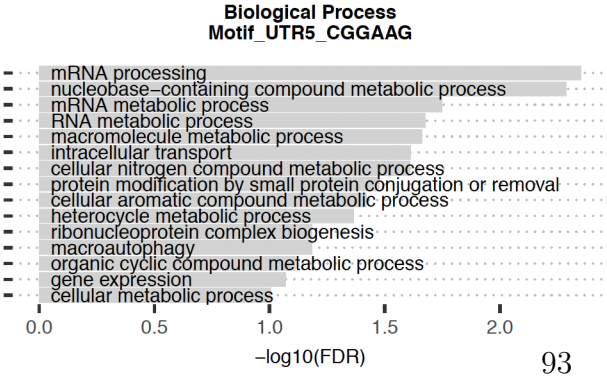
(c)


Biological Process
Motif_UTR5_AGCCCCC
regulation of epithelial cell proliferation
positive regulation of phosphate metabolic process
positive regulation of phosphorus metabolic process
regulation of cellular component organization
generation of neurons
regulation of cell differentiation
regulation of ion transport
cell development
regulation of blood vessel endothelial cell migration
blood vessel development
positive regulation of developmental process
cardiovascular system development
circulatory system development
blood vessel morphogenesis
response to abiotic stimulus
0.0
0.5
1.0
−log10(FDR)

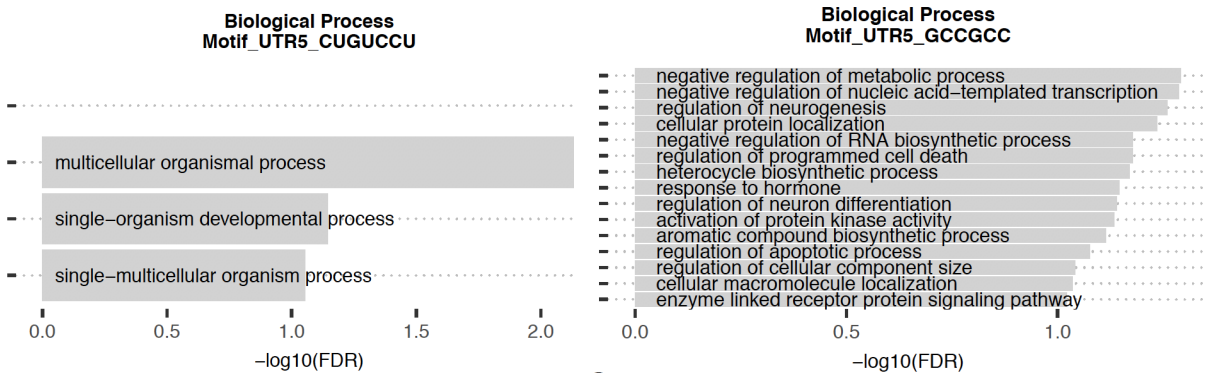
(d)


Biological Process
Motif_UTR5_CACGU
ribosome biogenesis
0.0
0.5
1.0
1.5
−log10(FDR)

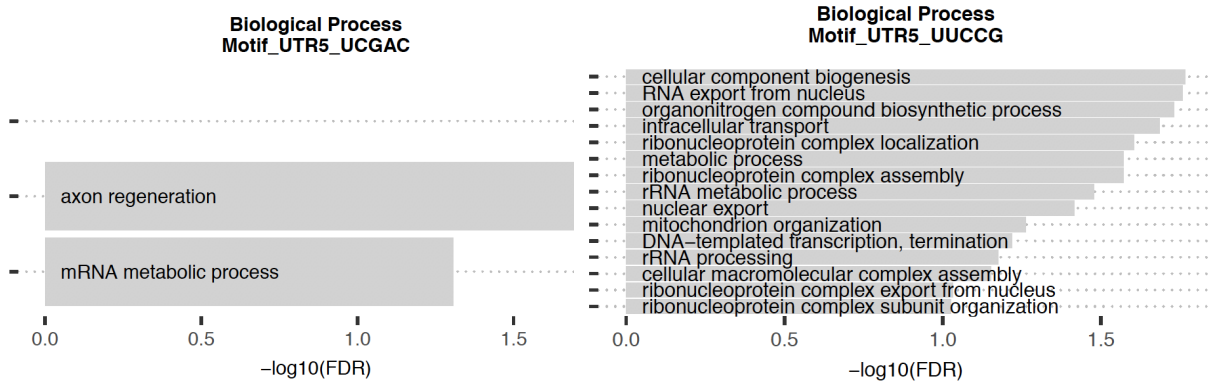
(e)


Biological Process
Motif_UTR5_CGGAAG
mRNA processing
nucleobase−containing compound metabolic process
mRNA metabolic process
RNA metabolic process
macromolecule metabolic process
intracellular transport
cellular nitrogen compound metabolic process
protein modification by small protein conjugation or removal
cellular aromatic compound metabolic process
heterocycle metabolic process
ribonucleoprotein complex biogenesis
macroautophagy
organic cyclic compound metabolic process
gene expression
cellular metabolic process
0.0
0.5
1.0
1.5
2.0
−log10(FDR)

(f)

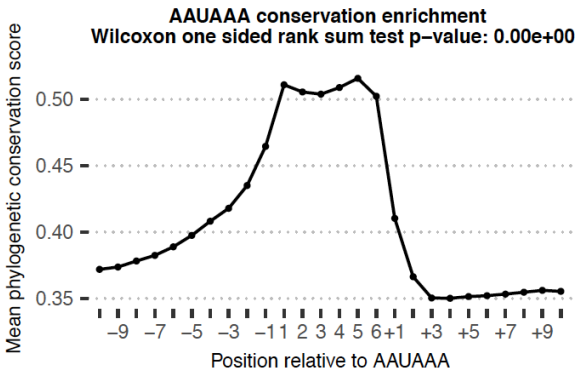


(g) (h)

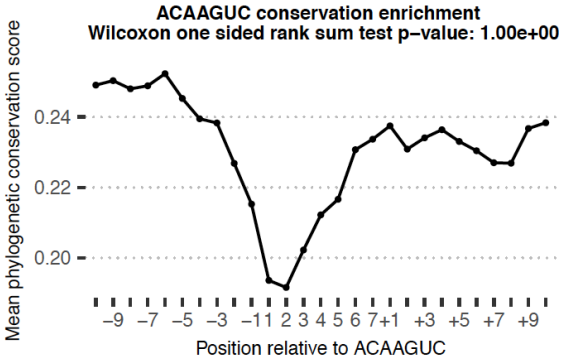


(i) (j)

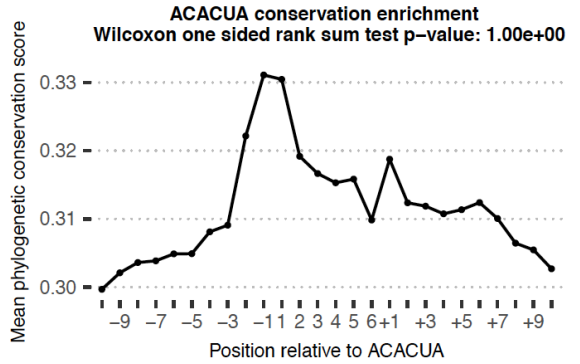
Figure A.0: Gene ontology terms enriched for genes that have the consensus 5' UTR motifs. Gene ontology terms that are enriched for set of genes that contain consensus sequences of the de-novo identified k-mers in 5' UTR that are predictive of PTR ratios.


AAUAAA conservation enrichment
Wilcoxon one sided rank sum test p−value: 0.00e+00
Mean phylogenetic conservation score
0.50
0.45
0.40
0.35
−9 −7 −5 −3 −11 2 3 4 5 6+1 +3 +5 +7 +9
Position relative to AAUAAA

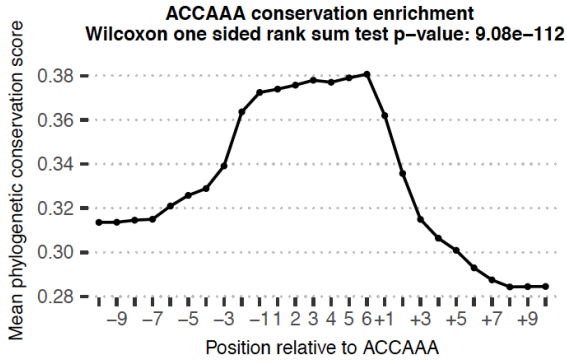
(a)


ACAAGUC conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.00e+00
Mean phylogenetic conservation score
0.24
0.22
0.20
−9 −7 −5 −3 −11 2 3 4 5 6 7+1 +3 +5 +7 +9
Position relative to ACAAGUC

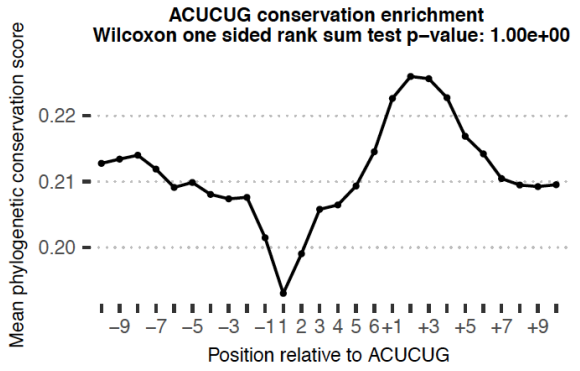
(b)


ACACUA conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.00e+00
Mean phylogenetic conservation score
0.33
0.32
0.31
0.30
−9 −7 −5 −3 −11 2 3 4 5 6+1 +3 +5 +7 +9
Position relative to ACACUA

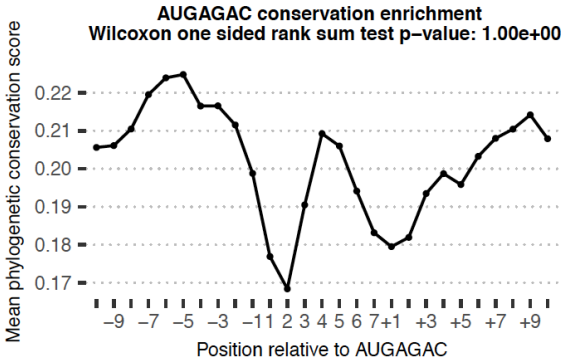
(c)


ACCAAA conservation enrichment
Wilcoxon one sided rank sum test p−value: 9.08e−112
Mean phylogenetic conservation score
0.38
0.36
0.34
0.32
0.30
0.28
−9 −7 −5 −3 −11 2 3 4 5 6+1 +3 +5 +7 +9
Position relative to ACCAAA

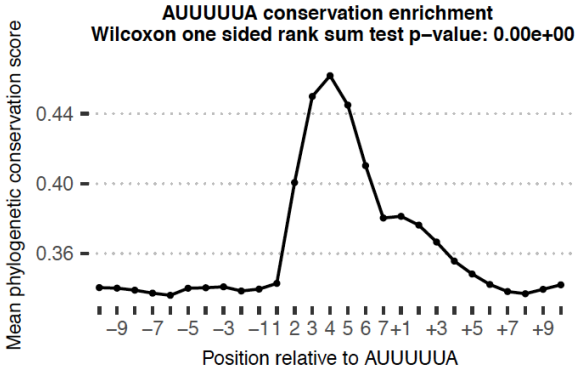
(d)


ACUCUG conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.00e+00
Mean phylogenetic conservation score
0.22
0.21
0.20
−9 −7 −5 −3 −11 2 3 4 5 6+1 +3 +5 +7 +9
Position relative to ACUCUG

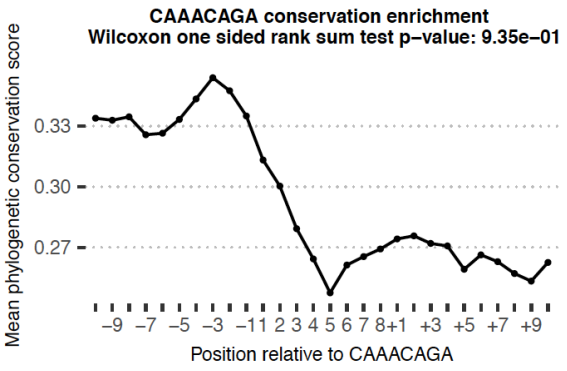
(e)


AUGAGAC conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.00e+00
Mean phylogenetic conservation score
0.22
0.21
0.20
0.19
0.18
0.17
−9 −7 −5 −3 −11 2 3 4 5 6 7+1 +3 +5 +7 +9
Position relative to AUGAGAC

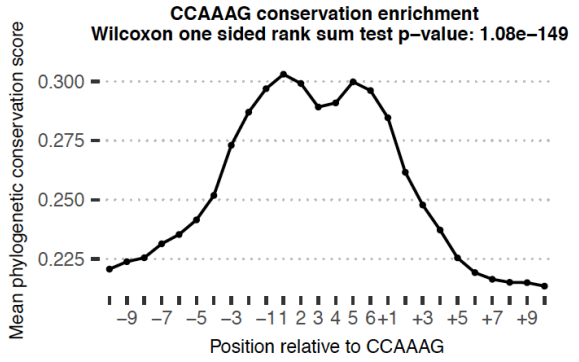
(f)


AUUUUUA conservation enrichment
Wilcoxon one sided rank sum test p−value: 0.00e+00
Mean phylogenetic conservation score
0.44
0.40
0.36
−9 −7 −5 −3 −11 2 3 4 5 6 7+1 +3 +5 +7 +9
Position relative to AUUUUUA

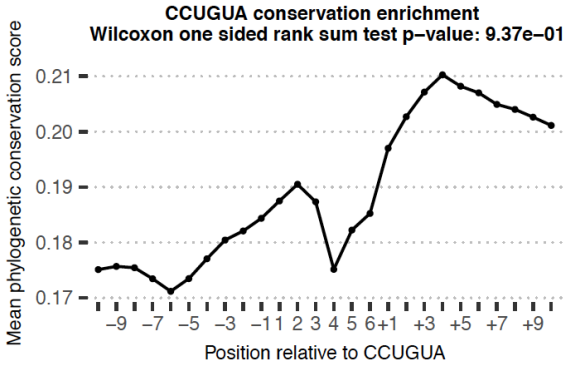
(g)


CAAACAGA conservation enrichment
Wilcoxon one sided rank sum test p−value: 9.35e−01
Mean phylogenetic conservation score
0.33
0.30
0.27
−9 −7 −5 −3 −11 2 3 4 5 6 7 8+1 +3 +5 +7 +9
Position relative to CAAACAGA

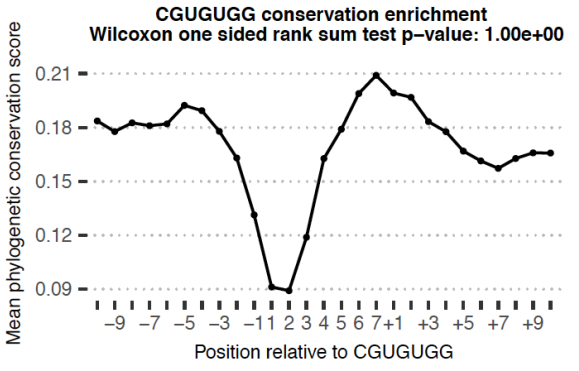
(h)


CCAAAG conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.08e−149
Mean phylogenetic conservation score
0.300
0.275
0.250
0.225
−9 −7 −5 −3 −11 2 3 4 5 6+1 +3 +5 +7 +9
Position relative to CCAAAG

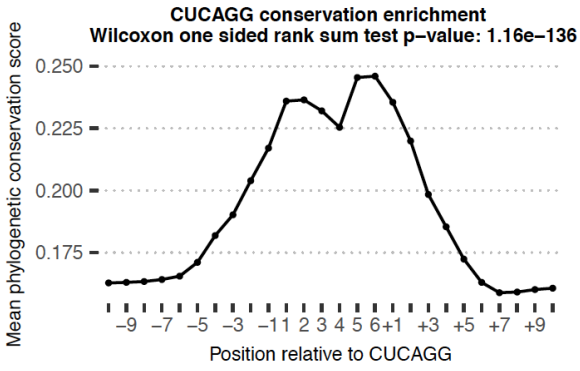
(i)


CCUGUA conservation enrichment
Wilcoxon one sided rank sum test p−value: 9.37e−01
Mean phylogenetic conservation score
0.21
0.20
0.19
0.18
0.17
−9 −7 −5 −3 −11 2 3 4 5 6+1 +3 +5 +7 +9
Position relative to CCUGUA

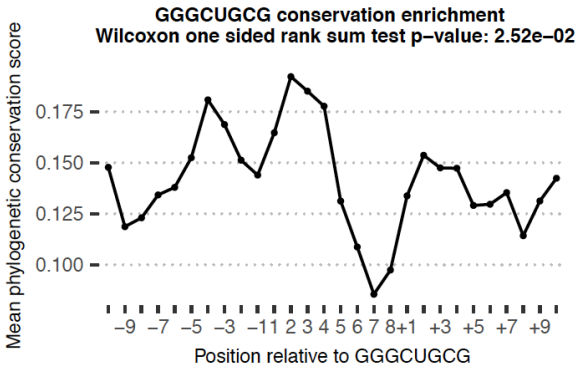
(j)


CGUGUGG conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.00e+00
Mean phylogenetic conservation score
0.21
0.18
0.15
0.12
0.09
−9 −7 −5 −3 −11 2 3 4 5 6 7+1 +3 +5 +7 +9
Position relative to CGUGUGG

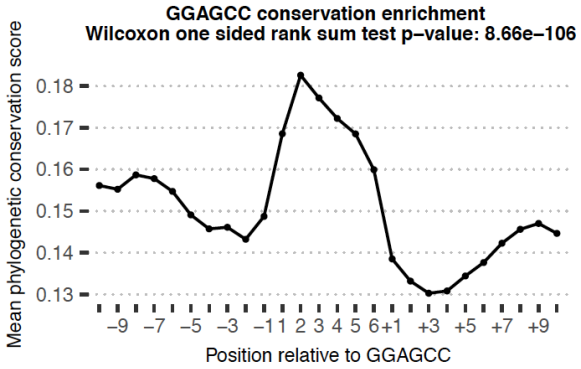
(k)


CUCAGG conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.16e−136
Mean phylogenetic conservation score
0.250
0.225
0.200
0.175
−9 −7 −5 −3 −11 2 3 4 5 6+1 +3 +5 +7 +9
Position relative to CUCAGG

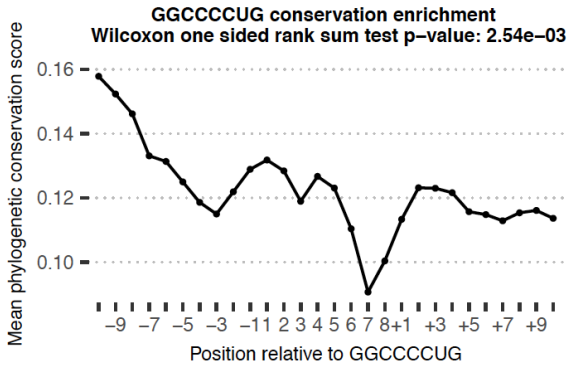
(l)


GGGCUGCG conservation enrichment
Wilcoxon one sided rank sum test p−value: 2.52e−02
Mean phylogenetic conservation score
0.175
0.150
0.125
0.100
−9 −7 −5 −3 −11 2 3 4 5 6 7 8+1 +3 +5 +7 +9
Position relative to GGGCUGCG

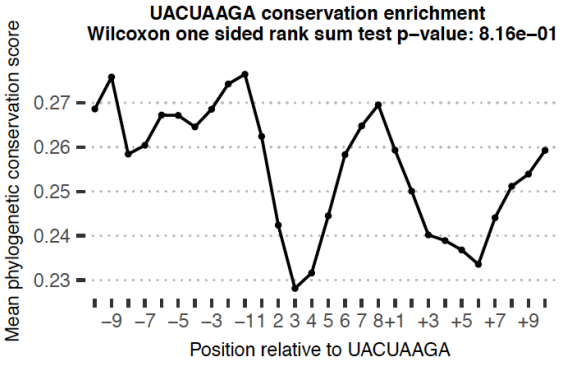
(m)


GGAGCC conservation enrichment
Wilcoxon one sided rank sum test p−value: 8.66e−106
Mean phylogenetic conservation score
0.18
0.17
0.16
0.15
0.14
0.13
−9 −7 −5 −3 −11 2 3 4 5 6+1 +3 +5 +7 +9
Position relative to GGAGCC

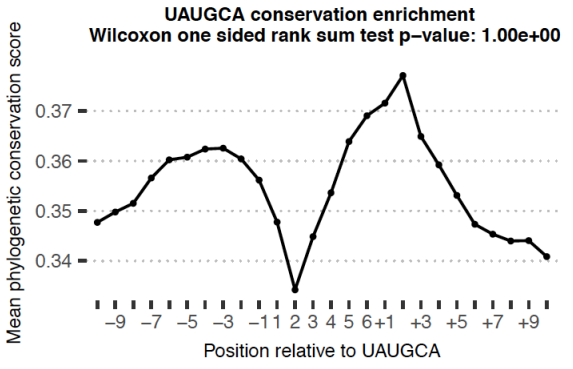
(n)


GGCCCCUG conservation enrichment
Wilcoxon one sided rank sum test p−value: 2.54e−03
Mean phylogenetic conservation score
0.16
0.14
0.12
0.10
−9 −7 −5 −3 −11 2 3 4 5 6 7 8+1 +3 +5 +7 +9
Position relative to GGCCCCUG

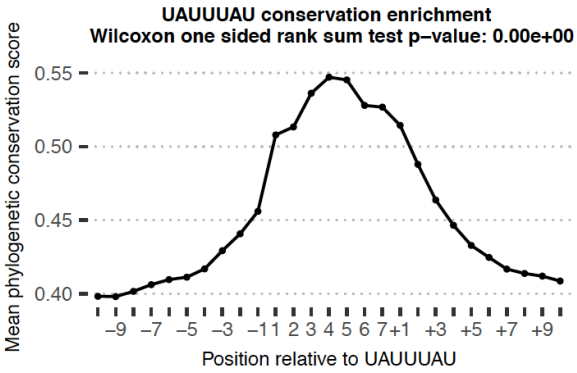
(o)


UACUAAGA conservation enrichment
Wilcoxon one sided rank sum test p−value: 8.16e−01
Mean phylogenetic conservation score
0.27
0.26
0.25
0.24
0.23
−9 −7 −5 −3 −11 2 3 4 5 6 7 8+1 +3 +5 +7 +9
Position relative to UACUAAGA

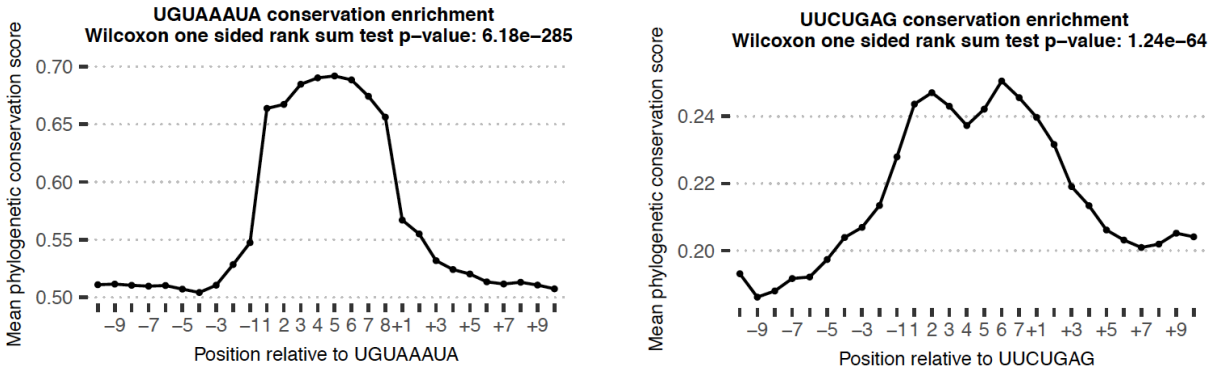
(p)


UAUGCA conservation enrichment
Wilcoxon one sided rank sum test p−value: 1.00e+00
Mean phylogenetic conservation score
0.37
0.36
0.35
0.34
−9 −7 −5 −3 −11 2 3 4 5 6+1 +3 +5 +7 +9
Position relative to UAUGCA

(q)

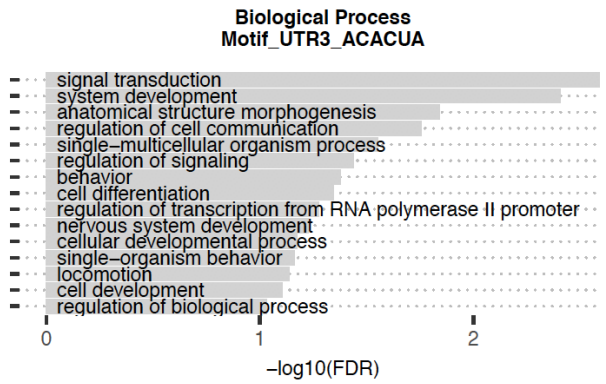

UAUUUAU conservation enrichment
Wilcoxon one sided rank sum test p−value: 0.00e+00
Mean phylogenetic conservation score
0.55
0.50
0.45
0.40
−9 −7 −5 −3 −11 2 3 4 5 6 7+1 +3 +5 +7 +9
Position relative to UAUUUAU

(r)

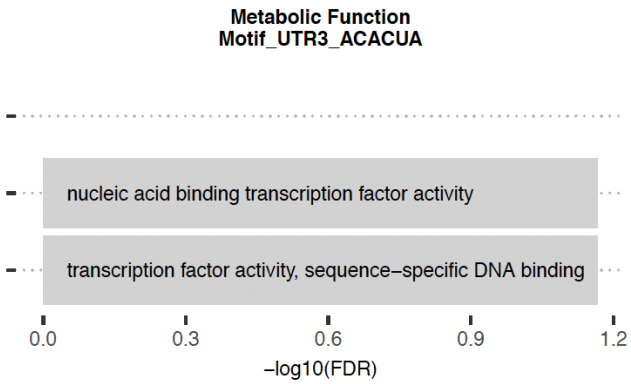


(s) **(t)**

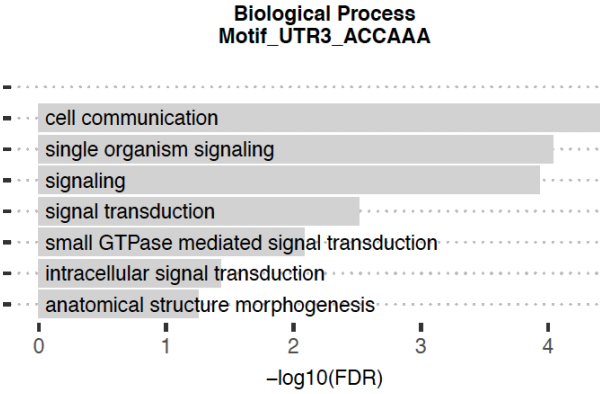
Figure A.-2: 3' UTR motifs' conservation scores. Average 100-vertebrate PhastCons score (y-axis) per position relative to the exact motif match instances in 3' UTR (x-axis). P-values assess significance of the average 100-vertebrate PhastCons scores at the motif sites compared to the two 10-nucleotide flanking regions


Biological Process
Motif_UTR3_ACACUA
signal transduction
system development
anatomical structure morphogenesis
regulation of cell communication
single−multicellular organism process
regulation of signaling
behavior
cell differentiation
regulation of transcription from RNA polymerase II promoter
nervous system development
cellular developmental process
single−organism behavior
locomotion
cell development
regulation of biological process
0
1
2
−log10(FDR)

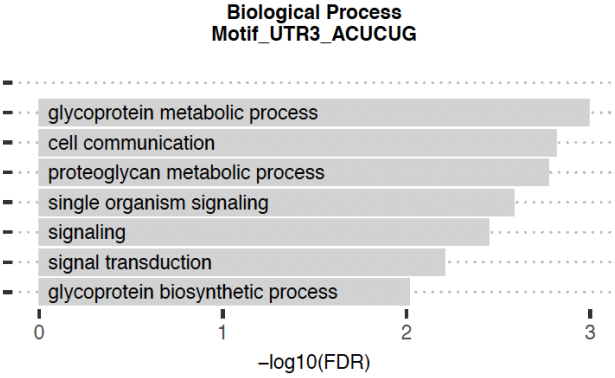
(a)


Metabolic Function
Motif_UTR3_ACACUA
nucleic acid binding transcription factor activity
transcription factor activity, sequence−specific DNA binding
0.0
0.3
0.6
0.9
1.2
−log10(FDR)

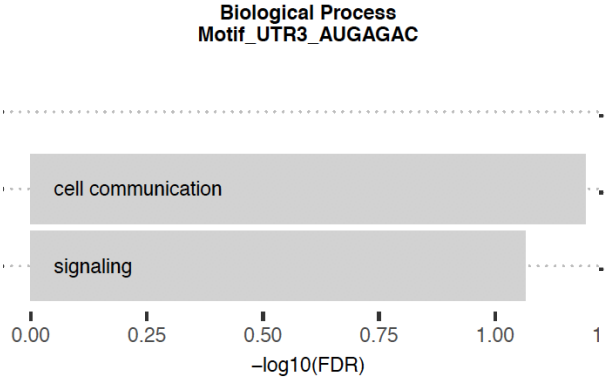
(b)


Biological Process
Motif_UTR3_ACCAAA
cell communication
single organism signaling
signaling
signal transduction
small GTPase mediated signal transduction
intracellular signal transduction
anatomical structure morphogenesis
0
1
2
3
4
−log10(FDR)

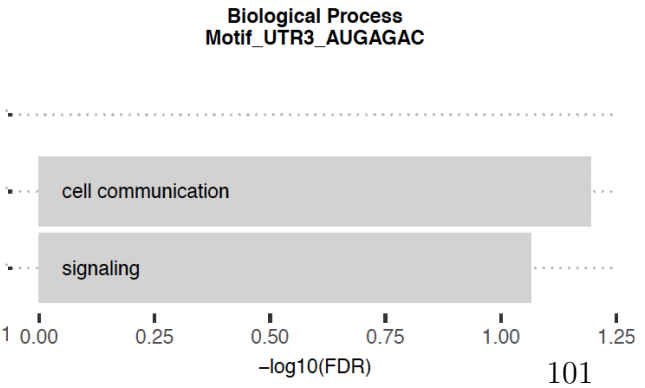
(c)


Biological Process
Motif_UTR3_ACUCUG
glycoprotein metabolic process
cell communication
proteoglycan metabolic process
single organism signaling
signaling
signal transduction
glycoprotein biosynthetic process
0
1
2
3
−log10(FDR)

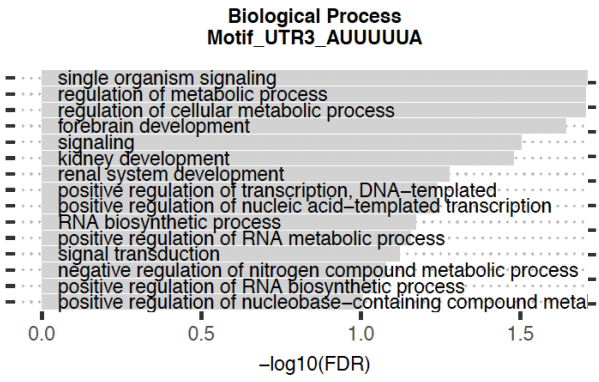
(d)


Biological Process
Motif_UTR3_AUGAGAC
cell communication
signaling
0.00
0.25
0.50
0.75
1.00
1
−log10(FDR)

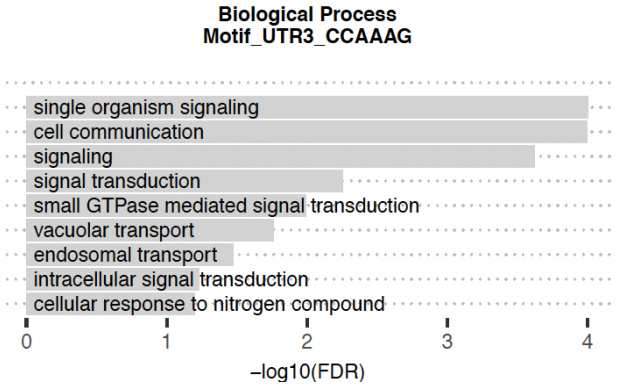
(e)


Biological Process
Motif_UTR3_AUGAGAC
cell communication
signaling
0.00
0.25
0.50
0.75
1.00
1.25
−log10(FDR)

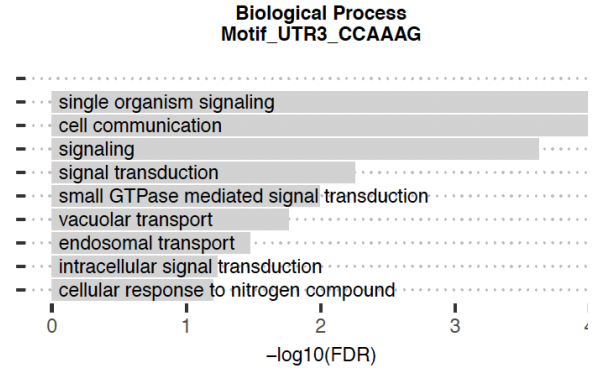
(f)


Biological Process
Motif_UTR3_AUUUUUA
single organism signaling
regulation of metabolic process
regulation of cellular metabolic process
forebrain development
signaling
kidney development
renal system development
positive regulation of transcription, DNA−templated
positive regulation of nucleic acid−templated transcription
RNA biosynthetic process
positive regulation of RNA metabolic process
signal transduction
negative regulation of nitrogen compound metabolic process
positive regulation of RNA biosynthetic process
positive regulation of nucleobase−containing compound meta
0.0
0.5
1.0
1.5
−log10(FDR)

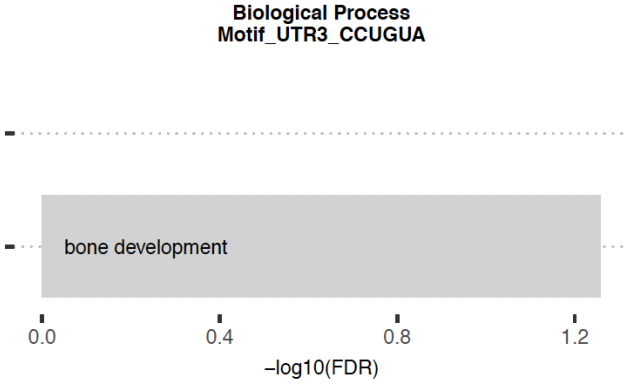
(g)


Biological Process
Motif_UTR3_CCAAAG
single organism signaling
cell communication
signaling
signal transduction
small GTPase mediated signal transduction
vacuolar transport
endosomal transport
intracellular signal transduction
cellular response to nitrogen compound
0
1
2
3
4
−log10(FDR)

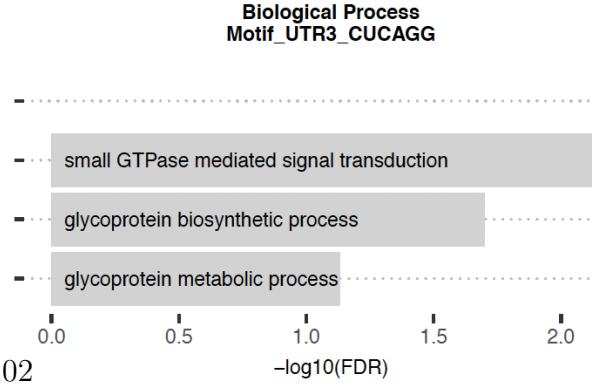
(h)


Biological Process
Motif_UTR3_CCAAAG
single organism signaling
cell communication
signaling
signal transduction
small GTPase mediated signal transduction
vacuolar transport
endosomal transport
intracellular signal transduction
cellular response to nitrogen compound
0
1
2
3
−log10(FDR)

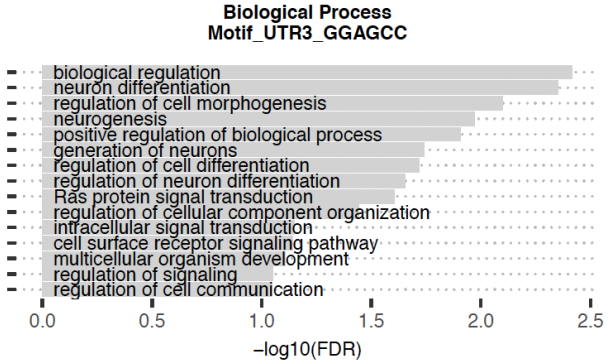
(i)


Biological Process
Motif_UTR3_CCUGUA
bone development
0.0
0.4
0.8
1.2
−log10(FDR)

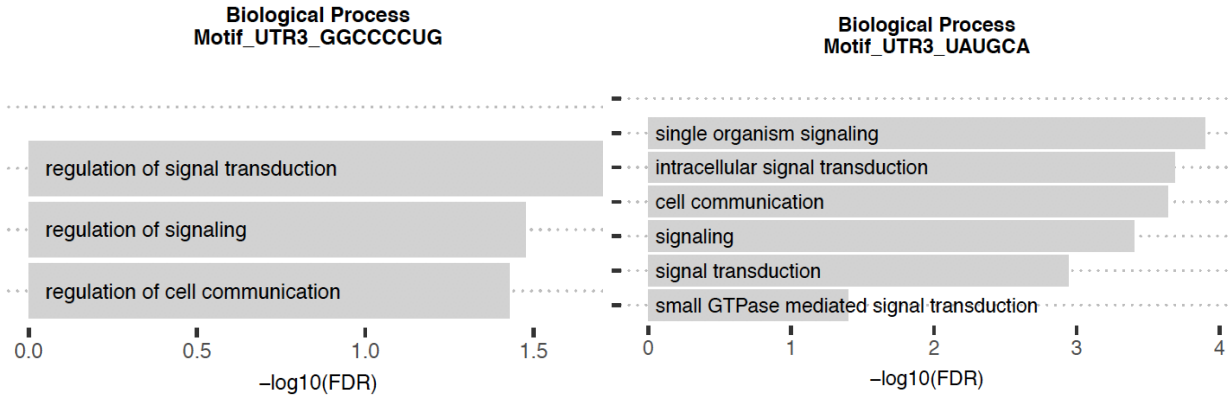
(j)


Biological Process
Motif_UTR3_CUCAGG
small GTPase mediated signal transduction
glycoprotein biosynthetic process
glycoprotein metabolic process
0.0
0.5
1.0
1.5
2.0
−log10(FDR)

(k)

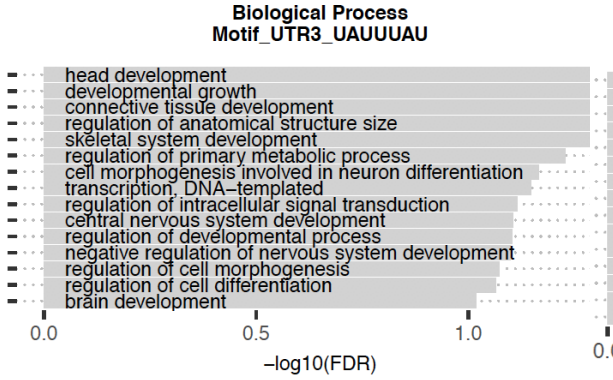

Biological Process
Motif_UTR3_GGAGCC
biological regulation
neuron differentiation
regulation of cell morphogenesis
neurogenesis
positive regulation of biological process
generation of neurons
regulation of cell differentiation
regulation of neuron differentiation
Ras protein signal transduction
regulation of cellular component organization
intracellular signal transduction
cell surface receptor signaling pathway
multicellular organism development
regulation of signaling
regulation of cell communication
0.0
0.5
1.0
1.5
2.0
2.5
−log10(FDR)

(l)

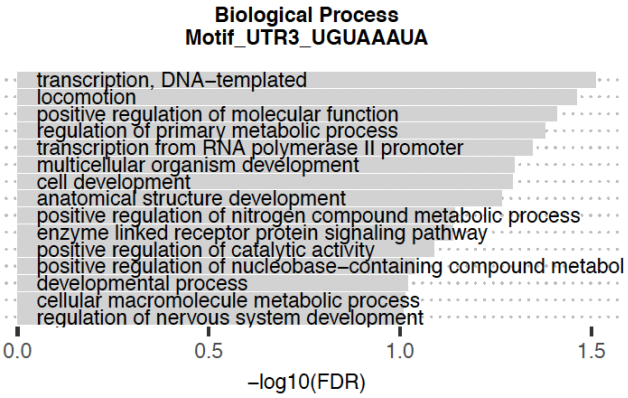


(m)

(n)



(o)



(p)

Figure A.-3: Gene ontology terms enriched for genes that have the consensus 3' UTR motifs. Gene ontology terms that are enriched for set of genes that contain consensus sequences of the de-novo identified k-mers in 3' UTR that are predictive of PTR ratios.

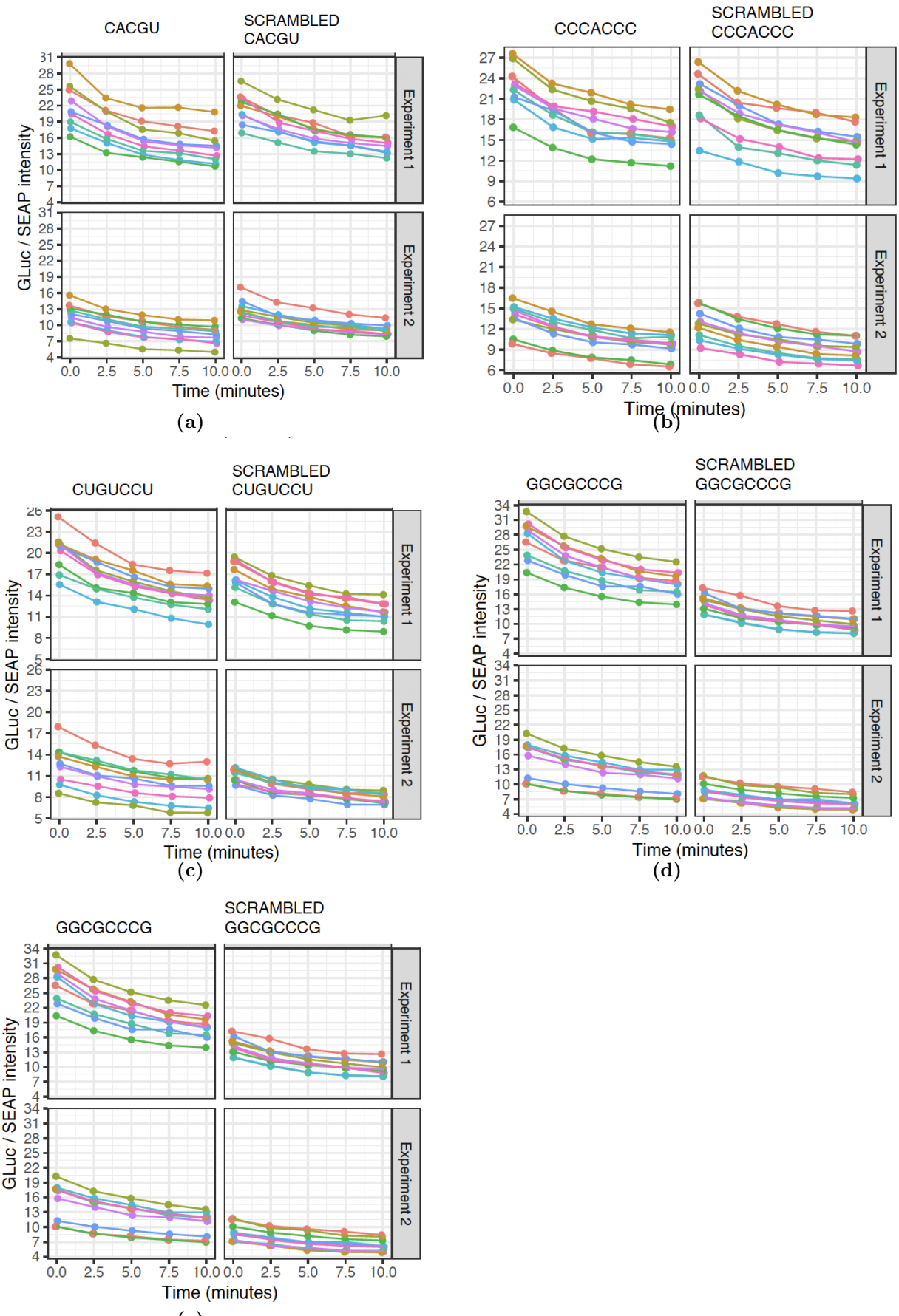


Figure A.-2: Selected 5' UTR motifs' test experiments. Time course Gluc/SEAP intensity values per 5' UTR motif and its scrambled version.

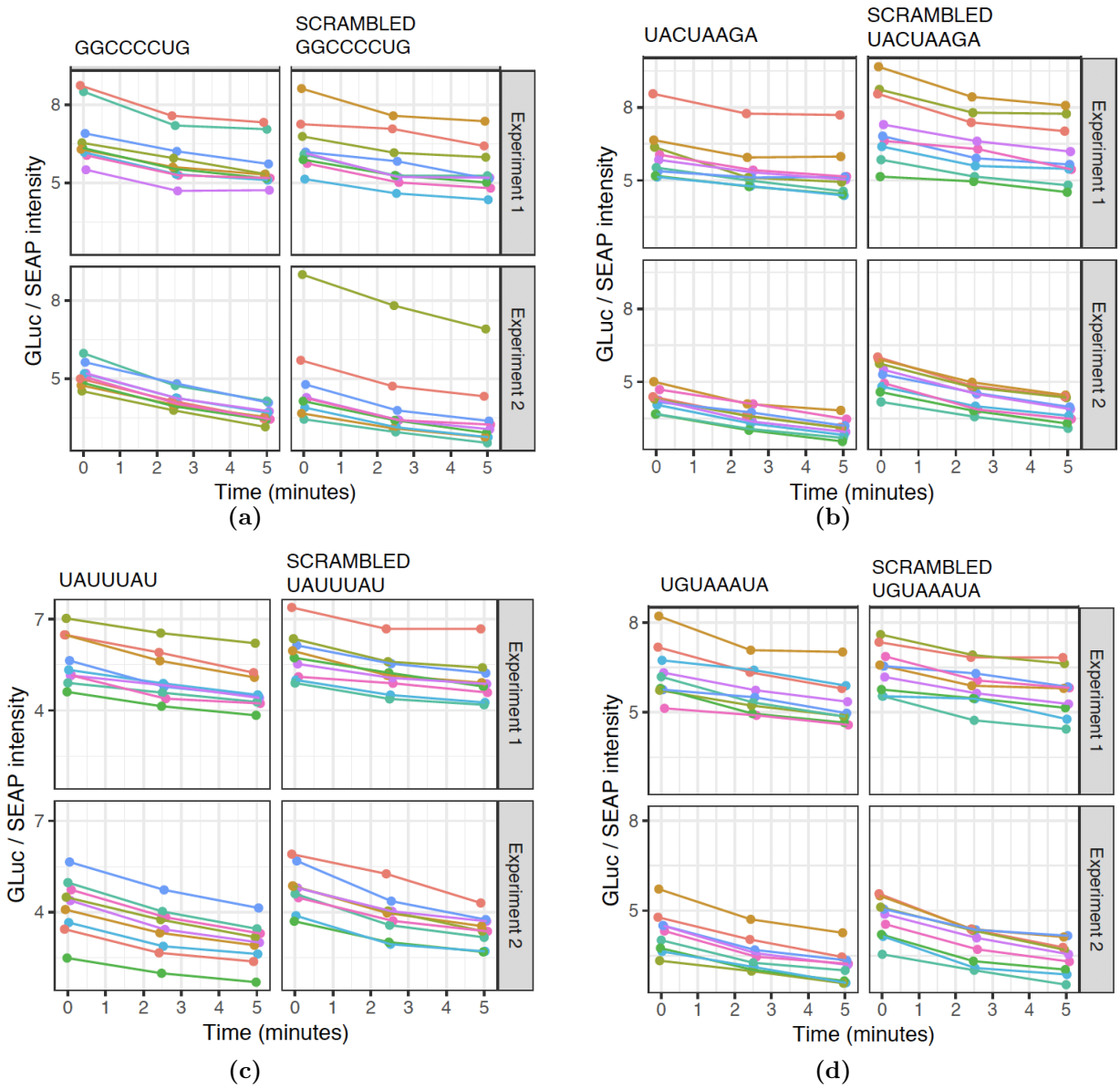


Figure A.-1: Selected 3' UTR motifs' test experiments. Time course Gluc/SEAP intensity values per 3' UTR motif and its scrambled version.

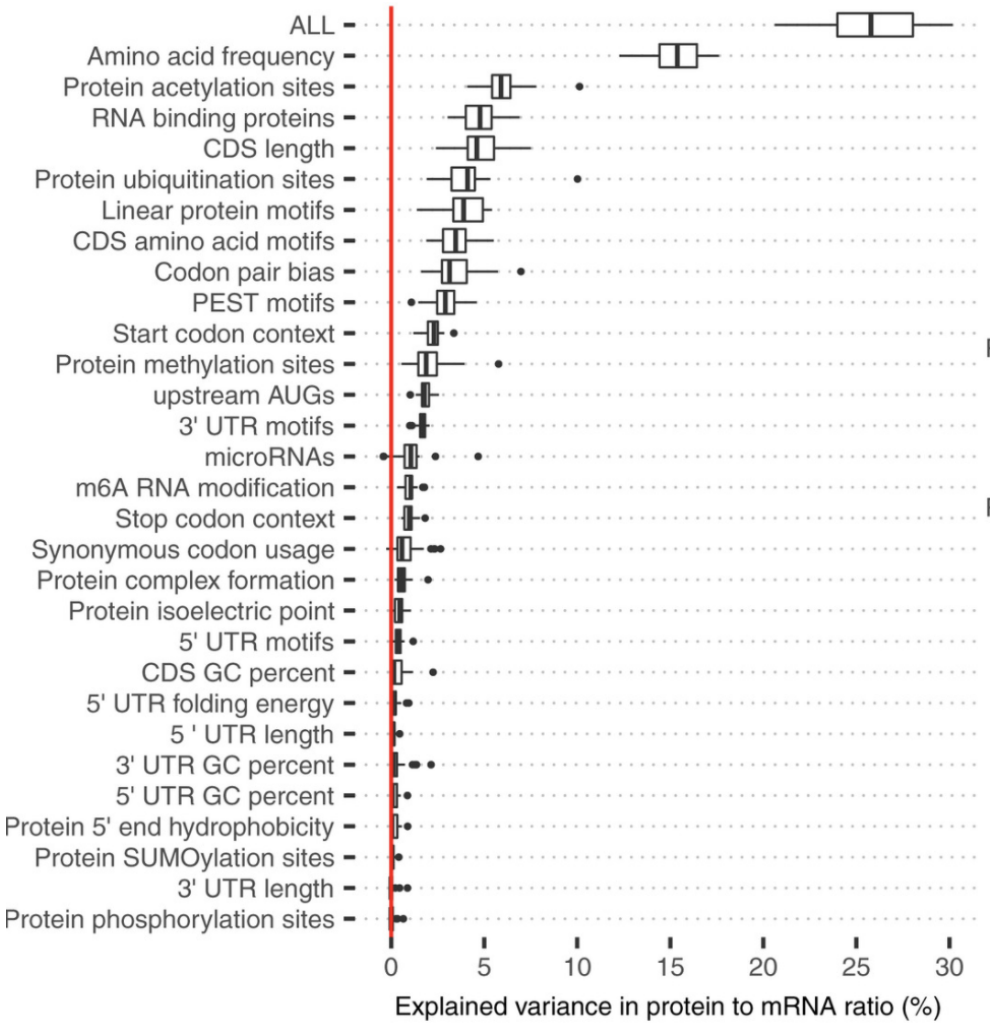
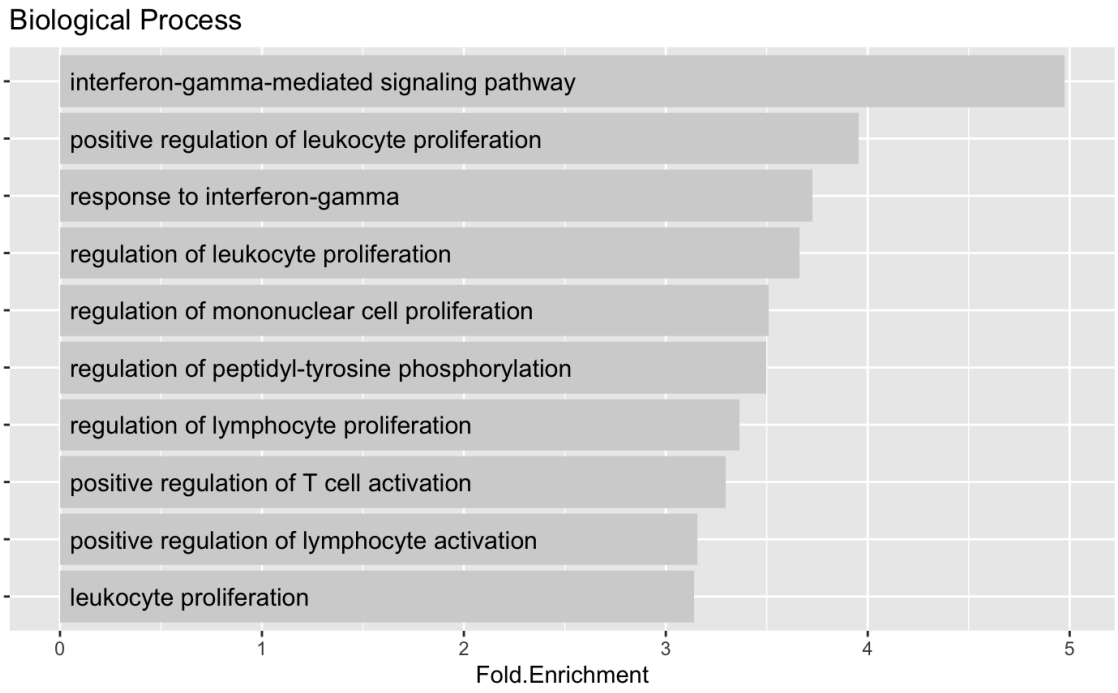
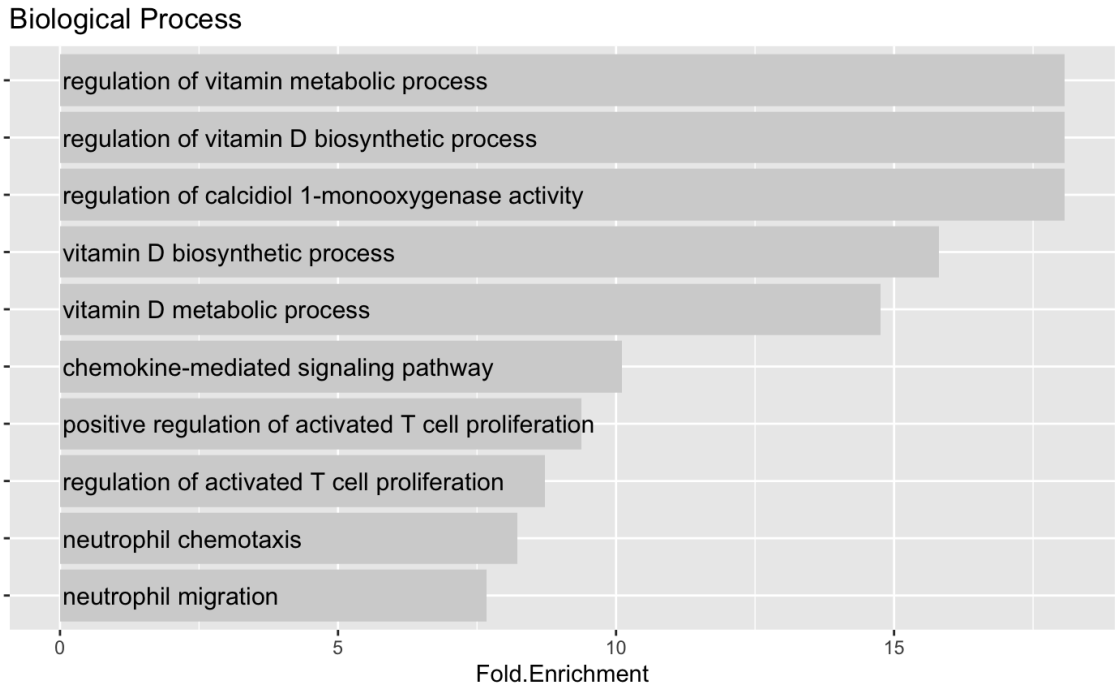


Figure A.0: The distribution of the total explained variance by the linear model that combines all of the listed sequence features is displayed in the first line. Shown are the quartiles (boxes and vertical lines) and furthest data points still within 1.5 times the interquartile range of the lower and upper quartiles (whiskers).



(a)



(b)

Figure A.1: GO terms enriched for the genes in group 4 and 5.

List of Figures

References

[1] Eraslan, B. *et al.* Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Molecular Systems Biology* **15** (2019). URL https://onlinelibrary.wiley.com/doi/abs/10.15252/msb.20188513.

[2] Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Molecular Systems Biology* **15** (2019). URL https://onlinelibrary.wiley.com/doi/abs/10.15252/msb.20188503.

[3] McManus, J., Cheng, Z. & Vogel, C. Next-generation analysis of gene expression regulation-comparing the roles of synthesis and degradation. *Molecular BioSystems* **11**, 2680–2689 (2015). URL http://dx.doi.org/10.1039/C5MB00310E.

[4] Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

[5] Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* **20**, 207–220 (2019). URL https://doi.org/10.1038/s41576-018-0089-8.

[6] Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics* **15**, 453–468 (2014). URL http://www.nature.com/articles/nrg3684.

[7] Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology* **18**, 437–451 (2017). URL http://www.nature.com/doifinder/10.1038/nrm.2017.27.

[8] Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760–1774 (2012). URL http://genome.cshlp.org/cgi/doi/10.1101/gr.135350.111.

[9] Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics* **10**, 57–63 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19015660http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2949280.

[10] Wong, J. W. H. & Cagney, G. An Overview of Label-Free Quantitation Methods in Proteomics by Mass Spectrometry. 273–283 (2010). URL http://link.springer.com/10.1007/978-1-60761-444-9{_}18.

[11] Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**, 218–223 (2009). URL https://science.sciencemag.org/content/324/5924/218.

[12] Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods* **13**, 508–514 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27018577http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4887338http://www.nature.com/articles/nmeth.3810.

[13] Bertone, P. *et al.* Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. *Science* **306**, 2242–2246 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15539566http://www.sciencemag.org/cgi/doi/10.1126/science.1103388.

[14] Cheng, J. *et al.* Transcriptional Maps of 10 Human Chromosomes at 5-Nucleotide Resolution. *Science* **308**, 1149–1154 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/15790807http://www.sciencemag.org/cgi/doi/10.1126/science.1108625.

[15] David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 5320–5 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16569694http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1414796.

[16] Clark, T. A., Sugnet, C. W. & Ares, M. Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays. *Science* **296**, 907–910 (2002). URL http://www.ncbi.nlm.nih.gov/pubmed/11988574http://www.sciencemag.org/cgi/doi/10.1126/science.1069415.

[17] Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**, R18 (2011). URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-2-r18.

[18] Wilkins, M. R. *et al.* Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It. *Biotechnology and Genetic Engineering Reviews* **13**, 19–50 (1996). URL http://www.tandfonline.com/doi/abs/10.1080/02648725.1996.10647923.

[19] Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016). URL http://www.nature.com/articles/nature19949.

[20] Sinitcyn, P., Rudolph, J. D. & Cox, J. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annual Review of Biomedical Data Science* **1**, 207–234 (2018). URL https://www.annualreviews.org/doi/10.1146/annurev-biodatasci-080917-013516.

[21] STEITZ, J. A. Polypeptide Chain Initiation: Nucleotide Sequences of the Three Ribosomal Binding Sites in Bacteriophage R17 RNA. *Nature* **224**, 957–964 (1969). URL http://www.nature.com/articles/224957a0.

[22] Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Reviews Genetics* **15**, 205–213 (2014). URL http://www.nature.com/articles/nrg3645.

[23] O'Connor, P. B. F., Andreev, D. E. & Baranov, P. V. Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nature Communications* **7**, 12915 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27698342http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5059445http://www.nature.com/articles/ncomms12915.

[24] Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology* **19**, 327–341 (2018). URL http://www.nature.com/articles/nrm.2017.130.

[25] Marcotte, E. M. & Vogel, C. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature reviews. Genetics* **13**, 227–232 (2012).

[26] Jovanovic, M. *et al.* Dynamic profiling of the protein life cycle in response to pathogens. *Science* **347** (2015). arXiv:1011.1669v3.

[27] Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016). URL https://www.sciencedirect.com/science/article/pii/S0092867416302707?via{%}3Dihub.

[28] Lundberg, E. *et al.* Defining the transcriptome and proteome in three functionally different human cell lines. *Molecular Systems Biology* **6**, 450 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/21179022http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3018165http://msb.embopress.org/cgi/doi/10.1038/msb.2010.106.

[29] Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014). URL http://www.nature.com/doifinder/10.1038/nature13319.

[30] Edfors, F. *et al.* Gene-specific correlation of RNA and protein levels in human cells and tissues. *Molecular Systems Biology* **12**, 883 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27951527http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5081484http://msb.embopress.org/lookup/doi/10.15252/msb.20167144.

[31] Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? *Nature* **547**, E19–E20 (2017). URL http://www.nature.com/doifinder/10.1038/nature22293. 209.

[32] Franks, A., Airoldi, E. & Slavov, N. Post-transcriptional regulation across human tissues. *PLoS Computational Biology* **13**, 1–20 (2017).

[33] Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between Protein and mRNA Abundance in Yeast. *Molecular and Cellular Biology* **19**, 1720–1730 (1999).

[34] De Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels (2009).

[35] Maier, T., Güell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. *FEBS Letters* **583**, 3966–3973 (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19850042http://doi.wiley.com/10.1016/j.febslet.2009.10.036.

[36] Siwiak, M. & Zielenkiewicz, P. A comprehensive, quantitative, and genome-wide model of translation. *PLoS Computational Biology* **6**, 4 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20686685http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2912337.

[37] Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics* **13**, 227–232 (2012).

[38] Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2**, e270 (2014). URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3940484{&}tool=pmcentrez{&}rendertype=abstract. 1212.0587.

[39] Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M. & Drummond, D. A. Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast. *PLOS Genetics* **11**, e1005206 (2015). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005206.

[40] Kozak, M. Selection of initiation sites by eucaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin. *Nucleic acids research* **12**, 3873–93 (1984).

[41] Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science* **324**, 255–258 (2009).

[42] Kozak, M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**, 283–292 (1986). URL https://www.sciencedirect.com/science/article/pii/0092867486907622.

[43] Sørensen, M. A., Kurland, C. G. & Pedersen, S. Codon usage determines translation rate in Escherichia coli. *Journal of molecular biology* **207**, 365–77 (1989). URL http://www.ncbi.nlm.nih.gov/pubmed/2474074.

[44] Gardin, J. *et al.* Measurement of average decoding rates of the 61 sense codons in vivo. *eLife* **3** (2014). URL https://elifesciences.org/articles/03735.

[45] Hanson, G. & Coller, J. Translation and Protein Quality Control: Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews Molecular Cell Biology* **19**, 20–30 (2018). URL http://dx.doi.org/10.1038/nrm.2017.91.

[46] Varenne, S., Buc, J., Lloubes, R. & Lazdunski, C. Translation is a non-uniform process: Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *Journal of Molecular Biology* **180**, 549–576 (1984). URL https://www.sciencedirect.com/science/article/pii/0022283684900275.

[47] Plotkin, J. B. & Kudla, G. Synonymous but not the same: The causes and consequences of codon bias. *Nature Reviews Genetics* **12**, 32–42 (2011). URL http://dx.doi.org/10.1038/nrg2899. NIHMS150003.

[48] Quax, T. E., Claassens, N. J., Söll, D. & van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell* **59**, 149–161 (2015). 15334406.

[49] Qu, X. *et al.* The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature* **475**, 118–121 (2011).

[50] Artieri, C. G. & Fraser, H. B. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome research* **24**, 2011–21 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/25294246http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4248317.

[51] Sabi, R. & Tuller, T. Computational analysis of nascent peptides that induce ribosome stalling and their proteomic distribution in Saccharomyces cerevisiae. *RNA* **23**, 983–994 (2017).

[52] Dao Duc, K. & Song, Y. S. The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation. *PLOS Genetics* **14**, e1007166 (2018). URL https://dx.plos.org/10.1371/journal.pgen.1007166.

[53] Bonetti, B., Fu, L., Moon, J. & Bedwell, D. M. The Efficiency of Translation Termination is Determined by a Synergistic Interplay Between Upstream and Downstream Sequences inSaccharomyces cerevisiae. *Journal of Molecular Biology* **251**,

334–345 (1995). URL http://www.ncbi.nlm.nih.gov/pubmed/7650736https://linkinghub.elsevier.com/retrieve/pii/S002228368570438X.

[54] McCaughan, K. K., Brown, C. M., Dalphin, M. E., Berry, M. J. & Tate, W. P. Translational termination efficiency in mammals is influenced by the base following the stop codon. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 5431–5 (1995). URL http://www.ncbi.nlm.nih.gov/pubmed/7777525http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC41708.

[55] Poole, E. S., Brown, C. M. & Tate, W. P. The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. *The EMBO journal* **14**, 151–8 (1995). URL http://www.ncbi.nlm.nih.gov/pubmed/7828587http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC398062.

[56] Tate, W. P. *et al.* The translational stop signal: codon with a context, or extended factor recognition element? *Biochimie* **78**, 945–52 (1996). URL http://www.ncbi.nlm.nih.gov/pubmed/9150871.

[57] Baek, D. *et al.* The impact of microRNAs on protein output. *Nature* **455**, 64–71 (2008).

[58] Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008). URL http://www.nature.com/articles/nature07228.

[59] Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835–840 (2010). URL http://www.nature.com/articles/nature09267.

[60] Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nature Reviews Genetics* **15**, 829–845 (2014). URL http://dx.doi.org/10.1038/nrg3813. /www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3006164{&}tool=pmcentrez{&}rendertype=abstract.

[61] Hudson, W. H. & Ortlund, E. A. The structure, function and evolution of proteins that bind DNA and RNA. *Nature Reviews Molecular Cell Biology* **15**, 749–760 (2014). URL http://www.nature.com/articles/nrm3884.

[62] Cottrell, K. A., Szczesny, P. & Djuranovic, S. Translation efficiency is a determinant of the magnitude of miRNA-mediated repression. *Scientific Reports* **7**, 14884 (2017). URL http://www.nature.com/articles/s41598-017-13851-w.

[63] Geffen, Y. *et al.* Mapping the Landscape of a Eukaryotic Degronome. *Molecular Cell* **63**, 1055–1065 (2016). URL https://www.sciencedirect.com/science/article/pii/S1097276516304191?via{%}3Dihub.

[64] Bachmair, A., Finley, D. & Varshavsky, A. In vivo half-life of a protein is a function of its amino-terminal residue. *Science (New York, N.Y.)* **234**, 179–86 (1986). URL http://www.ncbi.nlm.nih.gov/pubmed/3018930.

[65] Kats, I. *et al.* Mapping Degradation Signals and Pathways in a Eukaryotic N-terminome. *Molecular Cell* **70**, 488–501.e5 (2018). URL https://www.sciencedirect.com/science/article/pii/S1097276518302363?via{%}3Dihub.

[66] Ravid, T. & Hochstrasser, M. Diversity of degradation signals in the ubiquitin–proteasome system. *Nature Reviews Molecular Cell Biology* **9**, 679–689 (2008). URL http://www.nature.com/articles/nrm2468.

[67] Maurer, M. J. *et al.* Degradation Signals for Ubiquitin-Proteasome Dependent Cytosolic Protein Quality Control (CytoQC) in Yeast. *G3: Genes—Genomes—Genetics* **6**, 1853–1866 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27172186http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4938640http://g3journal.org/lookup/doi/10.1534/g3.116.027953.

[68] Mészáros, B., Kumar, M., Gibson, T. J., Uyar, B. & Dosztányi, Z. Degrons in cancer. *Science signaling* **10**, eaak9982 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28292960.

[69] Steinman, R. M. & Nussenzweig, M. C. Avoiding horror autotoxicus: The importance of dendritic cells in peripheral T cell tolerance. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 351–358 (2002). URL www.pnas.orgcgidoi10.1073pnas.231606698.

[70] Banchereau, J. & Steinman, R. M. Dendritic cells and the control of immunity (1998). URL https://www.nature.com/articles/32588.

[71] Dhodapkar, M. V., Dhodapkar, K. M. & Palucka, A. K. Interactions of tumor cells with dendritic cells: Balancing immunity and tolerance (2008). URL https://pubmed.ncbi.nlm.nih.gov/17948027/.

[72] Reis E Sousa, C. Dendritic cells in a mature age (2006). URL https://www.nature.com/articles/nri1845.

[73] O'Shea, J. & Paul, W. E. Mechanisms underlying lineage commitment and plasticity of helper CD4 + T cells (2010). URL https://pubmed.ncbi.nlm.nih.gov/20185720/.

[74] Kristensen, A. R., Gsponer, J. & Foster, L. J. Protein synthesis rate is the predominant regulator of protein expression during differentiation. *Molecular Systems Biology* **9**, 1–12 (2013). URL http://dx.doi.org/10.1038/msb.2013.47.

[75] Ceppi, M. *et al.* Ribosomal protein mRNAs are translationally-regulated during human dendritic cells activation by LPS. *Immunome Research* **5** (2009). URL https://pubmed.ncbi.nlm.nih.gov/19943945/.

[76] Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011). URL http://www.cell.com/article/S0092867411011925/fulltexthttp://www.cell.com/article/S0092867411011925/abstracthttps://www.cell.com/cell/abstract/S0092-8674(11)01192-5.

[77] Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal* **33**, 981–993 (2014). URL http://emboj.embopress.org/cgi/doi/10.1002/embj.201488411.

[78] Aspden, J. L. *et al.* Extensive translation of small open reading frames revealed by poly-ribo-seq. *eLife* **3**, 1–19 (2014).

[79] Skabkin, M. A., Skabkina, O. V., Hellen, C. U. & Pestova, T. V. Reinitiation and other unconventional posttermination events during eukaryotic translation. *Molecular Cell* **51**, 249–264 (2013). URL /pmc/articles/PMC4038429//pmc/articles/PMC4038429/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4038429/.

[80] Young, D. J., Guydosh, N. R., Zhang, F., Hinnebusch, A. G. & Green, R. Rli1/ABCE1 Recycles Terminating Ribosomes and Controls Translation Reinitiation in 3′UTRs In Vivo. *Cell* **162**, 872–884 (2015). URL http://dx.doi.org/10.1016/j.cell.2015.07.041.

[81] Hargrove, J. L. & Schmidt, F. H. The role of mRNA and protein stability in gene expression. *The FASEB Journal* **3**, 2360–2370 (1989).

[82] Hausser, J., Mayo, A., Keren, L. & Alon, U. Central dogma rates and the trade-off between precision and economy in gene expression. *Nature Communications* **10**, 1–15 (2019).

[83] Fagerberg, L. *et al.* Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. *Molecular & Cellular Proteomics* **13**, 397–406 (2014). URL http://www.mcponline.org/cgi/doi/10.1074/mcp.M113.035600.

[84] Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010). URL https://science.sciencemag.org/content/329/5991/533https://science.sciencemag.org/content/329/5991/533.abstract.

[85] Newton, W. A., Beckwith, J. R., Zipser, D. & Brenner, S. Nonsense mutants and polarity in the Lac operon of Escherichia coli (1965). URL `https://pubmed.ncbi.nlm.nih.gov/5327654/`.

[86] Gowrishankar, J. & Harinarayanan, R. Why is transcription coupled to translation in bacteria? *Molecular Microbiology* **54**, 598–603 (2004). URL `http://doi.wiley.com/10.1111/j.1365-2958.2004.04289.x`.

[87] Beyer, A., Hollunder, J., Nasheuer, H. P. & Wilhelm, T. Post-transcriptional expression regulation in the yeast Saccharomyces cerevisiae on a genomic scale. *Molecular and Cellular Proteomics* **3**, 1083–1092 (2004). URL `http://www.mcponline.org`.

[88] Gu, W., Zhou, T. & Wilke, C. O. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Computational Biology* **6**, 1000664 (2010). URL `www.ploscompbiol.org`.

[89] Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* **14** (2018). URL `https://onlinelibrary.wiley.com/doi/abs/10.15252/msb.20178124`.

[90] Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57 (2009). URL `https://pubmed.ncbi.nlm.nih.gov/19131956/https://pubmed.ncbi.nlm.nih.gov/19131956/?dopt=Abstract`.

[91] Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**, 1–13 (2009). URL `https://pubmed.ncbi.nlm.nih.gov/19033363/https://pubmed.ncbi.nlm.nih.gov/19033363/?dopt=Abstract`.

[92] Radhakrishnan, A. & Green, R. Connections Underlying Translation and mRNA Stability (2016).

[93] Tani, H. *et al.* Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Research* **22**, 947–956 (2012). URL `https://pubmed.ncbi.nlm.nih.gov/22369889/https://pubmed.ncbi.nlm.nih.gov/22369889/?dopt=Abstract`.

[94] Schueler, M. *et al.* Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biology* **15** (2014). URL `https://pubmed.ncbi.nlm.nih.gov/24417896/https://pubmed.ncbi.nlm.nih.gov/24417896/?dopt=Abstract`.

[95] Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016). URL `https://pubmed.ncbi.nlm.nih.gov/27257258/https://pubmed.ncbi.nlm.nih.gov/27257258/?dopt=Abstract`.

[96] Dana, A. & Tuller, T. Mean of the typical decoding rates: A New translation efficiency index based on the analysis of ribosome profiling data. *G3: Genes, Genomes, Genetics* **5**, 73–80 (2015). URL http://www.cs.

[97] Zecha, J. *et al.* Peptide level turnover measurements enable the study of proteoform dynamics. *Molecular and Cellular Proteomics* **17**, 974–992 (2018). URL https://pubmed.ncbi.nlm.nih.gov/29414762/https://pubmed.ncbi.nlm.nih.gov/29414762/?dopt=Abstract.

[98] Mathieson, T. *et al.* Systematic analysis of protein turnover in primary cells. *Nature Communications* **9** (2018). URL https://pubmed.ncbi.nlm.nih.gov/29449567/https://pubmed.ncbi.nlm.nih.gov/29449567/?dopt=Abstract.

[99] Hendrix, D. K., Brenner, S. E. & Holbrook, S. R. RNA structural motifs: Building blocks of a modular biomolecule. *Quarterly Reviews of Biophysics* **38**, 221–243 (2005).

[100] Mao, Y., Liu, H., Liu, Y. & Tao, S. Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in Saccharomyces cerevisiae. *Nucleic Acids Research* **42**, 4813–4822 (2014).

[101] Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics* **15**, 469–479 (2014). URL http://dx.doi.org/10.1038/nrg3681.

[102] Kozak, M. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 8301–8305 (1990). URL https://www.pnas.org/content/87/21/8301https://www.pnas.org/content/87/21/8301.abstract.

[103] Barbosa, C., Peixeiro, I. & Romão, L. Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLoS Genetics* **9**, 1–12 (2013).

[104] Morris, D. R. & Geballe, A. P. Upstream Open Reading Frames as Regulators of mRNA Translation. *Molecular and Cellular Biology* **20**, 8635–8642 (2000). URL http://mcb.asm.org/.

[105] Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7507–7512 (2009). URL www.pnas.org/cgi/content/full/.

[106] Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**, 1034–1050 (2005). URL www.genome.org.

[107] Churbanov, A., Rogozin, I. B., Babenko, V. N., Ali, H. & Koonin, E. V. Evolutionary conservation suggests a regulatory function of AUG triplets in 5'-UTRs of eukaryotic genes. *Nucleic Acids Research* **33**, 5512–5520 (2005). URL `https://academic.oup.com/nar/article/33/17/5512/1067686`.

[108] Svidritskiy, E., Brilot, A. F., Koh, C. S., Grigorieff, N. & Korostelev, A. A. Structures of yeast 80S ribosome-tRNA complexes in the rotated and nonrotated conformations. *Structure* **22**, 1210–1218 (2014).

[109] Yu, C. H. *et al.* Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Molecular Cell* **59**, 744–754 (2015).

[110] Weinberg, D. E. *et al.* Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *bioRxiv* 021501 (2015). URL `http://biorxiv.org/content/early/2015/07/06/021501.abstract`.

[111] Yan, X., Hoek, T. A., Vale, R. D. & Tanenbaum, M. E. Dynamics of Translation of Single mRNA Molecules in Vivo. *Cell* **165**, 976–989 (2016).

[112] Wilson, D. N., Arenz, S. & Beckmann, R. Translation regulation via nascent polypeptide-mediated ribosome stalling (2016).

[113] Fang, L., Hou, S., Xue, L., Zheng, F. & Zhan, C. G. Amino-acid mutations to extend the biological half-life of a therapeutically valuable mutant of human butyrylcholinesterase. *Chemico-Biological Interactions* **214**, 18–25 (2014).

[114] Plotkin, J. B., Robins, H. & Levine, A. J. Tissue-specific codon usage and the expression of human genes. *Proc Natl Acad Sci USA* **101**, 12588–12591 (2004). URL `http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed{&}id=15314228{&}retmode=ref{&}cmd=prlinks`.

[115] Dittmar, K. A., Goodenbour, J. M. & Pan, T. Tissue-specific differences in human transfer RNA expression. *PLoS Genetics* **2**, 2107–2115 (2006).

[116] Gustafsson, C. *et al.* Engineering genes for predictable protein expression (2012).

[117] Dana, A. & Tuller, T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Research* **42**, 9171–9181 (2014).

[118] Charneski, C. A. & Hurst, L. D. Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. *PLoS Biology* **11**, e1001508 (2013). URL `https://dx.plos.org/10.1371/journal.pbio.1001508`.

[119] Hoekema, A., Kastelein, R. A., Vasser, M. & de Boer, H. A. Codon replacement in the PGK1 gene of Saccharomyces cerevisiae: experimental approach to study the role of biased codon usage in gene expression. *Molecular and Cellular Biology* **7**, 2914–2924 (1987). URL `http://mcb.asm.org/`.

[120] Presnyak, V. *et al.* Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111–1124 (2015).

[121] Bazzini, A. A. *et al.* Codon identity regulates <scp>mRNA</scp> stability and translation efficiency during the maternal-to-zygotic transition. *The EMBO Journal* **35**, 2087–2103 (2016). URL https://onlinelibrary.wiley.com/doi/10.15252/embj.201694699.

[122] Mishima, Y. & Tomari, Y. Codon Usage and 3' UTR Length Determine Maternal mRNA Stability in Zebrafish. *Molecular Cell* **61**, 874–885 (2016).

[123] Dehouck, Y. *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* **25**, 2537–2543 (2009). URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp445.

[124] Nick Pace, C., Martin Scholtz, J. & Grimsley, G. R. Forces stabilizing proteins (2014).

[125] Díaz-Villanueva, J., Díaz-Molina, R. & García-González, V. Protein Folding and Mechanisms of Proteostasis. *International Journal of Molecular Sciences* **16**, 17193–17230 (2015). URL http://www.mdpi.com/1422-0067/16/8/17193.

[126] Brown, A., Shao, S., Murray, J., Hegde, R. S. & Ramakrishnan, V. Structural basis for stop codon recognition in eukaryotes. *Nature* **524**, 493–496 (2015). URL https://www.nature.com/articles/nature14896.

[127] Arkov, A. L., Korolev, S. V. & Kisselev, L. L. Termination of translation in bacteria may be modulated via specific interaction between peptide chain release factor 2 and the last peptidyl-tRNASer/Phe. *Nucleic Acids Research* **21**, 2891–2897 (1993). URL https://academic.oup.com/nar/article/21/12/2891/1045771.

[128] Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020). URL https://doi.org/10.1038/s41586-020-2077-3.

[129] Kechavarzi, B. & Janga, S. C. Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biology* **15**, 1–16 (2014). URL https://link.springer.com/articles/10.1186/gb-2014-15-1-r14https://link.springer.com/article/10.1186/gb-2014-15-1-r14.

[130] Jacobsen, A., Wen, J., Marks, D. S. & Krogh, A. Signatures of RNA binding proteins globally coupled to effective microRNA target sites. *Genome Research* **20**, 1010–1019 (2010). URL http://www.genome.org/cgi/doi/10.1101/gr.103259.109.

[131] Chang, S. H. & Hla, T. Gene regulation by RNA binding proteins and microRNAs in angiogenesis (2011).

[132] Jiang, P. & Coller, H. Functional Interactions Between microRNAs and RNA Binding Proteins .

[133] Ciafrè, S. A. & Galardi, S. microRNAs and RNA-binding proteins. *RNA Biology* **10**, 934–942 (2013). URL `http://www.tandfonline.com/doi/abs/10.4161/rna.24641`.

[134] Chou, C. H. *et al.* MiRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Research* **46**, D296–D302 (2018). URL `http://mirtarbase.mbc.nctu.edu.tw/.`

[135] Tuller, T. & Zur, H. Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Research* **43**, 13–28 (2015). URL `https://academic.oup.com/nar/article/43/1/13/2903398`.

[136] Strickler, S. S. *et al.* Protein stability and surface electrostatics: A charged relationship. *Biochemistry* **45**, 2761–2766 (2006). URL `https://pubs.acs.org/sharingguidelines`.

[137] Chan, C.-H., Wilbanks, C. C., Makhatadze, G. I. & Wong, K.-B. Electrostatic Contribution of Surface Charge Residues to the Stability of a Thermophilic Protein: Benchmarking Experimental and Predicted pKa Values. *PLoS ONE* **7**, e30296 (2012). URL `https://dx.plos.org/10.1371/journal.pone.0030296`.

[138] Requião, R. D. *et al.* Protein charge distribution in proteomes and its impact on translation. *PLOS Computational Biology* **13**, e1005549 (2017). URL `https://dx.plos.org/10.1371/journal.pcbi.1005549`.

[139] Rogers, S., Wells, R. & Rechsteiner, M. Amino acid sequences common to rapidly degraded proteins: The PEST hypothesis. *Science* **234**, 364–368 (1986). URL `https://science.sciencemag.org/content/234/4774/364https://science.sciencemag.org/content/234/4774/364.abstract`.

[140] Dinkel, H. *et al.* ELM 2016 - Data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Research* **44**, D294–D300 (2016). URL `https://academic.oup.com/nar/article/44/D1/D294/2503097`.

[141] Giudice, G., Sánchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT—a database of RNA-binding proteins and associated motifs. *Database* **2016**, baw035 (2016). URL `https://academic.oup.com/database/article-lookup/doi/10.1093/database/baw035`.

[142] Proudfoot, N. Poly(A) signals (1991). URL `http://www.cell.com/article/009286749190495K/fulltexthttp://www.cell.com/article/009286749190495K/abstracthttps://www.cell.com/cell/abstract/0092-8674(91)90495-K`.

[143] Kruys, V., Marinx, O., Shaw, G., Deschamps, J. & Huez, G. Translational blockade imposed by cytokine-derived UA-rich sequences. *Science* **245**, 852–855 (1989). URL https://science.sciencemag.org/content/245/4920/852https://science.sciencemag.org/content/245/4920/852.abstract.

[144] Qi, M.-Y. *et al.* AU-Rich-Element-Dependent Translation Repression Requires the Cooperation of Tristetraprolin and RCK/P54. *Molecular and Cellular Biology* **32**, 913–928 (2012). URL http://mcb.asm.org/.

[145] Ma, W. J., Cheng, S., Campbell, C., Wright, A. & Furneaux, H. Cloning and characterization of HuR, a ubiquitously expressed Elav-like protein. *Journal of Biological Chemistry* **271**, 8144–8151 (1996).

[146] Parisi, M. & Lin, H. Translational repression: A duet of Nanos and Pumilio. *Current Biology* **10**, R81–R83 (2000).

[147] Piqué, M., López, J. M., Foissac, S., Guigó, R. & Méndez, R. A Combinatorial Code for CPE-Mediated Translational Control. *Cell* **132**, 434–448 (2008).

[148] Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419–1260419 (2015). URL https://www.sciencemag.org/lookup/doi/10.1126/science.1260419.

[149] Åberg, K., Saetre, P., Jareborg, N. & Jazin, E. Human QKI, a potential regulator of mRNA expression of human oligodendrocyte-related genes involved in schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 7482–7487 (2006). URL www.pnas.orgcgidoi10.1073pnas.0601213103.

[150] Teplova, M. *et al.* Structure-function studies of STAR family quaking proteins bound to their in vivo RNA target sites. *Genes and Development* **27**, 928–940 (2013). URL http://www.genesdev.org/cgi/doi/10.1101/gad.216531.113.

[151] Kanhoush, R. *et al.* Novel domains in the hnRNP G/RBMX protein with distinct roles in RNA binding and targeting nascent transcripts. *Nucleus* **1**, 109–122 (2010). URL http://www.tandfonline.com/doi/abs/10.4161/nucl.1.1.10857.

[152] Vogel, C. *et al.* Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology* **6**, 400 (2010). URL https://onlinelibrary.wiley.com/doi/10.1038/msb.2010.59.

[153] Wang, X. *et al.* N6-methyladenosine modulates messenger RNA translation efficiency. *Cell* **161**, 1388–1399 (2015). URL http://dx.doi.org/10.1016/j.cell.2015.05.014http://dx.doi.org/10.1016/j.cell.2015.05.014.

[154] Mueller, S. *et al.* Protein degradation corrects for imbalanced subunit stoichiometry in OST complex assembly. *Molecular Biology of the Cell* **26**, 2596–2608 (2015). URL https://www.molbiolcell.org/doi/10.1091/mbc.E15-03-0168.

[155] Ishikawa, K., Makanae, K., Iwasaki, S., Ingolia, N. T. & Moriya, H. Post-Translational Dosage Compensation Buffers Genetic Perturbations to Stoichiometry of Protein Complexes. *PLOS Genetics* **13**, e1006554 (2017). URL https://dx.plos.org/10.1371/journal.pgen.1006554.

[156] Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Research* **43**, D512–D520 (2015). URL http://www.proteinatlas.org/.

[157] Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes and Development* **27**, 2380–2396 (2013). URL https://pubmed.ncbi.nlm.nih.gov/24145798/.

[158] Ezkurdia, I. *et al.* Most highly expressed protein-coding genes have a single dominant isoform. *Journal of Proteome Research* **14**, 1880–1887 (2015). URL https://pubs.acs.org/doi/abs/10.1021/pr501286b.

[159] Wang, X., Hou, J., Quedenau, C. & Chen, W. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. *Molecular Systems Biology* **12**, 875 (2016). URL https://pubmed.ncbi.nlm.nih.gov/27430939/.

[160] Zacher, B., Lidschreiber, M., Cramer, P., Gagneur, J. & Tresch, A. Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle. *Molecular Systems Biology* **10**, 768 (2014). URL /pmc/articles/PMC4300491//pmc/articles/PMC4300491/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4300491/.

[161] Lareau, L. F., Hite, D. H., Hogan, G. J. & Brown, P. O. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife* **2014** (2014).

[162] Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**, 1367–1372 (2008). URL http://www.nature.com/articles/nbt.1511.

[163] Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts635.

[164] Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525–527 (2016). URL `http://www.nature.com/`.

[165] Lorenz, R. & Bernhart, S. H "o ner zu siederdissen c et al (2011) viennarna package 2.0. *Algorithms Mol Biol* **6**, 26.

[166] Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105–132 (1982).

[167] Rice, P., Longden, I. & Bleasby, A. Emboss: the european molecular biology open software suite. *Trends in genetics* **16**, 276–277 (2000).

[168] Dominissini, D. *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**, 201–206 (2012). URL `https://www.nature.com/articles/nature11112`.

[169] Ruepp, A. *et al.* CORUM: The comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Research* **38**, D497–D501 (2009). URL `http://www.geneontology.org/`.

[170] Avsec, Ž., Barekatain, M., Cheng, J. & Gagneur, J. Modeling positional effects of regulatory sequences with spline transformations increases prediction accuracy of deep neural networks. *Bioinformatics* **34**, 1261–1269 (2018). URL `https://academic.oup.com/bioinformatics/article/34/8/1261/4636216`.

[171] Eser, P. *et al.* Determinants of <scp>RNA</scp> metabolism in the Schizosaccharomyces pombe genome. *Molecular Systems Biology* **12**, 857 (2016). URL `https://onlinelibrary.wiley.com/doi/10.15252/msb.20156526`.

[172] Zhou, X., Carbonetto, P. & Stephens, M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genetics* **9**, 1003264 (2013). URL `www.plosgenetics.org. 1209.1341`.

[173] Kuhn, M. Building predictive models in r using the caret package journal of statistical software 28 (2008).

[174] Kremer, L. S. *et al.* Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature Communications* **8**, 1–11 (2017). URL `www.nature.com/naturecommunications`.

[175] Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology (2000). URL `http://www.flybase.bio.indiana.eduhttp://fruitfly.bdgp.berkeley.eduhttp://genome-www.stanford.eduhttp://www.informatics.jax.org`.

[176] de Klerk, E. & 't Hoen, P. A. Alternative mRNA transcription, processing, and translation: Insights from RNA sequencing (2015). URL `https://pubmed.ncbi.nlm.nih.gov/25648499/`.

[177] Sharp, P. M. & Li, W. H. An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution* **24**, 28–38 (1986). URL `https://link.springer.com/article/10.1007/BF02099948`.

[178] dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Research* **32**, 5036–5044 (2004). URL `http://flybase.bio.`

[179] Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural & Molecular Biology* (2012).

[180] Sémon, M., Lobry, J. R. & Duret, L. No evidence for tissue-specific adaptation of synonymous codon usage in humans. *Molecular biology and evolution* **23**, 523–529 (2006).

[181] Rudolph, K. L. *et al.* Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States. *PLoS Genetics* **12**, e1006024 (2016). URL `https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006024`.

[182] Frumkin, I. *et al.* Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E4940–E4949 (2018). URL `www.pnas.org/lookup/suppl/doi:10.1073/pnas.1719375115/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1719375115`.

[183] Gingold, H. *et al.* A dual program for translation regulation in cellular proliferation and differentiation. *Cell* **158**, 1281–1292 (2014).

[184] Lackner, D. H. *et al.* A Network of Multiple Regulatory Layers Shapes Gene Expression in Fission Yeast. *Molecular Cell* **26**, 145–155 (2007).