Sequence-based determinants of mRNA half-life in human cells

Pedro Tomaz da Silva
Electrical and
Computer Engineering
Instituto Superior Técnico
Lisboa, Portugal
Email: pedrotomaz20@gmail.com

***Abstract*—In humans, the DNA molecule encodes a program written in a 4-character language using a 3-billion-long-text, which defines the behavior and function of each cell in the organism.**

Understanding how this code (genotype) produces it's output (phenotype) is the topic of decades of research.

Portions of this code - genes - contain the instructions to build proteins. Regulating the amount of proteins in a cell at a given time is of utmost importance for its correct functioning.

The code for the production of a protein is copied from its storing location and delivered to its production site by a molecule termed messenger RNA (mRNA). The variable length 4-character-language sequence contained in the mRNA molecule partly determines the time window it stays functional and can be measured by its half-life. The longer the mRNA is available, the more proteins will be produced from it.

Here, we model mRNA half-life and its variations across different human cells through the task of predicting mRNA half-life from its sequence.

Linear regression modeling allowed us to evaluate the quantitative impact of each known mRNA sequence feature on half-life and its variation across cells. Together with high dimensional data analysis and visualization techniques, we uncovered a previously unknown connection between a cell's energy production and mRNA half-life through its translation.

A multi-task deep neural network was developed to predict a tissue's cell mRNA half-life variation and its overall performance indicates its further usability on other domains such as in tissue specific gene expression modeling.

Lastly, we developed deep convolutional neural network models for half-life prediction from sequence and subsequently interpreted them using the tools DeepLIFT and TF-MoDISco, revealing new possible sequence portions or motifs which potentially regulate mRNA half-life.

1. Introduction

The genome of an organism contains all the instructions which command its cells' response to environmental cues and their development throughout the life-cycle of the organism. The ultimate aim of the program encoded in the genome is to create the set of traits and responses which define a living being - the phenotype.

The genome program is encoded in the DNA molecule as a text written in a 4-character vocabulary corresponding to the DNA bases Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). In humans this text is 3 billion-characters long. Finding out how this text produces the phenotype corresponds to fitting a function of domain of minimum length 3×10^9 to the phenotypic space.

The main building blocks of the cell, RNAs and proteins, are the functional products whose instructions for their construction are encoded in certain portions of the genome called genes. The process comprising the steps which create a functional product from a gene is called gene expression.

The central dogma of molecular biology states that firstly, gene expression starts in the nucleus of the cell with the transfer of the information of a gene from the densely packed DNA molecule to an RNA molecule through a process termed transcription. Analogously to computer architecture, such process can be interpreted as a reading of the gene program from the disk (DNA molecule) into the RAM (RNA molecule). Secondly, the resulting RNA molecule, termed RNA transcript, is processed in a step called splicing, were certain portions of the RNA sequence, the introns, are removed and the remaining ones, the exons, are put together. Thirdly, in case the functional product of the gene is a protein, this RNA transcript is transported to the cytoplasm of the cell and subsequently used as a template to produce proteins in a process called translation (see Fig. 1). In this sense the RNA molecule acts as a message carrier, delivering the instructions for the design of a protein to its production site. For that reason, this RNA molecule is termed messenger RNA (mRNA).

The regulation of protein levels is vital to cellular functioning. mRNA degradation is one mechanism which allows for the regulation of protein amounts. Once in the cell's cytoplasm, the mRNA molecule can be used to produce multiple equal proteins through a molecular decoding machine called ribosome. Over time, the mRNA molecule degrades, making it unable to be used again. Therefore, the time window an mRNA molecule stays available for translation will influence the number of produced proteins encoded

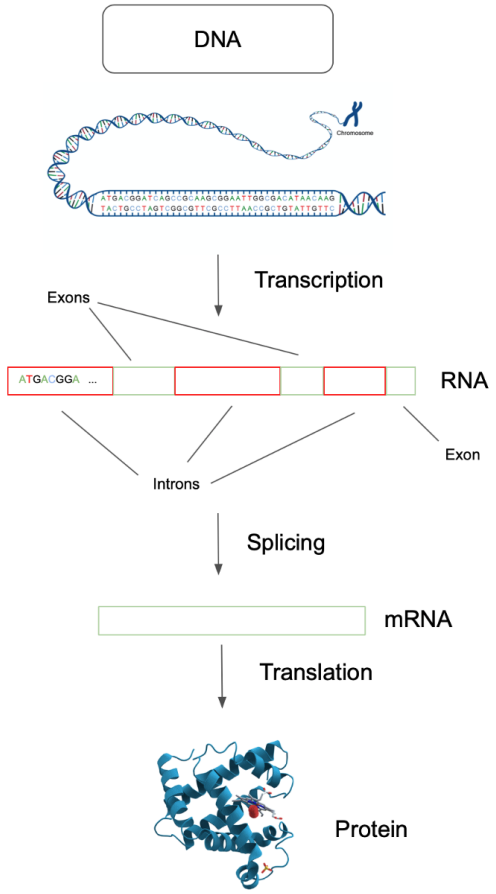

Figure 1. The steps from a gene to a protein. The DNA molecule is stored in the nucleus of the cell on chromosomes. A portion of this DNA molecule - gene - is transcribed into RNA, processed in a process called splicing and subsequently used as a template to form multiple equal proteins.

from it.

After termination of mRNA transcription, the abundance of mRNAs over time can be described by:

$$\text{mRNA abundance}(t) = \text{mRNA abundance}_{t_i} \times e^{-\gamma \times t} \quad (1)$$

where, t stands for the time interval from the last instant with steady state mRNA abundance, t_i, to the current instant and γ is the degradation rate [1]. The time interval which encompasses the reduction of the amount of available-to-translate specific mRNAs in the cell to its half is termed mRNA half-life. The mRNA degradation rate is proportional to the inverse of its half-life:

$$\text{mRNA half-life} = \frac{ln(2)}{\gamma} \quad (2)$$

Each mRNA molecule can be divided into 5 regions: 5'Cap, 5'UTR, coding sequence, 3'UTR and poly-A tail. The 5'UTR, coding sequence and 3'UTR regions contain the code encoded by the gene and carried by the mRNA molecule in the form of a unique set of Adenine (A), Cytosine (C), Guanine (G) and Uracil (U) nucleic acids or bases bound in a single strand. The base thymine (T), the DNA molecule equivalent of uracil (U) is sometimes used interchangeably to refer to the base U.

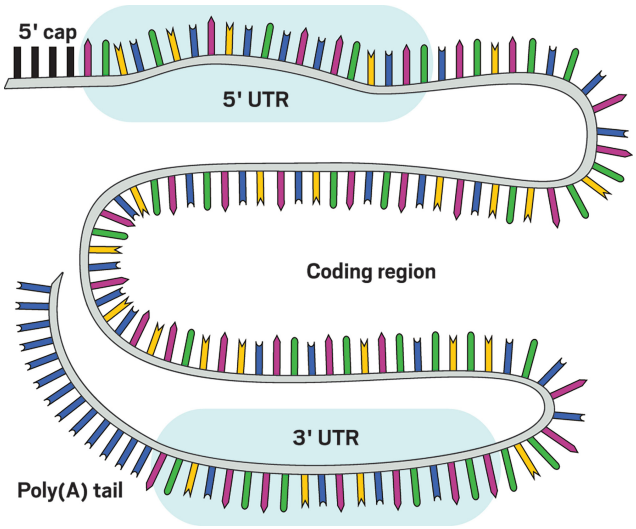

Figure 2. Schematic illustration of an mRNA molecule and its structure. Each blue, red, green and yellow positions in the figure represent the bases A, U, G and C respectively. [2]

The coding sequence of an mRNA contains the set of bases which encode the design of a protein. Amino-acids, the unit-blocks of proteins are encoded in the coding sequence of the mRNA in sets of 3 bases termed codons. Codons are disposed in the coding sequence as sets of adjacent non-overlapping base triplets.

Genneraly, the coding sequence starts with the triplet AUG, also called start codon, which encodes the amino-acid methionine. The ending of the coding sequence is generally marked by one of UAA, UAG, UGA which are termed stop codons. These codons do not encode any amino-acid, their only function is to mark the end of the coding sequence.

Given that there are 4 possible bases for each mRNA sequence position and that each codon is composed of 3 bases, the amount of different codons is $4^3 = 64$. Because the number of different amino-acids in a protein is 20, a number of codons will encode the same amino-acid. This property is termed codon degeneracy. A codon belonging to a set of codons which encode the same amino-acid is defined as synonymous codon.

The 5'UTR or 5' untranslated region, is the mRNA sequence which encompasses the first base from the 5' end of the mRNA to the last base before the start codon. This coding region is particularly important for the assembly of the ribosome - the translational molecular machine - on the mRNA.

The 3'UTR or 3' untranslated region is composed of the sequence from the first base after the stop codon to the last base of the gene-encoded mRNA sequence (right before the poly-A tail).

Some proteins or small non coding RNAs (microRNAs) bind to specific sequence portions of the 5'UTR or 3'UTR termed motifs. This RNA binding molecules directly influence the fate of the mRNA and the translation process. Both the 5'UTR and 3'UTR can encode diverse motifs suitable for regulation of the mRNA in different cell conditions [3].

The sequence in the mRNA defines its interaction with other molecules in the cell such as proteins or RNAs, its molecular structure and in part the translation process, all of which have a direct impact on mRNA half-life. Some elements of the sequence have already been found to be associated with half-life, however a quantitative measure of the influence of these elements in mRNA half-life in human cells to our knowledge does not exist and many are yet to be discovered.

In [4] a model for mRNA half-life prediction on yeast - the highly studied unicellular organism used in baking and the production of alcoholic beverages for thousands of years - which uses only mRNA sequence features was able to explain 59 % of half-life variability between mRNAs. This surprising result set the way for the extension of the quantitative modeling and evaluation of the sequence impact on mRNA half-life in human cells.

Furthermore, the variation of mRNA half-life between human tissues is still far from being extensively studied and quantitatively evaluated.

Deep learning models are now being used to predict several properties and phenotypes of genes and RNA from their sequence. However no such models have focused on predicting half-life from mRNA sequence. Furthermore, many regulatory sequence motifs influencing half-life are still to be discovered.

The presented work leverages the modeling of mRNA half-life through deep learning and regression techniques to address how much of half-life variabilty in human cells we can predict from sequence; what roles the main mRNA known sequence features such as codons and UTRs have; how these features' influence vary between human cells from different tissues; and are there novel mRNA sequence motifs.

Addressing these questions, we developed a linear regression model explaining the quantitative influence of the most well known sequence features impacting half-life in a specific human cell line. We created a deep convolutional neural network for mRNA half-life prediction and applied the interpretation tools DeepLIFT [5] and TF-MoDISco [6] to evaluate the quantitative influence of each sequence position in the prediction output, which revealed possible novel regulatory motifs. We produced a multi-task neural network model for the prediction of mRNA half-life variation between cells from different human tissues. Lastly, we developed a new model-interpretation-based metric which characterizes the effect of the mRNA sequence translation in tissue specific variations of its half-life. Further inspection of this metric uncovered a previously unknown possible connection between a tissue's cell specific mRNA sequence translation effects on half-life and its energy production.

Overall, both the models and the new metric produced in this work can be integrated in approaches to evaluate the impact of mutations in the genetic code of individuals, which in turn helps diagnose and prevent diseases and develop drugs [7].

Furthermore, the quantification of the impact of several sequence features on half-life and the possible uncovering of a new energy-production related pathway add to our understanding of cell biology and can further be the focus of new research.

2. Materials and Methods

2.1. Modeling mRNA in a human cell-line

2.1.1. Data source and brief description. The half-life dataset was obtained from transient transcriptome sequencing (TT-seq) on K562 chronic myeloid leukemia human cells. This dataset consists of 9426 half-life values for each transcript major isoform. The used gene annotation and genomic sequence were GENCODE version 24 and the hg38 (GRCh38) genome assembly (Human Genome Reference Consortium) respectively. For more details see [8].

The half-life measurements follow a distribution approximately symmetric to the median in the logarithmic scale. The 75% quantile is located at 558.27 minutes (9h:18min). The median value is 329.08 minutes (5h:29min), and its standard deviation 967.04 minutes (16h:07min). The maximum half-life is approximately 795 hours or 33 days.

2.1.2. Feature extraction. For each transcript major isoform, each sequence was retrieved using the annotations from GENCODE version 24 and the human genomic sequence from GRCh38. The retrieval was made using the Python packages *pyranges*, *pybedtools*, *kipoiseq*.

The sequences are retrieved as strings with variable length encoding one of A, T, G, C in each position. For each transcript, the 5'UTR, coding sequence and 3'UTR are retrieved separately.

Only transcripts which have a coding sequence starting with the string triplet "ATG" (which corresponds to the start codon AUG) are used, in order to avoid incomplete or uncertain annotations.

Each codon content of the coding sequence was obtained by counting all the non overlapping different triplets starting from the first position of the coding sequence until the last. Every coding sequence was checked for having length which is a multiple of 3.

The frequency of a codon i in a coding sequence is defined as $\frac{\#\text{codon}_i}{\#\text{codons}}$, where $\#\text{codon}_i$ is the number of codons i in the coding sequence and $\#$codons is the number of all codons in the coding sequence.

The GC content of a sequence is defined as $\frac{\#G+\#C}{\#A+\#T+\#G+\#C}$, where A, C, G, T are the bases in the sequence and $\#A+\#T+\#G+\#C$ is equal to the sequence length .

uAUG is a binary variable defining the presence of a "ATG" triplet in the 5'UTR of the transcript.

uORF is an integer variable defining the amount of ORFs in the 5'UTR.

Kozak is a binary variable defining the presence of the sequence (A orG)CCAUGG around the start codon (AUG).

PUM motif is an integer variable defining the amount of UGUANAUA in the 3'UTR.

2.1.3. Ridge regression. The linear model used was a Ridge regression with regularization strength $\alpha = 0.01$. The explained variance score was used to evaluate the model's performance. After feature extraction, and data processing the number of data points was 6524 with 78 features each. The model was fitted in a k-fold cross-validation scheme with 10 folds. The performance of the model was evaluated as the mean of the explained variance scores obtained on the 10 folds. The model was ran through the *scikit-learn* Python package.

2.1.4. CNN Hyperparameter optimization. The choice of the batch size, maximum sequence length size and model architecture parameters - part of the model's hyperparameters - were optimized on the validation set using the mean squared error as evaluation metric, through bayesian hyperparameter optimization, implemented on the python package wandb (Weights and Biases). The architecture parameters comprised the number of (1-dimensional) convolutional layers, the number of filters on each layer, the size of the filters for all layers, the option to double the amount of filters relative to the previous layer, the option to halve the filters' size relative to the previous layer, the option to do a maxpooling operation after each convolutional layer, the option to perform global maxpooling after the last convolutional layer, the number of dense layers, the size of all dense layers and the option to halve the size of dense layers relative to the previous one.

After hyperparameter optimization, the final 5'UTR convolutional neural network had as input a batch with 9 sequences with maximum length 3625 bases during training. If a sequence length was smaller than the maximum length, then the sequence was padded with zeros until having maximum length. A sequence with length higher than maximum length was cut. The resulting convolutional neural network for the 5'UTR model had 1376 parameters and consisted of layers with the following ordering:

- convolutional layer with 11 filters of dimension 8×4 and activation function ReLU
- maxpooling layer with pooling size 2
- convolutional layer with 22 filters of dimension 4×4 and activation function ReLU
- global max pooling layer
- dense layer with output 1 neuron

After hyperparameter optimization, the final 3'UTR convolutional neural network had as input a batch with 25 sequences with maximum length 4180 bases during training. The resulting convolutional neural network had 10161 parameters and consisted of layers with the following ordering:

- convolutional layer with 16 filters of dimension 10×4 and activation function ReLU
- convolutional layer with 32 filters of dimension 10×4 and activation function ReLU
- global max pooling layer
- dense layer with 128 output neurons and activation function ReLU
- dense layer with output 1 neuron

2.1.5. DeepLift. The DeepLift algorithm was applied based on the available implementation at github on kundajelab/deeplift. Ten reference sequences were created from randomly shuffling the original one. The contribution scores using each reference where then averaged.

2.1.6. TF-MoDISco. TF-MoDISco was applied based on the implementation available at github on kundajelab/tfmodisco. The following values for the customizeable parameters were chosen:

```
sliding_window_size=10
flank_size=5
target_seqlet_fdr=0.15
trim_to_window_size=15
initial_flank_to_add=5
kmer_len=5
num_gaps=1
num_mismatches=0
final_min_cluster_size=60
```

2.2. Modelling the variation of mRNA half-life across human tissues

2.2.1. Data. The used dataset comes from the Genotype-Tissue Expression (GTEx) project version 7. It comprises 11688 RNA-seq samples from 714 individuals on 27 different major tissue types. These samples were collected after the death of the individual.

2.2.2. Processing of exonic and intronic coverage. Exons were flanked by 10 bases on each side. The reads mapping completely inside exons were selected as part of the the gene's exonic reads. The reads mapping completely inside introns were selected as the gene's intronic reads. Following a similar procedure as in [9], the exonic and intronic reads were separately normalized for library size. The sum of exonic and the sum of intronic reads of each gene were then log2 transformed and a pseudo-count of 1 was added to the log2 argument. We define $\log_2(\text{exon})$ and $\log_2(\text{intron})$ as the exonic and intronic reads transformation of the last step. In the end, a value of $\log_2(\text{exon})$ and $\log_2(\text{intron})$ was obtained for each gene in each RNA-seq sample.

For each gene in each sample the difference:

$$\log_2(\text{exon}) - \log_2(\text{intron}) = \log_2\left(\frac{\text{exon}}{\text{intron}}\right) \qquad (3)$$

was calculated. Such difference is termed exonic/intronic ratio. The exonic/intronic ratio of a gene in two different samples is related to $\Delta\log_2(\text{half-life})$ by:

$$\Delta\log_2(\text{half-life}) = \log_2\left(\frac{\text{exon}}{\text{intron}}\right)_{s_1} - \log_2\left(\frac{\text{exon}}{\text{intron}}\right)_{s_2} \quad (4)$$

where, s1 and s2 correspond to samples 1 and 2 respectively and $\Delta\log_2(\text{half-life})$ is the half-life log2 difference between a gene in sample 1 and 2.

In a last step, the exonic/intronic ratios of each gene were centered along all samples, meaning that for each gene, its mean exonic/intronic ratio along all samples was subtracted from the exonic/intronic ratio of each sample.

The average exonic/intronic ratio for each gene in each tissue was obtained by averaging the exonic/intronic ratio of each group of samples belonging to one specific tissue.

The transcript exonic/intronic ratio tissue-specific variation from the mean exonic/intronic ratio, or centered exon/intron ratio, is here termed as tissue-specific mRNA half-life variation.

Genes with average TPM (transcripts per million) lower than 2 on a tissue were assumed non-expressed genes and discarded.

2.2.3. Major transcript isoform selection. The major transcript isoform was selected per tissue, by picking the gene's transcript with the highest median TPM (transcripts per million) value across all samples belonging to a tissue. The TPM values for each transcript and sample is available at the GTEx website (GTEx version 7).

2.2.4. Feature extraction. For each transcript major isoform, each sequence was retrieved using the annotations from GENCODE version 19 and the human genomic sequence from GRCh37/hg19. The retrieval was made using the Python packages *pyranges*, *pybedtools*, *kipoiseq*.

Both the extraction process and the features were handled the same as in 2.1.2.

2.2.5. Multi-task DNN hyperparameter optimization. A multi-task deep neural network model was developed to predict each mRNA's half-life variation for each tissue plus the mean (28 tasks on total). The input to this model is a set of 69 features, namely the codon content, the GC content of the 5'UTR and the base 2 logarithm of the 5'UTR length, 3'UTR length and coding sequence length.

The final multi-task DNN model was trained using a batch with 10 samples. The model performing best on a validation set composed of mRNAs corresponding to chromosomes 4,6,9,10 and 13 (around 22% of the 82% of mRNAs not belonging to the test set) was selected and finally, the evaluation was made on a test set with the mRNAs corresponding to chromosomes 3,18,19,20,21 (approximately 18% of the total amount of mRNAs). There are on average 9798 mRNA available per tissue. The final model is composed of 3 fully connected hidden layers with 440 neurons each and activation function rectified linear unit. The final layer outputs 28 values for each one of the tasks (tissue mRNA half-life variation + mean). The number of parameters for this model is 431228.

The loss function used was the mean squared error, taking into account that for each mRNA its half-life variation was often not available for some tissues and therefore those had to be masked. The multi-task DNN model was optimized using the Adam optimizer with learning rate 1e-4.

2.3. A tissue-specific codon effect program

2.3.1. Linear regression model. The linear model used was a Ridge regression with regularization strength $\alpha = 0.01$. Its implementation followed the same characteristics as 2.1.3.

3. Modeling mRNA in a human cell-line

3.1. Association between mRNA half-life and codon content

The plot in figure 3 represents in the y axis the pearson correlation coefficient between the codon frequency in the coding sequence and half-life, also termed CSC (codon stability coefficient). It is possible to see that the frequency of a codon in the coding sequence associates with half-life differently, negatively or positively, depending on the specific codon. Furthermore, the figure indicates that the association of codons with half-life varies between synonymous ones.

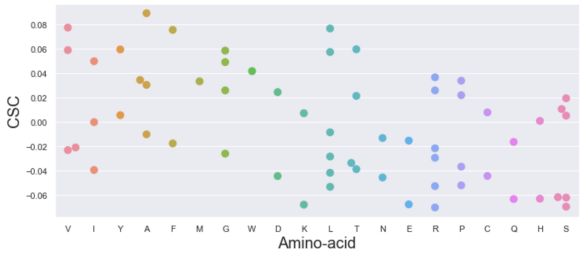

Figure 3. CSC grouped by amino-acid. Each color corresponds to codons encoding the same amino-acid.

3.2. Modeling results

3.2.1. Ridge regression. The mean explained variance score of the model is 0.153 and the mean Pearson correlation coefficient between the predicted and measure values is 0.393.

Table 1 contains an overview of the contribution of each feature to the average explained variance of the model on the test sets (folds). The individual value of a feature is the average explained variance of a model fitted only with that feature. The drop value is the difference between the explained variance of a model fitted on all features and the explained variance of a model fitted on all features but the one in the row. Positions with "-" correspond to features with negative individual values.

The codon content feature, which is the joint contribution of the codon frequencies for all codons, has both the highest individual and drop values, outperforming the other features by a large margin - approximately 6 times higher individual and drop values than the second best performing feature (log (3'UTR length)).

Some features like PUM motifs have a much higher individual value then drop value, indicating that their effect on half-life can be explained by other features.

Others like the stop codons TGA and TAG have a negative explained variance individually, indicating a possibly non-existent relevant contribution to the predicted log(half-life).

TABLE 1. INDIVIDUAL AND DROP EXPLAINED VARIANCE SCORE FOR EACH FEATURE IN THE RIDGE REGRESSION.

Feature	Individual	Drop
uAUG	9.88e-3	-4.24e-4
Stop codon TAA	7.10e-5	0.000468
Stop codon TAG	-	-
Stop codon TGA	-	-
log (3'UTR length)	0.0274	0.0154
log (5'UTR length)	0.0101	0.00122
log (CDS length)	0.0226	0.00547
GC content 5'UTR	0.0162	0.0112
GC content CDS	0.00310	-3.00e-06
GC content 3'UTR	0.00475	2.80e-05
uORF	0.0119	-2.81e-04
Kozak sequence	1.09e-3	-8.80e-5
PUM motifs	0.0123	0.000398
Codon content	0.116	0.0804

3.2.2. Convolutional Neural Networks. For the 5'UTR CNN, the best validation performance (mean squared error = 1.26) was achieved on epoch 134. The training process was stopped after reaching 40 epochs with no improvement on the validation set performance. The achieved explained variance on the test set was 3.329 % and the Pearson correlation coefficient between the measured and predicted half-lives was 0.186 (p-value = 2.126e-11).

On the 3'UTR CNN, the best validation performance (mean squared error = 1.27) was achieved on epoch 64. The training process was stopped after reaching 40 epochs with no improvement on the validation set performance. This model obtained an explained variance of 4.371 % and the Pearson correlation coefficient between the measured and predicted half-lives was 0.217 (p-value = 7.264e-14).

3.2.3. DeepLIFT. Using the obtained 3'UTR and 5'UTR convolutional neural network models, the contribution scores where calculated for each mRNA 3'UTR and 5'UTR sequence separately.

Figure 4 shows the contribution scores for the 3'UTR of an mRNA (PUSL1-201). It is possible to see several contiguous sequence regions with high positive or negative contribution scores.

3.3. TF-MoDISco

The TF-MoDISco algorithm was applied for each set of 3'UTR and 5'UTR DeepLIFT contribution scores. The 4

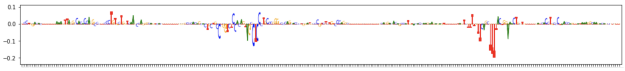

Figure 4. DeepLIFT contribution scores for the mRNA 3'UTR belonging to the PUSL1-201 mRNA. The y axis represents the contribution score and the x axis each position on the 3'UTR. The height of the letters reveals the magnitude of the contribution and the orientation (facing left or down) indicates the sign of the contribution.

motifs with the most amount of seqlets in the 3'UTR are represented in figure 5.

The motif with the most amount of seqlets for the 3'UTR had 662 seqlets and the fourth one had 516. In the 5'UTR set of sequences, the motif with highest amount of seqlets had 1361 and the fourth highest motif had 274 seqlets.

For each motif 2 plots are shown. One having the "real" contribution scores and the other the "hypothetical" ones. The "real" scores are obtained from considering the contribution score of each base present on the seqlet's UTR sequence. The "hypothetical" scores are obtained from considering the contribution score of every possible base for each seqlet position, regardless of it being in the actual sequence or not [10]. In this way, these scores provide extra infered/hypothesized information about the contribution of bases rarely or not seen in the sequence for certain positions based on the knowledge acquired by the model.

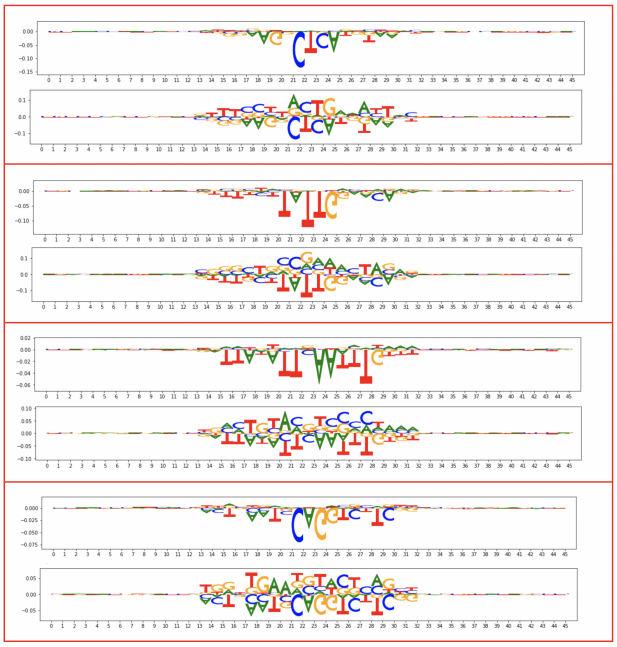

Figure 5. 3'UTR motifs corresponding to the most amount of seqlets ordered from highest on the top to lowest on the bottom. Each red box contains one motif, where the top sequence shows the motif with the "real" contribution scores for each seqlet and the the bottom sequence shows the motif with the "hypothetical" contribution scores. A letter facing up indicates a positive contribution to half-life a letter facing down indicates a negative contribution.

In order to further evaluate each motif effect on half-

life, a comparison between the distribution of half-lives with different number of motif instances in their corresponding mRNAs was made. Figure 6 represents such a comparison, for the motif with the most amount of seqlets on the 3'UTR (AGNCTCA). Notice that, as suggested by the negative motif scores resulting from TF-MoDISco, this motif is present on mRNAs with lower half-lives. The median halfl-life fold change between mRNAs having 2 or more instances of this motif and mRNAs having no instance is 0.73 (Wilcoxon ranksum test p-value = 2.42e-18).

Because the UTR length is correlated negatively with half-life (Spearman correlation coefficient = -0.194, p-value = 1.71e-57), and the probability of having any random motif in the UTR increases with its length, the length can confound the relationship of a certain motif with half-life. For this reason a new metric $f(\text{Half-life})$ was developed, which takes into account the length effect. This metric is calculated by first fitting a linear regression model to predict log2(half-life) from each UTR's length, and secondly by subtracting its predictions from the measured log2(half-life).

When evaluating the motif AGNCTCA this time with the corrected effect, $f(\text{Half-life})$, the distribution of having 2 or more instances of this motif compared with 0 instances is now more similar and the Wilcoxon ranksum test's p-value is higher but still significant (1.37e-3).

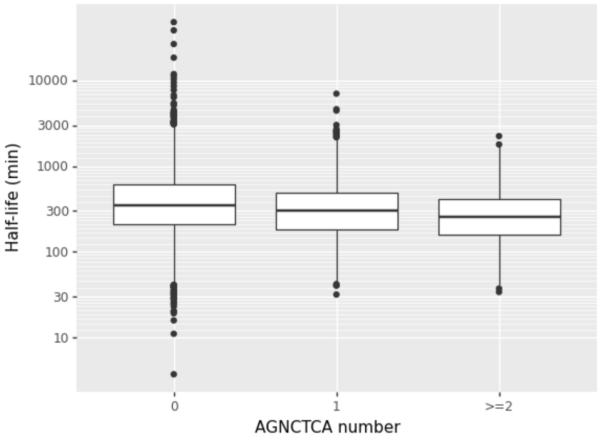

Figure 6. Boxplot depicting the distribution of half-life for mRNAs with 0,1 or 2 or more AGNCTCA motifs in the 3'UTR.

A similar analysis was made for the second motif with most seqlets for the 3'UTR - TATTG and for 2 of the top motifs with more seqlets on the 5'UTR. The sign of mRNA half-life effects of these motifs agree with the sign indicated by the TF-modisco motif scores.

Of noting is the statistical significance of the motif "C (C or A) GCGC", as measured by the p-value of a Wilcoxon ranksum test between the distributions of half-lives of mRNAs with 0 motif instances and with greater or equal to 2 motif instances. If using the half-life values with no correction for length, then this motif appears to have no association with half-life (p-value = 0.636). On the other hand, when using the half-life corrected by the length effect ($f(\text{Half-life})$), the motif association with half-life is positive and seems to be significant (p-value = 0.0123).

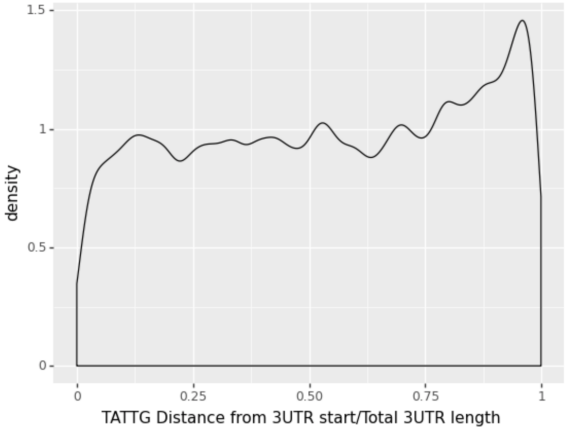

Figure 7. Density plot showing the distribution of the relative position of TATTG on the 3'UTR of mRNAs. The relative position is computed as the quotient between the distance from the beginning of the 3'UTR (after the stop codon) and the total 3'UTR length.

By looking at what position these motifs are found in their respective UTRs, some patterns were found. Figure 7 shows the distribution of the motif TATTG position relative to the total length of the 3'UTR. A similar analysis was made for the motif AAAA. For each distribution its significance was tested using a Wilcoxon ranksum test comparing the distribution and a uniform random distribution with the same length. The AAAA motif appears to have a preference for a location on the 5'UTR close to the start codon, and the TATTG motif appears to have a preference for a 3'UTR location close to the poly-A tail.

4. Modeling tissue-specific mRNA half-life variations

The best validation performance of the multi-task deep neural network (mean squared error = 1.26) was achieved on epoch 134. The training process was stopped after reaching 150 epochs with no improvement on the validation set performance.

The Pearson correlation coefficient between measured and predicted values for each tissue is shown in figure 8.

5. A tissue-specific codon effect program

5.1. Average human individual

We set out to explore how the specific content of codons in an mRNA influences its half-life variation on each tissue for the average human individual.

To that extent, for each tissue, a linear regression with Ridge regularization was developed, taking as input the frequencies of each codon in the mRNA but the 3 stop ones, and as target output the mRNA half-life variation for

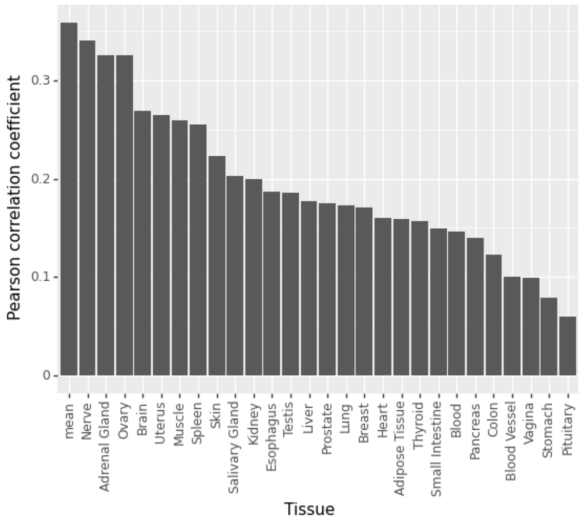

Figure 8. Pearson correlation coefficient per tissue and for the mean value (mean Exonic/Intronic ratio).

that tissue. In order to separate the codon content influence on the target variable from the influence of the coding sequence's GC content, this feature was added as an input to each regression model. Then, after regression, the used linear predictor includes the weights corresponding to all features but GC content.

For each tissue, t_i, the final obtained linear predictor can be defined as:

$$\Delta\log(\text{mRNA half-life})_{t_i} = \beta^{t_i}_{\text{AAA}} f_{\text{AAA}} + ... + \beta^{t_i}_{\text{TTT}} f_{\text{TTT}} + \beta^{t_i}_0 \tag{5}$$

where $\beta^{t_i}_0$ is the intercept, $\beta^{t_i}_{\text{codon}_k}$, $k \in 1, 2, ..., 61$ is the regression coefficient corresponding to codon_k on the predictor for tissue t_i, and f_{codon_k} is the frequency of codon k in the mRNA as defined in 2.1.2.

In total, 27 linear regression models were fitted accounting for all available tissues. Each fitted model was evaluated on a test set with mRNAs belonging to chromosomes 3, 18, 19, 20, 21. The highest Pearson correlation coefficient (0.355) between predicted and measure values is on Nerve and the median one is 0.172.

The linear model described in 5 can be used to further inspect the effect of each codon in ΔmRNA half-life, as captured by the model. As

$$\frac{\partial \Delta\log(\text{mRNA half-life})_{t_i}}{\partial f_{\text{codon}_k}} = \beta^{t_i}_{\text{codon}_k} \tag{6}$$

such can be done by analyzing the regression coefficients $\beta^{t_i}_{\text{codon}_k}$.

Changing the content of a codon k by $\Delta f_{\text{codon}_k}$ while keeping constant the content of every other codon will change $\Delta\log(\text{mRNA half-life})_{t_i}$ by $\Delta f_{\text{codon}_k} \beta^{t_i}_{\text{codon}_k}$.

Therefore, the sign of $\beta^{t_i}_{\text{codon}_k}$ indicates the positive or negative effect of codon_k in $\Delta\text{mRNA half-life}_{t_i}$ and the magnitude of $\beta^{t_i}_{\text{codon}_k}$ indicates the strength of this effect.

The tissue-specific codon effects on $\Delta\log(\text{mRNA half-life})$ as measured by $\beta^{t_i}_{\text{codon}_k}$ can be visualized in the form of the clustered heatmap of figure 9.

Further analysis of the heatmap reveals two distinct tissue clusters: group α and group γ. Group α is composed of the tissues heart, adrenal gland, brain, liver, esophagus, kidney and muscle, while group γ comprises the remaining tissues.

Furthermore, the heatmap codon clustering suggests two codon patterns, easier to notice when looking at the overall red and blue coloring on the tissues of group α or the third hierarchical level of the codon's dendogram.

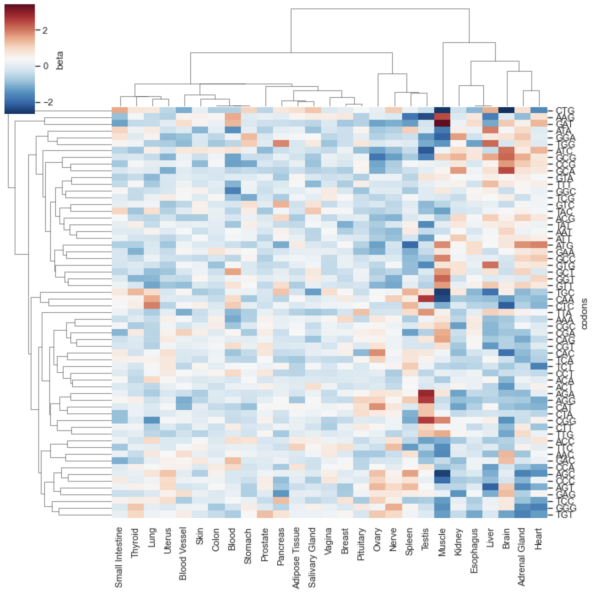

Figure 9. Clustered heatmap depicting the relationship of β across different tissues and codons.

A tissue i specific codon effects can be described as an n-dimensional vector composed of $\beta^{t_i}_{\text{codon}_k}$, where $n =$ number of codons $= 61$.

By finding the principal components of the n-dimensional space, each tissue's coordinates where projected into the two principal components accounting for the highest percentage of explained variance, 31.5% for PC1 and 18.0% for PC2.

Figure 10 shows each tissue's specific codon effects expressed in two coordinates PC1 and PC2. From this representation we can see a distinction between tissue group α and γ coordinates in the PC1 axis. Furthermore, along the PC2 axis, muscle shows a distinct value from all the tissues in group α (more than 3 times higher than the closest tissue in PC2 - esophagus). Such value is closer to testis, indicating that the codon effects of testis and muscle on ΔmRNA half-life share similarities in a domain different from the one encoding the tissue α - tissue β dichotomy.

PC1 accounts for approximately one third of the total explained variance, and therefore captures the bulk of tissue-specific mRNA half-life variations due to codon content. In

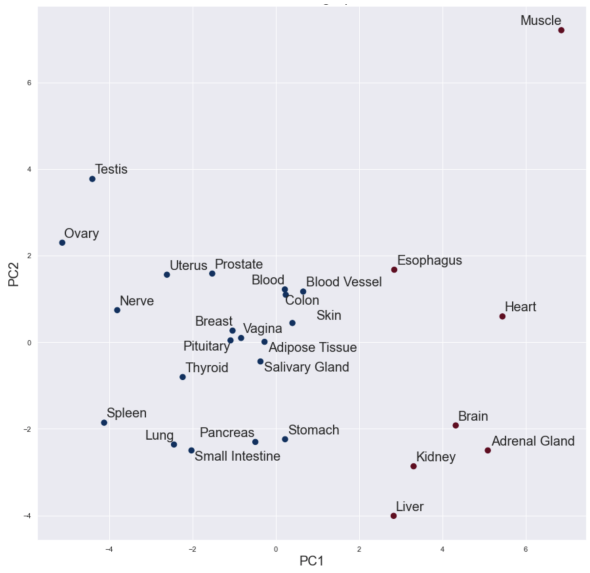

Figure 10. Tissues projected into the PC1 and PC2 components. The colors represent 2 clusters obtained by applying k-means clustering on the tissue's coordinates in the n-dimensional space.

light of this fact, we created a new metric termed tissue codon signature, defined for each tissue as its first principal component value.

We set out to further explore the newly developed tissue codon signature metric by investigating its relationship with transcript amounts.

To accomplish that, we retrieved a gene's tissue's transcript amounts, defined by a vector with length equal to the number of tissues and whose value's are the gene's tissue-specific TPM. Afterwards we defined a vector containing the codon signature value of each tissue. We inspected the relationship between these 2 vectors by computing the Spearman correlation coefficient.

All 38793 genes's TPMs were then ordered by their Spearman correlation with the codon signature. Surprisingly, the majority of the top correlating genes are related to mitochondrial pathways and more than half of those are encoded by the mitochondrial DNA. From the 20 top correlating genes all show such properties.

5.2. Tissue-specific codon signatures across human individuals

Having analyzed the tissue-specific codon signatures on the average human individual, we set out to further inspect the codon signatures between different human individuals. In order to accomplish that, a linear regression model was fitted to predict Δlog(mRNA half-life) of each sample, that is to say each tissue of each individual. Similarly to the steps described on the previous section, $\beta^{t_i}_{\text{codon}_k}$ was obtained for each individual's tissue. Furthermore, a principal component analysis was performed, describing each tissue-individual pair (sample) in the space with the 2 principal components with the most explained variance. A visualization of this representation is shown in figure 11. It is possible to see that overall, the samples cluster together into tissues, pointing that the codon signatures present more variability between tissues than individuals.

Figure 11. Samples projected into the PC1 and PC2 components. The colors map to tissues. Only some tissues were plotted in order to allow a better visualization.

Previously, we have shown the consistency of the codon signature metric across same tissue types of different individuals. Now, we explore how this metric relates to specific individual traits such as age and sex, and a specific sample acquisition characteristic termed ischemic time.

Overall, the age of the individual correlates negatively with codon signature (Spearman correlation: 0.049 P-value 2e-4), with the strength of the effect largely depending on the tissue. The strongest negative Spearman correlation of value -0.35, was found for blood vessel tissue.

Overall, the individual's sex didn't show any correlation with codon signature. A Wilcoxon ranksum test between the distribution of codon signatures of both sexes for all samples produced a p-value of 0.72. Per tissue, sex differences between codon signature distributions were also not found to be significant.

Lastly, we looked at ischemic time, which is the time interval between the actual or presumed death of the individual and the stabilization of the tissue sample. Overall, ischemic time shows a negative correlation with the codon signature (Spearman correlation coefficient = -0.15, p-value = 2.34e-34). This correlation strongly varies across tissues, showing a high negative correlation on heart and lung tissue (Spearman correlation = -0.69 for heart and -0.64 for lung).

6. Conclusion

Quantitative modeling of selected mRNA sequence features through linear regression indicates codon content ex-

plains by far the highest amount of half-life variability between mRNAs in human cells, in line with other quantitative modeling results on the yeast organism [4].

Inspecting the codon stability coefficients (CSC), showed that different codons associate positively or negatively with half-life. Interestingly, it showed synonymous codons have an heterogeneous association with half-life. In fact, codons encoding the same amino-acid are sometimes found in different ends of the codon stability coefficient spectrum. Such synonymous codon heterogeneity allows for an mRNA to have different half-lifes while still encoding the same protein. In fact, as shown in [11], replacing an mRNA coding sequence with different but synonymous codons, can increase or decrease half-life tenfold.

Both convolutional models on the 5'UTR and 3'UTR were successful in capturing more variance than the combined extracted features used in the linear model for these UTRs. In fact, for both the 5'UTR and 3'UTR, its GC content and length together account for approximately one third of the variance explained by the convolutional neural network on the test set. This indicates that both models were able to capture new features, possibly new regulatory motifs.

The four most supported candidate motifs obtained from TF-MoDISco were all found to be significantly associated with half-life, with the sign of the association in accordance to what was indicated by the TF-MoDISco output scores. This effect was also found significant when this association was corrected for the UTR length, which when higher is often associated with lower half-lives. Furthermore, two of this motifs UAUUG on the 3'UTR and AAAA on the 5'UTR were found to have a preference for a certain position inside the UTR, namely a preference for a position near the poly-A tail for UAUUG, and close to the start codon for AAAA. This fact endorses both motifs to be of relevant biological significance. A next step to further evaluate these motifs could be a study of their conservation throughout different species. Lastly the function and biological significance of these motif candidates can finally be checked by an experimental assay. To our knowledge, none of this motifs were already discovered.

Modeling mRNA half-life variations between mRNAs in different tissues could be achieved with the multi-task DNN. The reasoning behind developing a multi-task deep neural network model was to leverage the possible interrelationships between tissues to increase the predictive power of the final model. The resulting model was able to predict mRNA half-life variations with performance varying significantly between tissues. In the end, the resulting multi-task deep neural network performance endorses the usage of this trained model in other settings, such as its integration in models to predict tissue-specific mRNA levels.

The linear regression model interpretation allowed us to evaluate the relationship between each codon and the tissue-specific differences in half-life it associates with. It provided a quantitative description of such relationship in the form of $\beta^{t_i}_{\text{codon}_k}$, which we termed tissue-specific codon effect.

Interestingly, codon effects largely vary across tissues. Their analysis suggests that these variations follow a pattern in which groups of tissues share similar codon effects. In particular, one group (tissue group α) seems to be composed of tissues associated with high energy demands.

The newly developed metric, codon signature, explains most of the codon effect variability between tissues and therefore allows us to characterize each tissue in terms of the particular set of codon effects.

The correlation between a tissue's codon signature and transcript amounts uncovered a previously unknown connection between mitochondrial activity and codon effects. Such connection opens up new research directions and questions.

The half-life of an mRNA is affected by its codon content through the translation rate, in which the decoding rate of each codon is a determinant. By capturing the association between Δhalf-life on tissue i and the frequency of codon k, $\beta^{t_i}_{\text{codon}_k}$ can convey information on the influence codon k has on its decoding rate/time on tissue i compared to other tissues.

In this sense, the codon signature metric not only encodes the relationship between codon content and tissue-specific mRNA half-life variation but can also encode the particular influence each codon has on its decoding rate, conditioned on the tissue and relative to the average across tissues.

Mitochondria produce energy in the cell by making the energy molecule ATP, using oxygen and bio-molecules such as derivatives of glucose. As measured by the amount of transcripts of mitochondrial genes and reported in [12], some tissues contain a higher amount and activity of mitochondria. Therefore, the rate of ATP production and the concentration of ATP molecules can vary across tissues.

Codon decoding is the most energy demanding step of translation [13]. For the decoding of one codon two GTP and ATP energy molecules are required. We hypothesize that the GTP/ATP availability can be a time-limiting step on decoding, whose influence varies by nature of the mitochondrial ATP production of the tissue or, in other words, its overall energy production. A lower decoding rate makes for a lower mRNA half-life and lower translated proteins as outcome.

The association of ischemic time and age with codon signature is consistent with this hypothesis. The longer the ischemic time, the fewer the available oxygen in the cells, which in turn decreases the cell's ability to generate ATP through mitochondria. This effect can explain the fact that ischemic time correlates negatively with codon signature. Meaning that, under ischemia, the codon signature of high energy production tissues changes in the direction of the codon signature of lower energy production tissues. Furthermore, mitochondrial dysfunction is one of the main contributors to the human aging process [14]. The older the individual gets, the lesser the quality of the mitochondria and therefore its energy production capability. Indeed, the computed codon signature metric decreases with age.

In the end, we suggest that, depending on the cell's energy production, a specific set of codons will have an enhanced positive/negative impact on mRNA half-life and translation rate. Such sets can possibly drive the expression

of groups of genes with corresponding similar mRNA codon contents. Further analysis should be made to uncover which genes have a codon content that make their expression most sensible to cellular energy production.

Ultimately, our analysis raises the unprecedented hypothesis that the compound expression of genes involved in certain pathways in the cell can be in part tuned/coordinated by the cell's energy production through mRNA translation, and that such can explain part of the gene expression variability characterizing cells in different tissues or conditions.

References

[1] A. Zeisel, W. J. Köstler, N. Molotski, J. M. Tsai, R. Krauthgamer, J. Jacob-Hirsch, G. Rechavi, Y. Soen, S. Jung, Y. Yarden *et al.*, "Coupled pre-mrna and mrna dynamics unveil operational strategies underlying transcriptional responses to stimuli," *Molecular systems biology*, vol. 7, no. 1, p. 529, 2011.

[2] R. Cross, "Can mrna disrupt the drug industry?" *Chemical & Engineering News*, vol. 96, p. 35, 2018.

[3] C. Mayr, "Regulation by 3 - untranslated regions," *Annual review of genetics*, vol. 51, pp. 171–194, 2017.

[4] J. Cheng, K. C. Maier, Ž. Avsec, P. Rus, and J. Gagneur, "Cis-regulatory elements explain most of the mrna stability variation across genes in yeast," *Rna*, vol. 23, no. 11, pp. 1648–1659, 2017.

[5] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[6] A. Shrikumar, K. Tian, Ž. Avsec, A. Shcherbina, A. Banerjee, M. Sharmin, S. Nair, and A. Kundaje, "Technical note on transcription factor motif discovery from importance scores (tf-modisco) version 0.5. 1.1," *arXiv preprint arXiv:1811.00416*, 2018.

[7] J. Shendure, G. M. Findlay, and M. W. Snyder, "Genomic medicine–progress, pitfalls, and promise," *Cell*, vol. 177, no. 1, pp. 45–57, 2019.

[8] L. Wachutka, L. Caizzi, J. Gagneur, and P. Cramer, "Global donor and acceptor splicing site kinetics in human cells," *Elife*, vol. 8, p. e45056, 2019.

[9] D. Gaidatzis, L. Burger, M. Florescu, and M. B. Stadler, "Analysis of intronic and exonic reads in rna-seq data characterizes transcriptional and post-transcriptional regulation," *Nature biotechnology*, vol. 33, no. 7, pp. 722–729, 2015.

[10] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *arXiv preprint arXiv:1704.02685*, 2017.

[11] V. Presnyak, N. Alhusaini, Y.-H. Chen, S. Martin, N. Morris, N. Kline, S. Olson, D. Weinberg, K. E. Baker, B. R. Graveley *et al.*, "Codon optimality is a major determinant of mrna stability," *Cell*, vol. 160, no. 6, pp. 1111–1124, 2015.

[12] E. Fernández-Vizarra, J. A. Enríquez, A. Pérez-Martos, J. Montoya, and P. Fernández-Silva, "Tissue-specific differences in mitochondrial activity and biogenesis," *Mitochondrion*, vol. 11, no. 1, pp. 207–213, 2011.

[13] M. Leibovitch and I. Topisirovic, "Dysregulation of mrna translation and energy metabolism in cancer," *Advances in biological regulation*, vol. 67, pp. 30–39, 2018.

[14] S. Srivastava, "The mitochondrial basis of aging and age-related disorders," *Genes*, vol. 8, no. 12, p. 398, 2017.